Psychological Review
2015, Vol. 122, No. 3, 429–460

© 2015 American Psychological Association
0033-295X/15/$12.00 http://dx.doi.org/10.1037/a0039348

Separating Decision and Encoding Noise in Signal Detection Tasks

Carlos Alexander Cabrera and Zhong-Lin Lu
The Ohio State University

Barbara Anne Dosher
University of California, Irvine

In this article we develop an extension to the signal detection theory framework to separately estimate internal noise arising from representational and decision processes. Our approach constrains signal detection theory models with decision noise by combining a multipass external noise paradigm with confidence rating responses. In a simulation study we present evidence that representation and decision noise can be separately estimated over a range of representative underlying representational and decision noise level configurations. These results also hold across a number of decision rules and show resilience to rule miss-specification. The new theoretical framework is applied to a visual detection confidence-rating task with 3 and 5 response categories. This study compliments and extends the recent efforts of researchers (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008; Rosner & Kochanski, 2009; Kellen, Klauer, & Singmann, 2012) to separate and quantify underlying sources of response variability in signal detection tasks.

Keywords: decision noise, internal noise, external noise, signal detection theory, confidence rating

Signal detection theory (SDT; Green & Swets, 1966; Peterson, Birdsall, & Fox, 1954; Tanner & Swets, 1954) remains one of the most influential models of cognitive science. Disparate areas of psychological research have adopted SDT as an explanatory framework for a broad range of topics including sensation and perception (Tanner & Swets, 1954), category perception (Macmillan, Kaplan, & Creelman, 1977), recognition memory (Wickelgren & Norman, 1966), attention (Lu & Dosher, 1999), perceptual learning (Dosher & Lu, 1998, 1999), group decision behavior (Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001), neurophysiology (Britten, Shadlen, Newsome, & Movshon, 1992), and clinical applications (McFall & Treat, 1999). Many studies have found application for SDT in areas far beyond traditional psychological studies (Hutchinson, 1981; McClelland, 2011).

The fundamental assumptions of SDT include a representation stage and a response stage. The representation stage assumes a noisy transformation mediating the mapping between an external stimulus and an internal response along a decision axis. Over the course of many trials, a specific stimulus elicits internal responses with some mean level of activation (corresponding to stimulus strength) and some variability (corresponding to the noise in the internal response), so that the observer's internal representation takes the form of a probability density function. Stimuli of different strengths lead to probability density functions with different means along the decision axis and potentially different variances as well. The response stage assumes that observers use criteria to partition the decision axis to map internal responses to observable decisions (Figure 1, top panel).

This relatively simple model has recently been described as one of the most successful "theoretical frameworks" and "mathematical models" in psychology (Benjamin et al., 2009; Kellen, Klauer, & Singmann, 2012). However, results from a number of studies have undermined some of the assumptions of SDT, most notably the assumption that decision criteria remain fixed upon a decision axis over the sequence of trials in an experiment (Benjamin, Tullis, & Lee, 2013; Mueller & Weidemann, 2008; Wickelgren, 1968). An alternative possibility is that decision criteria fluctuate from trial to trial over the course of the experiment (Figure 1, bottom panel). Evidence that challenges the noiseless decision mechanism may appeal to a reevaluation of the principle measures of sensitivity and bias, as decision noise may modify the interpretation of these estimates and the conclusions drawn from them. Experimental methods capable of distinguishing representation and decision noise in signal detection tasks will serve to estimate decision noise and to evaluate the impact of criterion variability on SDT parameter estimates. So far, such methods are few and restrictive, so that it is often impossible to know whether reevaluation is even necessary for many SDT tasks. In this article, we build such a framework to separately estimate decision and representation noise components at the decision stage.

We begin with an overview of the SDT framework and a review of the empirical evidence suggesting that decision boundaries are

Carlos Alexander Cabrera and Zhong-Lin Lu, Laboratory of Brain Processes (LOBES), Center for Cognitive and Brain Sciences, Department of Psychology, The Ohio State University; Barbara Anne Dosher, Department of Cognitive Sciences and Institute of Mathematical Behavioral Sciences, University of California, Irvine.

This research was supported by National Institute of Health Grant MH081018 and National Eye Institute Grant EY017491. Matlab scripts related to this research are available for download at http://lobes.osu.edu/downloads/MCR.zip. We would like to express our gratitude to Bosco Tjan, David Kellen, and three anonymous reviewers for their especially helpful suggestions and comments. Additional thanks to lab members of LOBES and T-Lab for their very useful feedback during the research and writing of this study.

Correspondence concerning this article should be addressed to Carlos Alexander Cabrera, Department of Psychology, The Ohio State University, 60 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210. E-mail: cabrera.36@osu.edu

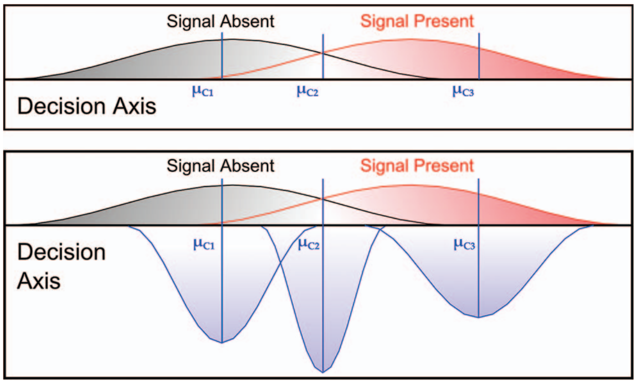


Figure 1. Top: decision axis under a classical confidence rating framework. Representations of signal-absent and signal-present distributions take the form of Gaussian probability density functions. The subject uses static criteria to partition the decision axis to map internal representations to overt responses. Bottom: a modified confidence rating framework in which the criteria are formulated as probability density functions with means μ_{C1}, μ_{C2}, and μ_{C3} because of trial by trial variability in decision processes. In this and later figures, probability density functions for criteria noise are shown reflected below the decision axis for clarity. See the online article for the color version of this figure.

variable or noisy, along with a review of recent efforts to identify and quantify decision noise in categorical judgment tasks with at least three stimulus classes (Rosner & Kochanski, 2009). We then develop a new framework that combines a decision noise model for a confidence rating procedure with a multipass external noise paradigm (Burgess & Colborne, 1988; Green, 1964; Lu & Dosher, 2008). Using simulations, we demonstrate the feasibility of parameter recovery that estimates the separate contributions of decision and representation noise for three different decision rules. Our development applies to tasks with only two stimulus classes over a range of possible underlying noise configurations, that is, different relative levels of representation and criterion noise. We then illustrate this method with an application using a multipass visual detection experiment with external noise. Finally, we consider some ideas for future studies as well as limitations of this framework. Details of our experiment along with derivations and a more formal analysis of this framework are provided in Appendix A–C.

SDT and Static Criteria

In a typical yes/no signal detection experiment, an observer monitors an observation interval for the presence of a designated signal stimulus. The observer responds affirmatively if she believes the signal was present during this interval. The observer cannot respond with perfect accuracy on every trial, sometimes correctly reporting the presence of a signal when a signal stimulus in fact occurred, but sometimes incorrectly affirming the presence of a signal when a signal was not present. The hit rate (HR) is the relative frequency of saying "yes" when a signal is present; the false alarm rate (FAR) is the relative frequency of saying yes when a signal is not present. Miss and correct rejection rates are the relative frequencies of saying "no" when a signal is present and when a signal is absent. Manipulation of the observer's yes rate by changing task instruction, pay-off structure, or stimulus base rates elicits different values of HR and FAR, and the HR plotted against the FAR defines the receiver operating characteristic (ROC; Figure 2, left; Green & Swets, 1966).

The data from empirical ROCs often comprise the fundamental features researchers wish to model in signal detection tasks. In most applications, SDT posits internal representations in the form of Gaussian random variables with mean values positioned along a decision axis and monotonically related to stimulus strength (Graham, 1989). Consequently, the representational distributions of two stimuli of different strength often overlap, leaving some nonzero likelihood that a stimulus sample from either stimulus class (signal present or signal absent) could have generated the internal response in a given trial. Many signal detection models assume that the observer responds by

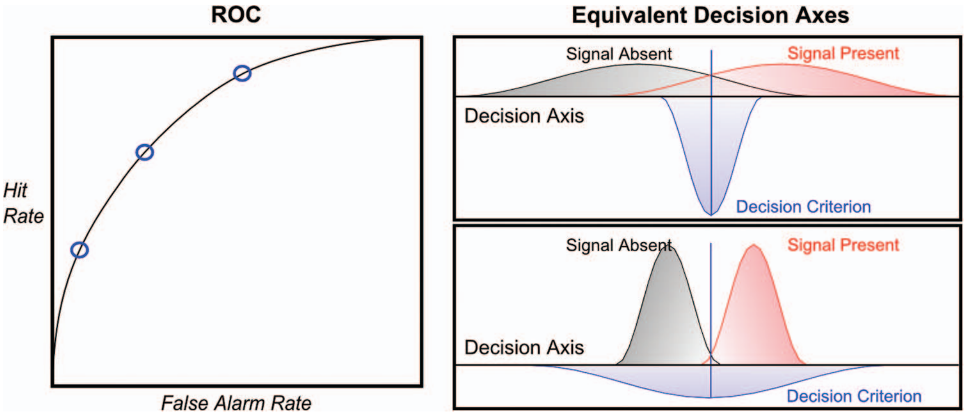


Figure 2. Left: A receiver operating characteristic (ROC) with three different decision criteria. When the signal strength is low, performance decreases, values of hit rate and false alarm rate converge, and the ROC curve approaches the unity slope. With higher signal strength, hit rate and false alarm rate diverge, so the ROC curve moves up and to the left. Right: underlying distributions of stimulus representations at the decision stage shown with high encoding noise and low decision noise (top panel) and an alternative representation with lower encoding noise and higher decision noise (bottom panel), each leading to the same performance outcome. See the online article for the color version of this figure.

establishing a boundary or criterion along the decision axis, and chooses yes when the value of the sampled internal representation exceeds this criterion, and chooses no otherwise (Figure 2, right panels). Internal representations from signal present trials exceeding the criterion contribute to HR, and representations of signal absent trials exceeding the criterion contribute to FAR. Insofar as distributions of internal representations really do approximate Gaussian probability density functions, HR and FAR may be transformed into standardized scores (z-scores) to indicate the position of the criteria along the decision axis in units of the *SD* of the underlying distributions (see Appendix A). Empirical *z*-transformed ROC (zROC) functions are often approximately linear, consistent with the Gaussian distribution assumption (Macmillan & Creelman, 2004). The classical SDT model does not incorporate trial-by-trial variability in the criterion position, so all response variability accrues from variations in the internal representations of the stimuli (Benjamin et al., 2009).

While some simple SDT applications assume equal variances for signal present and signal absent distributions, researchers frequently relax this equal variance assumption to account for the nonunity slopes often observed in many empirical zROCs. Meanwhile, the static criterion assumption has rarely been relaxed. Early formulations of SDT excluded decision noise for two reasons (Tanner & Swets, 1954). First, because a static decision mechanism was optimal and part of a cognitive operation, an observer would not willingly choose to vary its operation from trial to trial, because this variable strategy would lead to lower overall performance (Benjamin et al., 2013; Mueller & Weidemann, 2008). And second, typical analyses of signal detection data simply could not differentiate between noise arising from representational and decision-related processes (Figure 2, right panels; see Wickelgren, 1968).

Evidence for Criterion Variability

Though practical considerations led to omissions of criterion variability in early applications of signal detection theory, in fact, lines of evidence suggesting a variable decision process predate even the Thurstonian framework (Fernberger, 1920). Later, reduced performance on absolute identification in experiments with increased stimulus range was attributed to increased variance in identification criteria (the range effect; Pollack, 1952). Early research in auditory amplitude identification led to the explanation that the change in response variability arose because of subjects exhibiting a range-dependent criterion noise (also interpreted as memory noise; see Durlach & Braida, 1969). Later research suggested an independence between the range effect and the total number of response categories (Braida & Durlach, 1972) and specifically implicated the criterial range as the source of the performance decrement (Gravetter & Lockhead, 1973), though not to the exclusion of representation-related mechanisms as well (Luce & Nosofsky, 1984; Luce, Nosofsky, Green, & Smith, 1982; Nosofsky, 1983). Additionally, investigators have invoked criterion noise to help explain anomalies in the shape of the ROC curve (Mueller & Weidemann, 2008; Murray, Bennett, & Sekuler, 2002; Wickelgren, 1968); discrepancies in distribution-free estimates of response bias in confidence rating tasks (Mueller & Weidemann, 2008); performance decrements related to larger rating scales in confidence ratings tasks (Benjamin et al., 2013); and feedback-associated manipulation (Carterette, Friedman, & Wyman, 1966) and learning (Friedman, Carterette, Nakatani, & Ahumada, 1968) in auditory amplitude detection. Others have suggested that decision noise results from criterion-setting mechanisms for reconstructing stimulus representations at the decision level (Parks, 1966); and that criterion noise is related to nonoptimal criterion shifting (Thomas, 1973, 1975). For a more extensive review, see Benjamin et al. (2009).

Although we have presented a small sample here, evidence arising from these disparate research areas has generated a great body of literature implicating the presence of criterion variability. Along with these empirical results, a literature of theoretical contributions has also emerged (e.g., Kac, 1962; Treisman, 1984; Treisman & Williams, 1984). Strictly speaking, to whatever extent quantitative models can account for the phenomena of criteria shifting, we can no longer refer to this as “noise” in the proper sense of the word. We here follow earlier writers who have disambiguated “systematic” noise from “unsystematic,” “irreducible,” or “random” noise (Levi, Klein, & Chen, 2005; Rosner & Kochanski, 2009). We now turn to the research efforts to separate and measure decision noise.

Decision Noise Methods and Models

Analysis of the categorical judgment task showed that standard signal detection experimental procedures could not generally distinguish representational noise from decision noise without significant simplifying assumptions (Rosner & Kochanski, 2009; Torgerson, 1958). The first serious research effort to understand the influence of decision noise began with Wickelgren and his study of response predictions for a variety of signal detection task conditions in the presence of significant criterion noise (although see also Tanner, 1961, for consideration of decision noise under a less rigid interpretation of decision criterion in a two-alternative forced choice task). In a seminal paper, Wickelgren (1968) examined the ramifications of decision noise for subject performance in yes/no and confidence rating tasks. He derived functional forms for the zROC and showed that observers with nontrivial decision noise could produce linear zROCs as long as decision noise remained constant across criteria and task structure did not alter representational characteristics (see also Benjamin et al., 2009). Static criteria with Gaussian representational distributions lead to linear zROCs, but linear zROCs do not necessarily imply static criteria. Wickelgren also considered the implications of attenuated criterion noise at a primary decision boundary relative to the remaining criterion boundaries in bipolar confidence rating tasks and the data signature this affords in a zROC curve (see also Mueller & Weidemann, 2008; Murray et al., 2002). In particular, he observed that subjects could exhibit a peaked zROC when criterion noise at the primary decision boundary is significantly less than the decision noise at the remaining boundaries. Reviewing studies with greater numbers of category boundaries, he often identified larger peaks, leading to the speculation that increasing the number of category boundaries could increase decision noise. This finding was consistent with Miller’s famous article on information retrieval (Miller, 1956) and the criterial range interpretation of the range effect (Gravetter & Lockhead, 1973) insofar as additional criteria lead to broader criterion spread across the decision axis.

Wickelgren’s close examination of the shape of subjects’ ROCs and zROCs became a standard diagnostic approach for criterion variability in signal detection type tasks. However, because data collection in typical yes/no tasks requires bias manipulations that might alter either representational or decision processes, researchers preferred confidence rating procedures for their greater assurances of representation and decision noise stability over the duration of the experiment. However, even studies using rating procedures may have fallen short of unambiguous estimates of representation and decision variability owing to tradeoffs between these parameters in estimation (e.g., Benjamin et al., 2009; Mueller & Weidemann, 2008).

Nosofsky (1983) developed a multiple presentation method to examine the range effect with an identification task. On individual trials in his study, subjects made multiple responses to repeated identical presentations of a stimulus from one of the available stimulus classes. Although he treated each response as independent of the others, he assumed that noisy internal representations were averaged while decision noise remained constant across presentation repetitions. By separately measuring sensitivity for each presentation repetition, he demonstrated nontrivial decision and representational noise with both components increasing with larger criterion range.

Benjamin et al. (2009) developed an Ensemble Recognition task similar to the multiple presentation method of Nosofsky to examine the effects of decision noise in memory recognition. In this study, subjects were first presented a study list of words they would later be asked to recognize during a test phase. During the test phase individual trials contained ensembles of one, two, or four words. Each ensemble contained either one, two, four, or no words from the previously examined study list. The Ensemble Recognition framework assumed that each word of each trial ensemble led to internal activations independent of the other words, and that either the sum or the average of these activations would comprise the internal representation at the decision stage. Similar to Nosofsky, these authors assume that the decision noise remained constant while the summing or averaging would lead to adding or averaging of the representational noise. The averaging model performed best in model selection tests and estimated a very significant role for decision noise in word recognition.

More recently, Kellen et al. (2012) offered a critique of the conclusions drawn from the Ensemble Recognition study and provided new reports on the question of decision noise in memory recognition using a model generalization framework. This approach involves combining a four-alternative forced choice (4AFC) task with a rating procedure under the traditional assumptions that internal representations are identical under the two regimes and that response bias does not play a role in subject response during forced choice tasks. They jointly fit their elaborated SDT model with decision noise to data from both the 4AFC and the confidence rating tasks but found virtually no significant decision noise influencing subject performance in their memory recognition experiments.

Rosner and Kochanski (2009) developed a categorical judgment model to separately estimate criterion noise at decision boundaries. They corrected an error in an earlier formal description of a categorization task that allowed for decision noise in absolute identification and confidence rating tasks (Torgerson, 1958). However, Rosner and Kochanski showed that the earlier formulation failed to account for the fact that truly independent noisy criteria might overlap from trial to trial and could result in predictions of negative response frequencies. Their revised formalization accounts for this overlap and can be reduced to two special cases: in the absence of decision noise the model simplifies to the traditional SDT model, and in the absence of representation noise the model simplifies to a complimentary SDT model (a formulation that ascribes all response variability to noisy criteria). Using simulated experiments, Rosner and Kochanski showed parameter recovery was possible for a range of assumed parameter configurations. They argued that the general formulation of the model disambiguated the conflated parameters, and that acquiring sufficient degrees of freedom in data posed the only constraint to full parameter estimation. In particular, a categorization task with N stimulus classes and $M + 1$ response categories requires identification of the means and variances of $2N$-2 stimulus parameters (assuming a reference stimulus class with mean 0 and variance 1) and $2M$ criterion parameters. This categorization task has NM independent data points, so that full model identification is possible only when $NM > 2(N + M)$-2; that is, when both $N > 2$ and $M > 2$. For the standard signal detection paradigm with two stimulus classes ($N = 2$), a solution is available only if the criterion variances are assumed equal at all category boundaries.

A New Approach

Intuitions and Rationale

We now develop a framework combining two well-known experimental paradigms to estimate both representational and decision noise components in signal detection type tasks with only two stimulus classes, S_0 and S_1 (where 0 refers to signal absent trials and 1 refers to signal present trials). The first paradigm is a confidence-rating task in which subjects provide a rating R_{i} indicating their degree of certainty that the present trial contains a signal stimulus (Egan, Shulman, & Greenberg, 1959). The second component is the multipass procedure, an external noise paradigm involving multiple presentations of identical stimuli (Burgess & Colborne, 1988; Green, 1964; Lu & Dosher, 2008). We show that this combination sufficiently constrains elaborated signal detection models by providing measures of agreement in addition to rating frequencies.

Here we offer some basic intuitions to illustrate our strategy for dissociating representation and decision noise components. To begin with, we simplify our exposition by considering response variability with a single criterion C with stimulus class S_{h}, where $h = 0$ or 1. If an observer responds differently to two or more trial presentations with identical stimuli, we attribute the change in response to internal noise. Researchers have explored this basic idea by adding external noise to stimulus presentations to estimate internal noise (Barlow, 1957; Lu & Dosher, 1999, 2008; Pelli, 1990). Examples of external noise include random assignment of contrast increments or decrements to individual pixels in a visual stimulus, samples of “white noise” added to an auditory stimulus, or any other random trial-by-trial perturbations to the stimulus. Multiple presentation methods that utilize external noise assume that the total noise degrading subject performance is a composite of component noise sources. The first component, with *SD* σ_{ext}, reflects a variability in the subject’s internal representation of the external noise that is entirely correlated with the variability in the physical stimuli. This assumption implies that identical samples of external noise lead to internal representations that are partly composed of identical offsets along the decision axis. Therefore, a given sample offset reflected by this *consistent* noise component depends entirely on the specific noisy stimulus that evoked it.[1] The second component, with *SD* σ_{E_h}, signifies the internal noise induced during trials of stimulus class *h* and reflects random perturbations arising from the encoding of both signal (if present) and external noise in trial stimuli. Finally, random trial-by-trial sampling of a variable criterion with *SD* σ_C constitutes a third component. The distributional parameters of the encoding noise component may be functionally related to features of a stimulus class (e.g., contrast level), but it is still stochastic in nature and results in random perturbations of the internal representation to identical stimuli. The criterion variability, by assumption, neither depends on individual stimulus samples nor on the general stimulus class. We refer to these secondary noise components as *random* noise (Levi & Klein, 2003) insofar as they operate independently of any external noise samples (drawn from a single distribution). Therefore, the total response variability σ_{T_h} during trial presentations of stimulus S_h, is the combined result of the perturbations arising from consistent and random noise components.

$$\sigma^2_{T_h} = \sigma^2_{ext} + \sigma^2_{E_h} + \sigma^2_C \qquad (1)$$

In a multipass paradigm, subjects perform a signal detection task over multiple passes of trials. Each trial from the first pass includes an independent sample of external noise. However, subsequent passes of trials contain the same stimuli and exactly identical samples of external noise as in the first pass (see Figure 3). Although two passes suffice to obtain an estimate of agreement, in practice experiments often include additional passes for better accuracy and precision. Because any change in overt response to identical presentations of a stimulus reflects a change in the internal state of the observer, variability in response to identical stimuli reflects internal noise (Burgess & Colborne, 1988; Green, 1964; Lu & Dosher, 2008). Researchers can assess to what extent subject responses agree over multiple presentations of identical samples of noisy stimuli and this agreement can be used as an additional constraint to determine the ratio $(\sigma^2_{E_h} + \sigma^2_C)^{1/2}/\sigma_{ext}$ (see Appendix A). Low ratios of internal to external noise will lead to greater agreement between responses to identical stimuli, while higher ratios lead to a decline in agreement. The estimated statistic of agreement depends on the task specifications but can be measured with percent agreement (Burgess & Colborne, 1988; Lu & Dosher, 2008; Spiegel & Green, 1981), correlation (Levi & Klein, 2003), or covariance between responses to corresponding trials on successive passes.

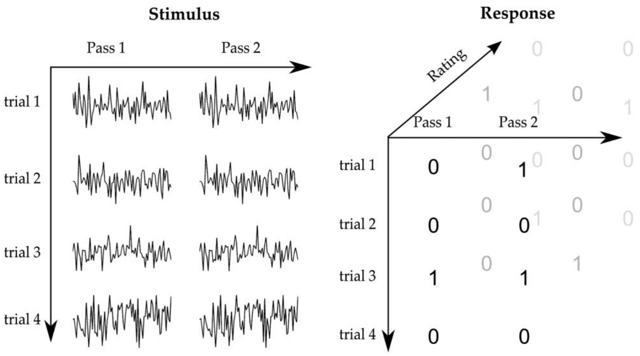


Figure 3. Left: a multipass procedure contains at least two runs with identical samples of external noise added to corresponding trial stimuli within each pass. Corresponding trials need not be presented according to the same stimulus schedule for each pass, but we match external noise samples with trial order here for the purpose of illustration. Right: Measures of agreement (percent agreement, covariance, and correlation) between responses to corresponding trials across passes provide additional behavioral measure to help constrain observer models.

For multipass experiments involving only a single decision criterion, the observed response frequency and response agreement can provide estimates of the total internal to external noise ratio in addition to sensitivity and response bias (Burgess & Colborne, 1988; Green, 1964). The separate parameters of criterion and encoding variance, however, leaves many possible combinations of criterion and encoding noise that are compatible with the measured combination of HR, FAR, and agreement measures. In a multipass signal detection experiment with a single criterion, there are five parameters to estimate (encoding noise for each stimulus class, a mean value for the signal distribution, a criterion mean, and a criterion variance) with only four data points (HR, FAR, agreement on signal present trials, and agreement on signal absent trials).

Degrees of freedom increase with additional criteria in a rating experiment. Rosner and Kochanski (2009) demonstrated the possibility of independent estimates of criteria variability, criteria positioning, stimulus positioning, and stimulus representational noise (they did not distinguish between consistent and random components) in rating tasks with at least three stimulus levels and four response categories. Estimating these parameters with only two stimulus classes, however, requires additional constraining data measurements. In this article, we use a multipass confidence rating procedure (MCR) and we measure the covariance of responses to trials of a specific stimulus class across different passes as an index of response correlation between these passes. The full covariance matrix provides a compact summary of agreement measures for the same categorization of identical trials across passes (within-category covariance along the diagonal) as well as disagreement for different categorizations of identical trials (between category covariance off the diagonal). Conceptually, if trial-by-trial responses over each pass are taken as vector elements, then the covariance gives the (mean adjusted) dot product of these response vectors. A highly positive covariance estimate implies response agreement across passes. Very low covariance (near zero) implies lack of agreement. Highly negative covariance implies not only lack of agreement but strong disagreement across passes. With low to moderate levels of internal noise, we intuitively expect positive covariance values for within-category estimates along the diagonal of the covariance matrix. For between-category covariance estimates for adjacent regions of decision space (e.g., response assignments of "2" and "3" across passes) we might expect lower though still positive values. For between-category covariance estimates for response assignments of nonadjacent regions (e.g., response assignments of "2" and "5" across passes), we expect nearly zero or negative covariance estimates.

[1] To our knowledge, filtered or bandpass noise has not generally been used with the mutlipass paradigm. However, color or frequency spectrum notwithstanding, we see no difference in the principle assumption that trial sampled internal noise is comprised of stimulus dependent (consistent) and stimulus independent (random) components.

Here we show that the MCR procedure sufficiently constrains a class of decision noise models to identify all relevant parameters even when the task involves only two stimulus classes. Under the MCR procedure, each stimulus class gives us M independent response frequencies as well as M independent agreement measures for identical responses between passes. In addition to the covariance of responses for the same rating category across passes (within-category covariance: e.g., response category "2" in the first pass and "2" again in subsequent passes), the covariance of responses for different rating categories across passes may provide even stronger constraints for model fits to data (between-category covariance: e.g., response category "2" in the first pass and "3" in subsequent passes). In total, the MCR provides $M(M + 3)$ data points ($2M$ response frequencies and $M(M + 1)$ covariance estimates) to fit $2M + 3$ free parameters: M criterion positions, M criterion variances, an encoding variance for the signal absent trials, an encoding variance for the signal present trials, and the mean position of the signal stimulus along the decision axis (see Table 1). Therefore, the MCR procedure may provide sufficient constraints to recover all decision noise parameters for a rating task with as few as three response categories (corresponding to $M = 2$).

To illustrate this point, Figure 4 (left) shows two overlapping and nearly identical ROCs generated using very different underlying internal noise components. In one case, the encoding noise is equal for signal-absent and signal-present trials while decision noise is small for all criteria. In the second case, the encoding noise for signal-present trials is half that for signal absent trials, while the decision noise varies markedly across criteria and even well exceeds the encoding noise at one of the decision boundaries. However, despite these very different noise profiles, the resulting ROC's are essentially the same. On the other hand, the covariance measures estimated from an MCR procedure are drastically different (Figure 4, right) and may provide additional constraints to disambiguate the underlying noise components. While a greater number of independent data points relative to the number of free parameters provides a necessary condition for fitting those parameters within the context of a model, this is not sufficient all on its own (Busemeyer & Diederich, 2010). Even with more data points relative to free parameters, the data may fail to fully constrain the model and disambiguate the parameters, so that successful model identification depends on more than degrees of freedom alone.

We will provide evidence that the MCR framework allows for full parameter recovery from simulated data over a wide range of conditions. However, we first seek an intuitive demonstration of the relationship between observed data and underlying noise components. While some changes to covariance data are straightforward (e.g., representational noise for a specific stimulus class selectively depresses covariance estimates for responses to that specific stimulus class, but nontrivial decision noise at even a single criterion boundary will lead to changes in covariance and z-scores at all criteria owing to positional overlap), the pattern of expected values becomes more complex with the introduction of decision noise. In Figure 5, we examined changes to expected values of response frequencies and covariance structure for a three-category rating task in which we selectively increase the variability for one of the criteria from zero to match the level of variability in the stimulus representation. For this very simple example, we assumed that observers map internal representations to responses according to a corrected law of categorical judgment as described by Rosner and Kochanski (2009; see Decision Rules below). This decision rule determines response assignment by subtracting each trial-sampled representation from trial-sampled criteria and choosing the category where the difference between representation and corresponding criterion gives the least positive value; when all values are negative, the representation is assigned to the highest response category.

We begin from the standard SDT account with no decision noise. In this case we assume that two static criteria, each positioned at the mean of the signal-absent and signal present distributions, divide the decision space into three response categories (Figure 5, top-left). Our example assumes a $d' = 1$ with equal representational noise for the two evidence distributions. In contrast, we juxtapose a second scenario in which we selectively increase the decision noise for the more lax criteria to match the representational noise, without modifying any of the other parameters. The joint distributions accounting for both the variability in the criterion as well as variability in the signal-absent and signal-present representations are shown as concentric circles (Figure 5, left middle and bottom). The vertical axis represents positions of the noisy criterion, the horizontal axis reflects positions of the noisy internal representations, and the solid blue lines reflect the position of the means of the noisy and static criteria with respect to the noisy criterion (horizontal blue lines) and representational (vertical blue lines) distributions. Finally, we superimpose rating response column and row labels A, B, C, and D for regions of the joint distributions according to the decision rule described above. For example, when trial samples of both the noisy criterion and representation exceed the stricter (and static) criterion in region DD, some trial representations will be classified as "1"s instead of "3" depending on whether the sampled criterion exceeds the sampled representation. Similarly, trial representations will always be classified with a response category of "2" anytime a sampled criterion exceeds the static criterion while the sampled representation does not (regions AD, BD, and CD). Each column of these joint distributions illustrates how some representations falling along the decision axis become reassigned depending on the position of the trial sampled criterion. In column C, for example, all representations remain with a response assignment of "2" except in row C where some will be reassigned to a response of "1."

Figure 5 (right) also shows the corresponding changes to the zROC and covariance in the classical SDT treatment with no decision noise (shown as circles) and with the targeted increase in decision noise at the most lax criteria (shown as "×" symbols). In the case of the zROC plot, we can see how the introduction of decision noise at the more lax criterion results in small but noticeable change in position for the stricter criterion in z-space. Column D in the joint distributions shows that response assignments of "3" can only decrease with increased decision noise at the more lax

Table 1
Degrees of Freedom in Rating Procedure Tasks

Procedure	Data points		Free parameters
Rating	$2M$	$<$	$2M + 3$
MCR	$2 \times 2M$	$>$	$2M + 3$

Note. MCR = multipass confidence rating.

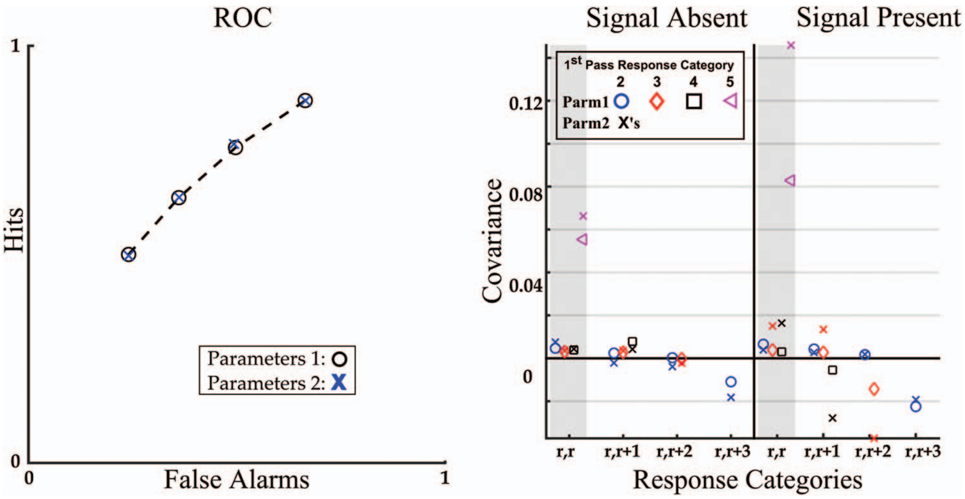


Figure 4. Left: Two overlapping receiver operating characteristics generated using a decision rule described by Rosner and Kochanski (2009; see decision rules below) and assuming two different underlying parameter sets. Parameters 1 (open symbols): encoding noise is 1 for both signal absent and signal present trials; mean of the signal distribution is 1; criteria are located at −0.62, 0, 0.5, 1 with criterion noise at 0.1 for all criteria. Parameters 2 (Xs): encoding noise is 0.8 for signal absent trials, 0.4 for signal present trials; signal mean is 0.92; criteria are located at −0.15, 0, 0.5, 0.77 with corresponding criteria noise of 0.125, 1, 0.3, 0.2. All quantities given in units of the consistent noise, σ_{ext}. Right: covariance outcomes using the same two underlying parameter sets result in discriminably different data patterns. Within-category covariances are denoted as [r,r] and lie within the gray bar. Between-category covariances lie outside the gray bar. Circles mark within- and between-category covariances for response "2"; diamonds for response "3"; squares for response "4"; and triangles shows within-category covariance for response "5." For example, between-category covariance for response categories "3" and "5" across passes are shown as a diamond and an X at the position "r, r + 2" along the abscissa. See the online article for the color version of this figure.

criterion, and no responses previously mapped to "1" or "2" will be reassigned to "3" according to the parameters we have chosen for this illustration. This net loss of assignments to "3" occurs for both signal-absent and signal-present trials and is reflected by a shift in the criterion estimate in the zROC toward the bottom left. Similarly, columns A and B show how the criterion variability on signal-absent trials results in a net decrease of response assignments mapped to "1" leading to a significant rightward shift in the more lax criterion estimate in zROC space: losses from region BB are canceled by gains in region CC, but region AA, BA, AD, and BD all lose response assignments of "1" without corresponding counterbalancing regions. These regional reassignments are also true for signal-present trials, but in this case the region CC represents a much higher likelihood under the joint density function than is counterbalanced by regions AA, BA, AD, and BD. These regional exchanges, coupled with an additional increase in "1" responses from region DD to counterbalance losses in region BB, result in a very slight net increase in response assignments of "1" with a corresponding subtle downward shift in the position of the more lax criterion in the zROC plot.

We can also observe this increased decision noise changes the covariance data, though overall response frequency will also affect this measure in addition to the correlation in responses across passes. For both signal-absent and signal-present trials, the covariances for response assignments of "3" decrease because of changes in lower correlations and lower response frequencies when trial samples of both criterion and representation fall within region DD. Within-category covariance for response assignments of "2" also decrease with increased decision noise for signal-absent trials since many of the regions previously assigned to "1" become remapped to "2" under the joint distribution. Although the remapping of these regions also occurs during signal-present trials, covariance for response assignments of "2" nets a small increases here because the overall response frequency increases with decision noise, but the shifted position of the signal-present joint distribution leads to a lower drop in correlation than occurs in signal-absent trials (note the lower impact of regions AD, BA, BB, and BB). On the other hand, the between-category covariance of responses "2" and "3" become increasingly negative on both signal-absent and signal-present trials. These negative covariances occur because response assignments of both "2" and "3" become increasingly associated with "1" on subsequent passes, thereby decreasing the "2–3" covariance from baseline.

Decision Rules

For any task amenable to analysis within the signal detection framework, SDT assumes observers generate responses by comparing internal representations of the trial stimulus with one or more decision criterion. A decision rule constitutes a specific protocol that determines how an observer assigns an internal representation to a response. With static criteria, most straightforward decision rules predict identical responses for any given trial-sampled representation. With noisy criteria, the situation may

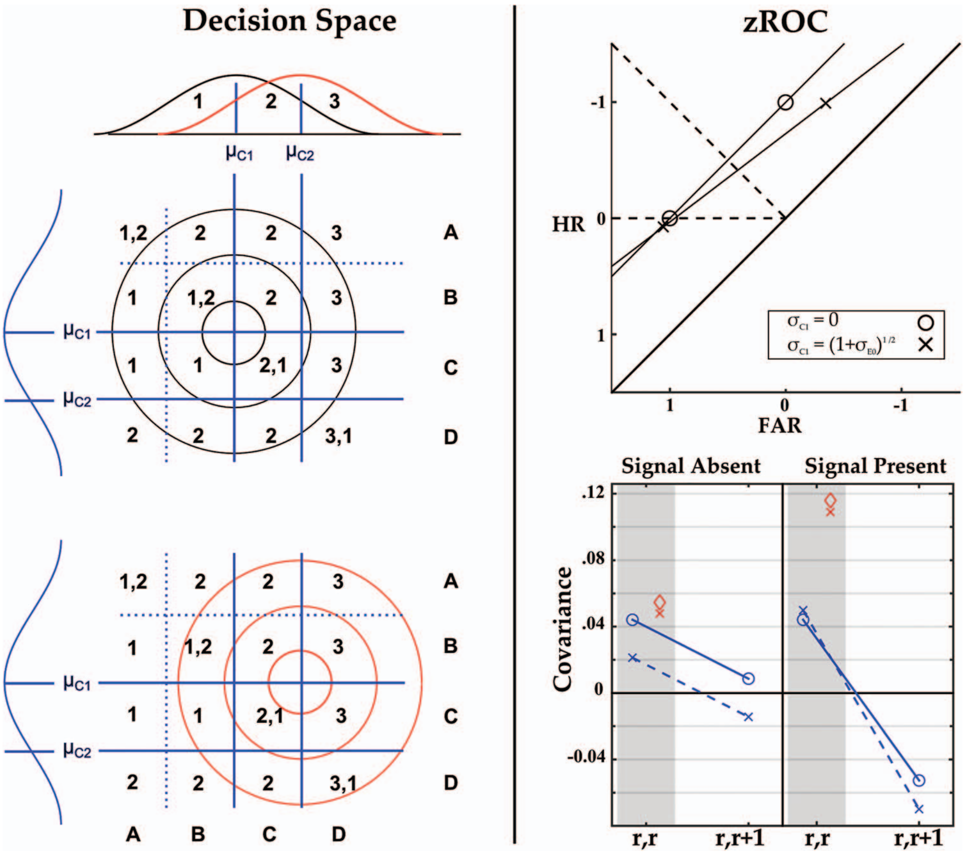


Figure 5. Left-top: decision space for classical confidence rating signal detection task with no decision noise. Criterion locations lie at the means of the signal-absent and signal-present distributions. Left-center and bottom: decision space showing joint distributions when decision noise equal to the representational noise is selectively added to the more lax criterion. The center of the concentric circles represents the mean position of the lax criterion along the ordinate, and the mean position of the signal-absent distribution (center) and signal-present distribution (bottom) along the abscissa. Straight vertical and horizontal lines represent mean criterion positions. Numbers overlaying joint distributions denote expected response categories for trial-sampled criteria and representations falling in these regions. Right: z-transformed receiver operating characteristic (zROC; top) and covariance data (bottom) for classical signal detection task without decision noise (circles and diamonds) and with decision noise equal to representational noise at the more lax criterion (Xs). Within-category covariance data lie within the gray bar, between-category covariance data lie outside the gray bar. Covariance data indicating a response of "2" in at least one pass are shown as circles (no decision noise) and vertically aligned Xs (a noisy lax criterion); within-category covariance for response "3" in both passes labeled as diamonds (no decision noise) and vertically aligned Xs (a noisy lax criterion). HR = hit rate; FAR = false alarm rate. See main text for more details. See the online article for the color version of this figure.

be quite complex. When the task involves only a single noisy criterion (yes/no, 2AFC, 2IFC with bias, etc.), no ambiguity arises in consideration of this comparison. Similarly, for tasks calling for multiple criteria (rating procedures, identification, classification, etc.), it is straightforward to map a trial-sampled representation to response as long as the noisy criteria do not overlap from trial to trial. We might even expect the operation of an enforcement mechanism maintaining ordinal relations between trial-sampled criteria (Treisman & Faulkner, 1984).

When noisy criteria have overlapping distributions, trial-sampled criteria may sometimes become disordered along the axis, requiring subjects to implement a more complicated decision rule. Simultaneous decision rules require the observers to compare the internal representation with available criteria all at once. These decision rules then determine a response category by making a unique selection among the results of these comparisons. The work in this article focuses on several forms of simultaneous decision rules.

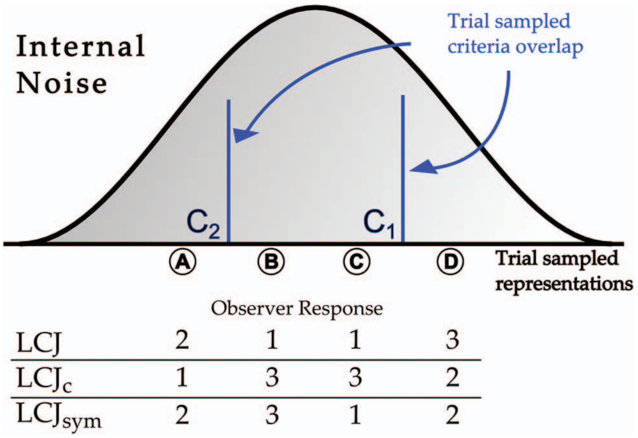
We first formulate the simultaneous decision rule used by Rosner and Kochanski: *subtract the position of the stimulus representation from each criterion boundary and respond with the category affording the least positive distance; if all differences are negative respond with category $M + 1$.* Following a similar notation used by Rosner and Kochanski, let $s_h \in G(0, 1)$ where $G(\mu, \sigma)$ is a Gaussian random variable with mean μ and variance σ^2. Then $s_h\sigma_{E_h}$ equals the random offset of the internal response from its mean position μ_{S_h} because of the subject's encoding noise during a trial of stimulus class S_h. Furthermore, let $c_i \in G(0, 1)$ and $c_i\sigma_{C_i}$ equal a trial-sampled offset of the ith criterion from its mean location μ_{C_i} because of the subject's internal decision noise at that boundary. We now assume a single external noise level $\sigma_{\text{ext}} = 1$, so that all parameters are estimated in reference to this term. We let s_{ext} equal an observer's *consistent* trial-by-trial offset to the internal representation because of presentation of a specific sample of Gaussian external noise, so that $s_{ext} \in G(0, 1)$.

The Rosner and Kochanski decision rule just described can be formalized as follows: for a trial-sampled stimulus of class h choose the category m when the following equation evaluates to true, or category $M + 1$ if the equation evaluates false for all m:

$$s_{ext} + s_h\sigma_{E_h} + \mu_{S_h} < c_m\sigma_{C_m} + \mu_{C_m} < \min_{\forall m' \neq m}\Big[c_{m'}\sigma_{C_{m'}} + \mu_{C_{m'}} \,|\, s_{ext} + s_h\sigma_{E_h} + \mu_{S_h} < c_{m'}\sigma_{C_{m'}} + \mu_{C_{m'}}\Big] \quad (2)$$

Klauer and Kellen (2012) proposed two alternative simultaneous decision rules. In the first of these alternatives, the decision rule determines the trial-by-trial response according to the rule: *subtract the* m *criterion boundaries from the trial-sampled stimulus representation and respond with the category m + 1 yielding the smallest positive distance; in the event all comparisons are negative, choose category 1*. The second rule determines the trial-by-trial response by computing the least absolute distance between criterion boundaries and the trial-sampled representation. Specifically, *subtract the stimulus representation from all M criterion boundaries, identifying the smallest absolute value of the difference between stimulus representation and criterion boundary m, and choose category m if the difference is positive and m + 1 otherwise*. This second rule also has the additional consequence that rating frequencies will be symmetrically distributed when the corresponding means of criteria distributions are symmetrically distributed about an evidence distribution. Given any $M > 1$ trial sampled criteria, these decision rules can be used to map any trial sampled internal representation to overt observer responses.

To distinguish these three decision rules, we follow Kellen et al. (2012) and denote Rosner and Kochanski's law of categorical judgment as LCJ (given by Equation 2); we denote the second (Klauer and Kellen's complimentary version of the LCJ) as LCJ$_\text{c}$, and the last as LCJ$_\text{sym}$ because of its symmetric treatment of criterial boundaries relative to trial sampled representations. Figure 6 contrasts the response mappings for each of these three decision rules when trial-sampled criteria overlap. For a given sample of criteria, the rules prescribe different response profiles for stimuli falling in a given region along the decision axis. Note that for any given overlapping criteria the LCJ and LCJ$_\text{c}$ prescribe entirely incongruent responses while LCJ$_\text{sym}$ shows some response agreement with both. These differences suggest the possibility that the LCJ will produce distinctly different data patterns in the aggregate from the LCJ$_\text{c}$ rule and moderately different patterns from the LCJ$_\text{sym}$ rule. With these three different decision rules in hand, we examined the possibility of parameter recovery in simulated MCR experiments using simultaneous decision rules that either matched or mismatched the rule used to generate simulated data.



Observer Response	A	B	C	D
LCJ	2	1	1	3
LCJ$_\text{c}$	1	3	3	2
LCJ$_\text{sym}$	2	3	1	2

Figure 6. Criterion overlap and stimulus-response mapping for three different decision rules. Random trial-by-trial sampling may lead to ordinal rearrangement of criteria (C1 and C2). The encircled letters A, B, C, and D denote different positions of trial sampled stimulus representations falling along the decision axis. An observer requires an explicit decision rule to map the internal representation to a response. Under each stimulus representation, the columns of the Observer Response shows how an observer operating under the LCJ, LCJ$_\text{c}$, and LCJ$_\text{sym}$ decision rules classifies each stimulus representation above. See main text for response mapping protocols. LCJ = law of categorical judgment; c = complimentary; sym = symmetrically adjusted. See the online article for the color version of this figure.

Simulation Study

In the present study, we recruit the power of external noise and the MCR method in a confidence rating task to disambiguate and estimate criterion noise under the various simultaneous decision rules LCJ, LCJ$_\text{c}$, and LCJ$_\text{sym}$. We derived the expected values of the response frequencies and covariance data conditioned on trial-by-trial samples of external noise. Here in the main text we show the equations describing LCJ. For a formal description of LCJ$_\text{c}$ and LCJ$_\text{sym}$, please see Appendix A.

For the LCJ decision rule, the expected response frequencies conditioned on the external noise sample s_{ext} for the hth stimulus class are given as,

$$P(R = m \,|\, s_{ext}, S_h) = \int \phi(c_m;\, \mu_{C_m}, \sigma_{C_m}) \int_{-\infty}^{\mu_{C_m}+c_m\sigma_{C_m}} \phi(s_{E_h}; s_{ext} + \mu_{S_h}, \sigma_{E_s}) \times \prod_{m' \neq m}\left[1 - \int_{\mu_{S_h}+s_{E_h}\sigma_{E_h}+s_{ext}}^{\mu_{c_m}+c_m\sigma_{C_m}} \phi(c_{m'}; \mu_{C_{m'}}, \sigma_{C_{m'}})dc_{m'}\right] ds_h dc_m \quad (3)$$

where $\varphi(x)$ is the Gaussian probability density function. We then easily determine $P(R = M + 1 \,|\, s_{ext}, S_h)$ as $1 - \sum_{m=1}^{M} P(R = m \,|\, s_{ext}, S_h)$. The first term in Equation 3 integrates over all possible values of the mth criterion. The middle term integrates over stimulus representation values up to that criterion. The third term estimates the probability that the response is consistent with any other criterion. We then integrate over all external noise samples s_{ext} to get the overall response frequency for this stimulus class h.

$$P(R = m \mid S_h) = \int P(R = m \mid s_{ext}, S_h)\phi(s_{ext})ds_{ext} \qquad (4)$$

Similarly, across any two passes i and j the covariance between any two response categories m and m' is,

$$Cov[R_i = m, R_j = m' \mid S_h] = \int P(R_i = m \mid s_{ext}, S_h) \times P(R_j = m' \mid s_{ext}, S_h)\phi(s_{ext})ds_{ext} - P(R_i = m \mid S_h)P(R_j = m' \mid S_h) \qquad (5)$$

We now show that data from the MCR experiment adequately constrains the models to uniquely identify individual representational and decision noise components. We approach this problem by examining the precision, accuracy, and goodness-of-fit of recovered model parameters from simulated data. For each decision rule adopted by our simulated observer we tested parameter recovery when fitting simulation data with matched models (e.g., LCJ fitted to data generated with a simulated observer using LCJ) as well as when fitted with mismatched models (LCJ$_\text{c}$ and LCJ$_\text{sym}$ fitted to data generated with simulated observer using LCJ). In the multipass framework, response frequencies and the covariances of responses across passes are estimated. This covariance data paired with the rating response sufficiently specifies the models for independent identification of encoding and decision noise contributions.

Method

Rationale. To demonstrate full parameter recovery for the model using our new framework, we simulated a number of MCR experiments under a range of noise configurations. Because MCR experiments schedule identical stimuli over each pass, data collection may require significant empirical investment. As the minimal data for acceptable model recovery was of interest, we examined not only the possibility but also the feasibility of parameter recovery at different numbers of trials and passes per simulated experiment.

Our simulations investigated several plausible configurations for the parameters of criterion and stimulus distributions using three response categories and two stimulus classes. We focus on the minimum number of stimuli and rating categories because earlier efforts toward parameter recovery became problematic with fewer response categories. We investigated configurations in which either the criterion noise variances or the encoding noise variances were equated along the decision axis (labeled *equ*), increased along the decision axis (labeled *asc*) or decreased along the decision axis (labeled *des*). We assume a single external noise variance of unity for all stimulus classes, with an external noise mean of zero. For any given variance configuration, $0 \leq max[\sigma^2_{E_0}, \sigma^2_{E_1}] \leq 1$ and $0 \leq max[\sigma^2_{C_1}, \sigma^2_{C_2}, \ldots, \sigma^2_{C_M}] \leq 1$. We also normalized the sum of the highest decision and encoding noise variances to equal the variance of the external noise. In other words, $max[\sigma^2_{E_0}, \sigma^2_{E_1}] + max[\sigma^2_{C_1}, \sigma^2_{C_2}, \ldots, \sigma^2_{C_M}] = \sigma^2_{ext}$. This constraint accords with the reports of previous authors that the total internal noise lies near this level for visual and auditory detection and discrimination experiments over a considerable range of external noise levels[2] (Burgess & Colborne, 1988; Green, 1964; Lu & Dosher, 2008). For all other noise components, we computed variances by applying logarithmic decrements in the *ascending* and *descending* conditions. We positioned each criterion mean along the decision axis at $\frac{1}{3}(\sigma^2_{ext} + \sigma^2_{E_0})^{1/2}$ and $\frac{2}{3}(\sigma^2_{ext} + \sigma^2_{E_0})^{1/2}$ so that we could ensure a robust level of trial-by-trial criterion overlap. Finally, we kept the position of the mean of the signal distribution at $(\sigma^2_{ext} + \sigma^2_{E_0})^{1/2}$. The various arrangements of parameter configurations is shown in Table 2 and Figure 7.

The simulated experiments emulated a confidence rating detection paradigm in which an observer maintains two criteria that define three response categories. The simulated observer implemented a LCJ decision rule for all noise level configurations. We also generated simulated data with the LCJ$_\text{c}$ and LCJ$_\text{sym}$ decision rules for a single parameter configuration in which decision and encoding noise are equal across criterion boundaries and stimulus classes. The probability of a signal present stimulus was 0.5. The simulated experiments varied the number of trials per pass and number of passes per experiment, in addition to a specific parameter configuration. The number of trials n per pass was 250, 500, or 1,000 and the number of passes was either four or six. We set the minimum number of passes to four to obtain variance estimates on covariance data for weighted-least squares model fitting.

Data analysis. The data were arranged in this way: for each stimulus class h, we have $M + 1$ subject response matrices $\boldsymbol{R}^{(m,h)}$ of size $T \times J$, where J is the number of passes, T is the number of trials per pass, and m is an available response category. Then each entry of $\boldsymbol{R}^{(m,h)}$ contains 1's for trial responses to stimulus class h classified as category m and 0's otherwise. Thus, we denote $r_j^{(m,h)}$ as the jth $T \times 1$ column vector of the matrix $R^{(m,h)}$ with the tth entry $r_{tj}^{(m,h)}$ equal to 1 or 0, signifying whether or not subjects classified the stimulus from the tth trial of the jth pass with a classification of m. The matrix corresponding to the lowest confidence rating $R^{(m=1,h)}$ was dropped because of its redundancy given the other response rates and fixed trial numbers.

For every simulated experiment, we computed the relative frequency of the mth classification rating during each pass j as

$$\hat{p}_j(r = m \mid S_h) = \frac{1}{T}\sum_{t=1}^{T} r_{tj}^{(m,h)} \qquad (6)$$

The average of each response rating across all passes is the best and final estimate of the rating response rate. That is

$$\hat{p}(r = m \mid S_h) = \frac{1}{J}\sum_{j=1}^{J} \hat{p}_j(r = m \mid S_h) \qquad (7)$$

Covariance was computed for every combination of passes for every rating category. For passes i and j, where $i \neq j$, and category ratings m and m', the covariance is given as,

$$Cov\left[r_i^{(m,h)}, r_j^{(m',h)}\right] = \frac{1}{T-1}\sum_{t=1}^{T}\left[r_{ti}^{(m,h)} - \hat{p}_i(r = m \mid S_h)\right] \times \left[r_{tj}^{(m',h)} - \hat{p}_j(r = m' \mid S_h)\right] \qquad (8)$$

[2] The dependence of internal noise on external noise is predicted from observer models that show internal noise increases with the total energy of the stimulus (Lu & Dosher, 2008).

Table 2
Parameter Configurations for Simulation Study

	Encoding noise	
Decision noise	Equal	Ascending
0		✓
Equal	✓	✓
Ascending	✓	✓
Descending		✓

We refer to the covariance as *within category* covariance when $m = m'$ and *between category* covariance when $m \neq m'$. For an MCR experiment with J passes, we have $\sum_{j=1}^{J-1} j$ observations of *within category* covariance estimates for each response rating m, and $2\sum_{j=1}^{J-1} j$ observations of *between category* covariance estimates for each response pairing of m and m'. We took the average of all pairwise estimates as our final covariance estimate between categories m and m'.

Weighted least-squares model estimation requires estimates of the variance for each of the final response rates. The variability of the response rates for each pass was estimated by the variance of each response rate across all passes:

$$Var[p_j(r = m \mid S_h) = \frac{1}{J-1}\sum_{j=1}^{J} [\hat{p}_j(r = m \mid S_h) - \hat{p}(r = m \mid S_h)]^2 \tag{9}$$

The final estimate of each response rate is the average of the response rates across passes, and the final estimate of variance for an averaged response rate across all passes is given by dividing the variance among individual passes by the total number of passes. That is,

$$Var[p(r = m \mid S_h)] = \frac{Var[p_j(r = m \mid S_h)}{J} \tag{10}$$

Variances for covariance data were computed by first taking the variance of each within and between pass estimate and then dividing by the $\sum_{j=1}^{J-1} j$ or $2\sum_{j=1}^{J-1} j$ possible pairing combinations, respectively.

Modeling. We fit the LCJ, LCJ$_c$, and LCJ$_{sym}$ to simulated data derived from each parameter configuration and LCJ decision rule, and to simulated data derived from one parameter configuration using the LCJ$_c$ and LCJ$_{sym}$ decision rules. Model fits used a Matlab simplex optimization routine (Nelder-Mead) and a weighted least-squares cost function. The cost function heavily penalized a possible solution if any variance parameters fell below zero or if the criterion means violated their ordinal relation. At the beginning of each parameter search routine, we generated initial starting parameters by independently perturbing the true means of each parameter using a Gaussian random number generator with a *SD* of 0.15σ_{ext}. Apart from penalties just stated, the constraints imposed on parameters of the simulated observer were not imposed upon the model during parameter recovery: candidate fits of criteria and signal distribution means were not restricted to specific positions along the decision axis nor were they restricted to maintain certain relative distances; nor were any decision and encoding noise variances constrained to sum to unity. We ran 250 experiments at each experimental condition and at each parameter configuration.

Results

We computed the median and 95% confidence interval (CI) for each model parameter using the 250 simulated runs at each parameter configuration and pass-trial combination. In every case, the actual parameter values of the simulated observer fell within

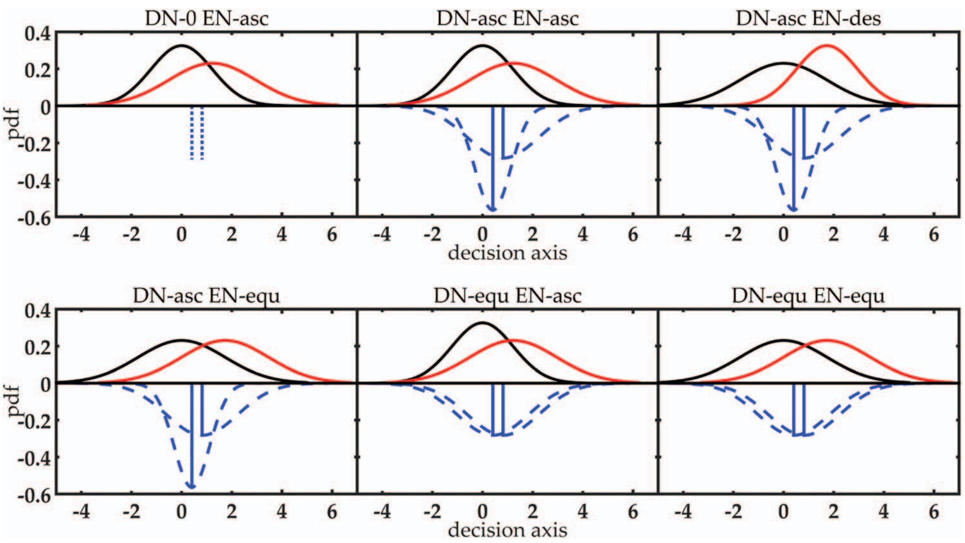


Figure 7. Probability density functions for six representative parameter configurations underlying response behavior for simulated observers. Density functions represent signal-absent trials (mean zero), signal-present trials (with greater mean values), and criterion noise (reflected downward across the decision axis). DN = decision noise; EN = encoding noise; asc = increased; des = decreased; equ = equated. See the online article for the color version of this figure.

the 95% CIs of the estimated values for each position and variance parameter. The median parameter values recovered from the matched model were very close to the parameter values used to generate the simulated data. These results stand in contrast to the attempted parameter recovery for decision protocols of the models mismatched against decision rule of the observer. In the case of $\mathrm{LCJ_c}$ fitted to the data simulated with LCJ, at least one generative parameter failed to fall within the 95% CI when simulations were run with four passes at 500 trials/pass or with six passes at 250 trials/pass. When we fitted $\mathrm{LCJ_{sym}}$ to the data simulated with LCJ, at least one generative parameter failed to fall within the 95% CIs when simulations were run with four passes at 500 trials/pass.

We also examined the precision and accuracy of our model fits as a function of trials per pass and passes per experiment. We calculated the standard error (*SE*) of individual recovered parameters by computing the *SD* of each fitted parameter across all experiments within a given noise configuration, trials/pass, and passes/experiment setting. Similarly, we estimated an individual parameter mean-squared error (*MSE*) by squaring the difference between the true parameter value adopted by the simulated observer from the corresponding fitted parameter in each experiment and averaging across all experiments within the given configuration, trials/pass, and passes/experiment setting. Mean *SE*s (averaged across all model parameters), as well as the *SE* of the most variable parameter, strictly decrease with increasing trials per pass and passes per experiment at each experimental configuration (see Figure 8). Mean MSEs (again, averaged across all model parameters) also exhibit a pattern of increasing accuracy (decreasing *MSE*) with greater numbers of trials and passes for the correctly matched decision rule (see Figure 9). The *MSE* of the most poorly fitted parameters (i.e., those parameters with the highest *MSE*) also decrease with increasing trials and increasing passes (a single exception occurs in the DN-*asc* EN-*des* configuration at 500 trials/pass comparing four vs. six passes per experiment).

We also examined fits at six passes/experiment for mismatched relative to matched models (see Figure 10). For both fits of $\mathrm{LCJ_c}$ and $\mathrm{LCJ_{sym}}$ to an observer using LCJ, the averages of the *MSE* for mismatched protocols do not generally monotonically decrease with trials/pass or passes/experiment. Furthermore, at six passes/experiment, fits for both mismatched models show a higher average *MSE* across all trials/experiment relative to *MSE* for the correctly matched model for all configurations except DN-0 EN-*asc*. The models perform equally well for simulations assuming zero decision noise because the models make identical predictions for negligible decision noise. For one parameter configuration, we used both $\mathrm{LCJ_c}$ and $\mathrm{LCJ_{sym}}$ as our simulation decision rule (Figure 10, bottom). Here too, accuracy improved for matched but not mismatched models with increasing trials.

An important concern is whether differences in parameter recovery between matched and mismatched models correspond to goodness-of-fit when actual underlying parameters are unknown. A weighted least squares estimate (χ^2) finds parameters that minimize the difference between simulated data and expected values of data based on recovered parameters. We computed χ^2 for each fit of matched and mismatched models to each simulated data set. We averaged across simulations from a given configuration and trials/pass setting using six passes/experiment from mismatched and correctly matched models. In this case, the average χ^2 fits for the correctly matched model remains nearly constant with increasing trials/experiment (see Figure 11). On the other hand, average χ^2 for mismatched models increases with increasing trials/experiment for all configurations except DN-0 EN-*asc*. In contrast to the other configurations, average χ^2 fits for DN-0 EN-*asc* are notably consistent across both matched and mismatched fits. For simulated

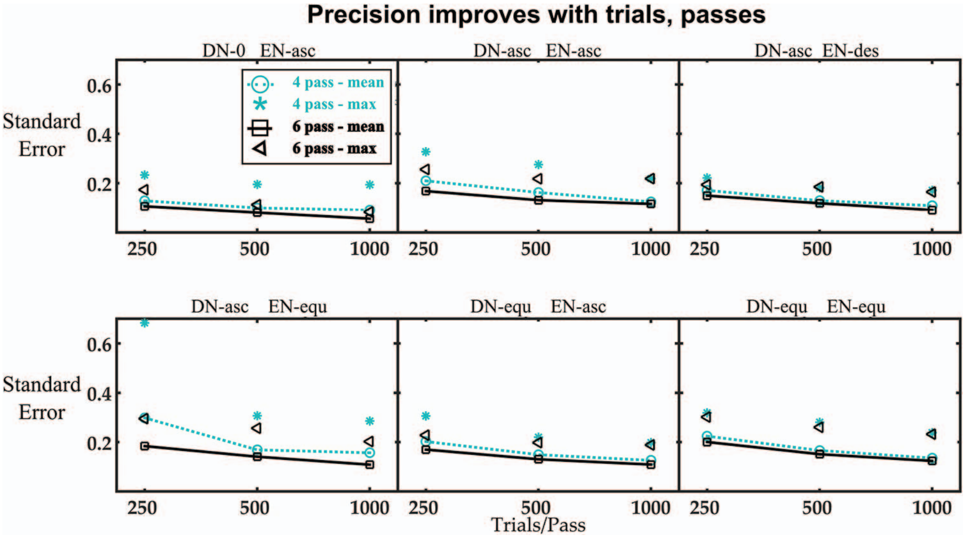


Figure 8. Standard error of parameter fits to data from simulated experiments for different pass-trial and parameter configurations. Average standard error across all parameters given for four passes/experiment (circles) or six passes/experiment (squares). Maximum standard error among parameters given by asterisks and triangles. All parameter configurations show less variability in parameter fits with increasing trials and passes. DN = decision noise; EN = encoding noise; asc = increased; des = decreased; equ = equated. See the online article for the color version of this figure.

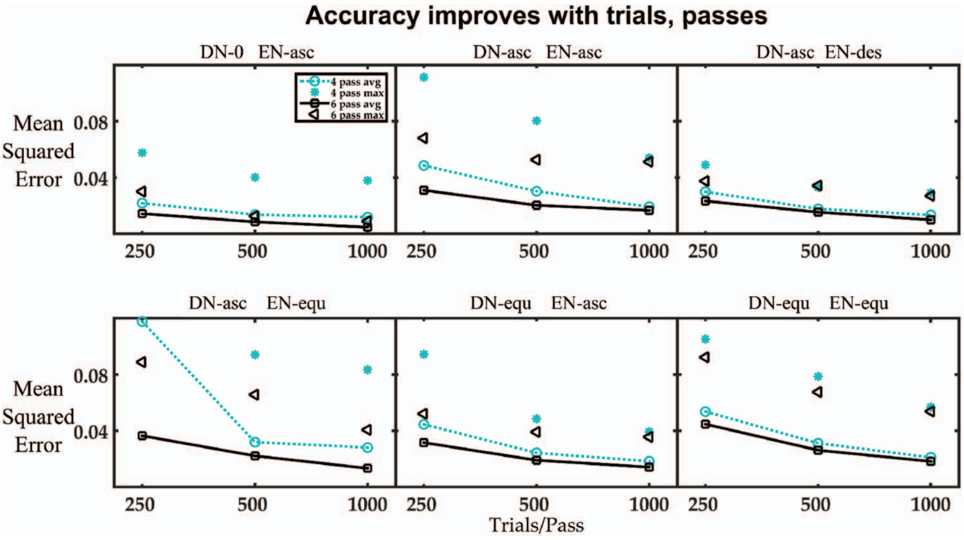


Figure 9. Average mean squared error of parameter fits to simulated data for various pass-trial and parameter configurations. Average mean squared error across all parameters given for four passes/experiment (circles) or six passes/experiment (squares). Maximum mean squared error among parameters given by asterisks and triangles. (Maximum for DN-asc EN-equ at 250 trials, four passes is 0.465; not shown to preserve scale). DN = decision noise; EN = encoding noise; asc = increased; des = decreased; equ = equated. See the online article for the color version of this figure.

observers with zero decision noise, fits show an increasing accuracy while the log of the mean χ^2 fits lie within a narrow range across all trials/experiment for all model protocols. We also investigated the frequency with which the model fits for correctly matched model resulted in lower weighted least square costs than fits for mismatched models. For every configuration except DN-0 EN-*asc*, χ^2 fits were lower for correctly matched models than mismatched models for at least 91% of the individual simulations with four passes and 250 trials/pass. This lower bound on success rate increased to 97% for individual simulations with six passes and 1,000 trials/pass.

We also examined *MSE* and χ^2 for model fits to data generated using the LCJ_c and LCJ_{sym} decision rules for a single parameter configuration, DN-*equ* EN-*equ* (Figure 11, bottom). Similar to results when using the LCJ as a generative model, *MSE* decreased with additional trials for correctly matched rules but did not generally show similar decreases with mismatched rules. Again, the χ^2 results for models matched to the generative model remained low with increasing trials, while the χ^2 increased with increasing trials for mismatched models. When using LCJ_c as the generative decision rule, χ^2 fits for correctly matched models were lower than mismatched models for at least 90% of the individual simulations with four passes and 250 trials/pass. This lower bound success rate increased to 99% of individual simulations with six passes and 1,000 trials/pass. However, when using the LCJ_{sym} as the generative decision rule, success rate decreased significantly for correctly matched models relative to mismatched models at 60% of individual simulations with four passes and 250 trials/pass, increasing to 80% with six passes and 1,000 trials/pass.

Discussion

Previous attempts to estimate decision noise in simple response signal detection type tasks with two stimulus classes have required strong simplifying assumptions about the various noise components. Here we demonstrate that an MCR procedure provides a sufficiently rich data set to effectively recover decision noise parameters in many representative parameter configurations without assuming specific relationships between noise components. More important, this framework uses a model that permits overlapping criterion distributions and a decision rule that deals with this possible overlap.

The results show that both the precision (1/*SE*) and the accuracy (1/*MSE*) of the parameters increase with the number of trials/pass and passes/experiment. Furthermore, model fitting is not only possible, but also feasible with a number of total trials amenable to typical experiments in psychophysical studies. For all parameter configurations, it appears that parameter recovery does no worse and often improves with total number of trials up to 2,000 total trials. However, within the range of 3,000 to 4,000 total trials, allocating less trials over more passes results in better average accuracy than a greater number of total trials distributed over less passes for some parameter configurations (cf., DN-asc EN-equ, and DN-0 EN-asc). Still, though the optimal allocation strategy may depend on the underlying parameter configuration, the accuracy generally appears to improve with total number of trials.

For the configuration assuming zero decision noise, our simulations showed that all three decision models gave accurate and precise fits to the data of simulated experiments. This result should come as no surprise because each of the protocols prescribes identical trial-by-trial responses to a trial-sampled representation when criteria remain static over the course of the experiment. However, the results for accuracy look quite different for mismatched model and simulation protocols for all configurations imposing nontrivial decision noise. In every configuration with decision noise the accuracy and χ^2 estimates are much worse relative to correctly matched model fits. In these cases, the accu-

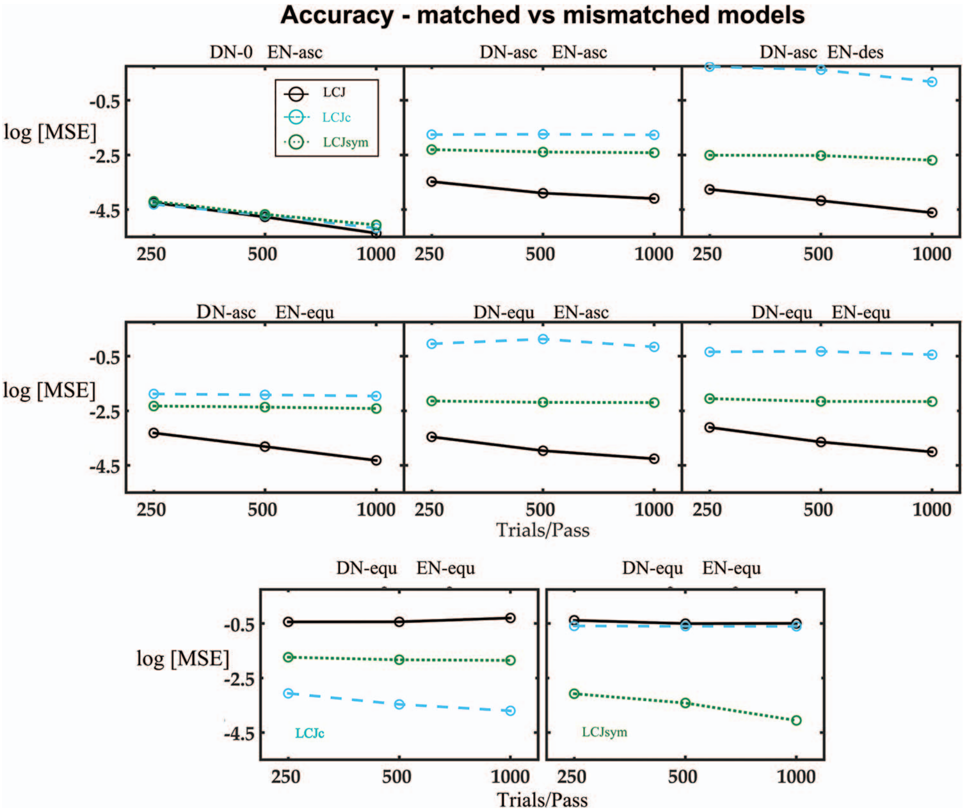


Figure 10. Top and middle rows: average log mean squared error (*MSE*) for model fits versus trials/pass (assuming six passes/experiment) for the LCJ, LCJ$_c$, and LCJ$_{sym}$ matched to data simulated using the LCJ decision rule. Bottom: average (*MSE*) for model fits to simulations when decision noise and encoding noise are equal across criteria and stimulus classes. Bottom left: LCJ, LCJ$_c$, and LCJ$_{sym}$ modeled to data simulated using the LCJ$_c$ decision rule. Bottom right: LCJ, LCJ$_c$, and LCJ$_{sym}$ modeled to data simulated using the LCJ$_{sym}$ decision rule. DN = decision noise; EN = encoding noise; asc = increased; des = decreased; equ = equated; LCJ = law of categorical judgment; c = complimentary; sym = symmetrically adjusted. See the online article for the color version of this figure.

racy generally fails to improve in any significant way with increasing trials/pass or passes/experiment and the χ^2 estimates become notably worse. The failure of these models to fit simulated data from mismatched protocols shows that the χ^2 estimates of recovered parameters for correctly matched pairings do not result from underconstrained models. It appears that some combinations of response frequencies and covariance data are simply not compatible with data sets generated by certain decision protocols. Therefore, fitting a decision rule model to data derived from an MCR experiment could recover erroneous estimates of the underlying parameters when the model rule fails to match the decision strategy of the observer. At least in some cases, however, mismatched models can be ruled out by comparison to fits of models more closely aligned with decision rules used by the observer. Some positive evidence exists suggesting that the experimenter may manipulate the observer's decision strategy by instruction and task structure (Treisman & Faulkner, 1985). However, a more parsimonious approach would attempt to disambiguate potential protocols through model selection techniques.

In a related study, we investigated the possibility of trade-offs between decision and encoding variance parameters. That is, for a given data set of response frequencies and covariance estimates, are variances associated with decision and encoding processes fungible? Using the LCJ decision rule, we generated expected values of response frequencies and covariance data using the same underlying parameter sets from our simulation study (see Table 2) for three response categories. We then independently perturbed these generative parameters using a Gaussian random number generator with a *SD* of $0.15\sigma_{ext}$. We then used these perturbed parameters as an initial guess in model fitting routines to assess how changes in model parameters led to differences between expected values in the data obtained from our generative parameters. We penalized violations of criterion ordering along the decision axis, but we did not constrain our model fitting with the same constraints imposed on our simulated observer: decision and encoding noise variances were not constrained to sum to unity. We obtained fits for 500 iterations at each parameter configuration. The norm of the difference between expected values resulting from the fitting routine and those given by the true generative parameters was always greater than zero when the search failed to converge on the true parameters. That is, we did not find any alternative model solutions that resulted in nonzero costs.

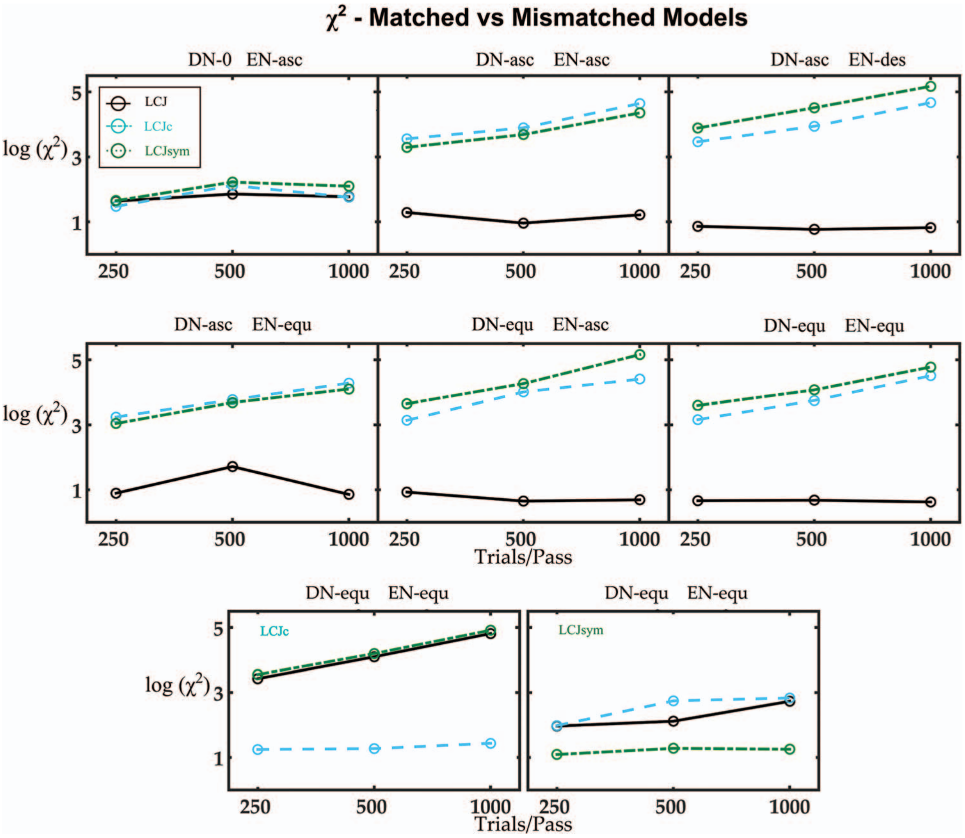


Figure 11. Top and middle rows: average log χ^2 for model fits versus trials/pass (assuming six passes/experiment) for the LCJ, LCJ$_c$, and LCJ$_{sym}$ matched to data simulated using the LCJ decision rule. Bottom: log χ^2 for model fits to simulations when decision noise and encoding noise are equal across criteria and stimulus classes. Bottom left: LCJ, LCJ$_c$, and LCJ$_{sym}$ modeled to data simulated using the LCJ$_c$ decision rule. Bottom right: LCJ, LCJ$_c$, and LCJ$_{sym}$ modeled to data simulated using the LCJ$_{sym}$ decision rule. DN = decision noise; EN = encoding noise; asc = increased; des = decreased; equ = equated; LCJ = law of categorical judgment; c = complimentary; sym = symmetrically adjusted. See the online article for the color version of this figure.

Finally, we compared the expected values of the LCJ for each of our representative parameter settings with those obtained when random numbers were given as parameter inputs to the model. The sum of squared differences between model outputs for the representative parameter sets and model outputs for random selected parameters generally increased with the euclidean distance between parameter sets. This relationship was not monotonic, but a general trend showed an increasing sum of squared error with increasing distance between parameters.

We have demonstrated the feasibility of recovering estimates for decision noise as well as encoding noise within an expanded signal detection framework for representative parameter configurations. These configurations imposed identical positioning of the criteria and signal distribution means, and caps on the total noise at the decision stage. While we do not believe that this circumstance poses any fundamental constraints on the application of our framework, more complex configurations might lead to more variable parameter estimation. For example, a higher overall total internal noise relative to external noise would necessitate a greater number of total trials to achieve comparable levels of accuracy and precision in parameter estimates. Nevertheless, the total internal noise levels assumed by our simulated observer lay well within the range often reported in multipass experiments (Burgess & Colborne, 1988; Green, 1964; Lu & Dosher, 2008). While simulation studies cannot guarantee that the parameters of the decision noise models considered here uniquely map to confidence rating and covariance estimates, we believe the demonstrations given here provide strong evidence for the efficacy of the procedure in resolving and identifying factors underlying response variability.

Application

We applied our framework to a simple visual detection confidence rating experiment to assess the degree to which decision noise contributes to response variability, and to investigate the dependence of noise components on the response structure of the task. We conducted a multipass, Gabor detection experiment with external noise in foveal vision (see Appendix C for additional details). Subjects performed in sessions with both three and five rating categories each day. For each subject and for each rating scale, we collected response frequencies and covariance estimates for signal absent and signal present trials across 5 days. We cumulatively summed response frequencies with traditional zROC plots and also plotted both within- and between-category covariance estimates for signal present and signal absent trials.

We found the best fitting lines for zROCs (fit to both coordinates) estimated with yes rates for experiments with three response categories fell above the best fitting line of the zROC determined by yes rates for experiments with five response categories for both subjects (see Figure 12). This result is consistent with the prediction of Benjamin et al. (2013) that more response categories are associated with more decision noise. We then fit our data with each of the three decision noise models (LCJ, LCJ_c, and LCJ_{sym}) and the classical signal detection model without decision noise cSDT. Our criteria for model selection among those with equal number of parameters (i.e., LCJ, LCJ_c, and LCJ_{sym}) was simply to choose the model with the lowest weighted least-squares cost function. For selection between these more complex models and the simpler reduced model cSDT, we used *F* tests for nested models (Wonnacot & Wonnacot, 1981). For both subjects, the decision noise models did not fit the data significantly better than the cSDT model without decision noise when subjects used only three response categories. With five response categories, the LCJ decision noise model fit the data better than the LCJ_c or LCJ_{sym} models, and it provided significantly better fits than the cSDT model for both subjects. For further verification of the LCJ model fits to data with five response categories, we randomly sampled, analyzed, and modeled subject responses to 80% of trial stimuli, and then computed the r^2 between predictions of the model with these parameters and the remaining 20% of the data. Repeating this procedure for over 100 repeated samples, we found median $r_{\text{ROC}} = 0.99$ in zROC data and $r_{\text{cov}} = 0.82$ in covariance estimates for subject CC, and median $r_{\text{ROC}} = 0.97$ in zROC data and $r_{\text{cov}} = 0.87$ in covariance estimates for subject YZ.

We also examined whether representational parameters at the decision stage remained constant across three and five response categories. We fit LCJ to subject data from the five-category rating experiment while jointly fitting the cSDT to the three category rating experiment. We either allowed all parameters to vary freely, or assumed that the represention-related parameters σ_{E_0}, σ_{E_1} and μ_{S_1} remained identical across response structures. For both subjects, fits using the representation-constrained model were statistically equivalent to the unconstrained model suggesting stationary representational distributions but decision noise increasing with the number of response categories. These preliminary findings suggest that decision noise may play a larger role in task processing when tasks require a large number of response categories.

Of course, other SDT models might be generating the observed data patterns—for example the data may be generated by a mixture model in which a sample representation from a signal-present trial may derive from one of two underlying distributions (DeCarlo, 2002). When a trial is well attended, the trial representation is sampled from a distribution with mean μ_{S_1} and variance $\sigma^2_{ext} + \sigma^2_{E_0}$. However, if the trial occurred during a lapse of attention then the trial representation is sampled from a distribution with mean 0 and variance $\sigma^2_{ext} + \sigma^2_{E_0}$. A mixture parameter λ determines the base rates for attended and unattended signal present trials. Relative to LCJ model, the mixture model also provided very good fits to the data, but the parameter λ changed inconsistently from three to five response categories for each subject. Cross-validation results from the mixture model and those obtained with the LCJ decision model resulted in very similar performance outcomes so we are unable to distinguish between these with our experimental data (see Appendix C for details).

Nevertheless, the success of the mixture model to account for the data patterns in an MCR experiment raises the question of whether the decision noise models might mischaracterize response variability generated from attentional lapses as variability arising from decision mechanisms. We carried out a preliminary study by

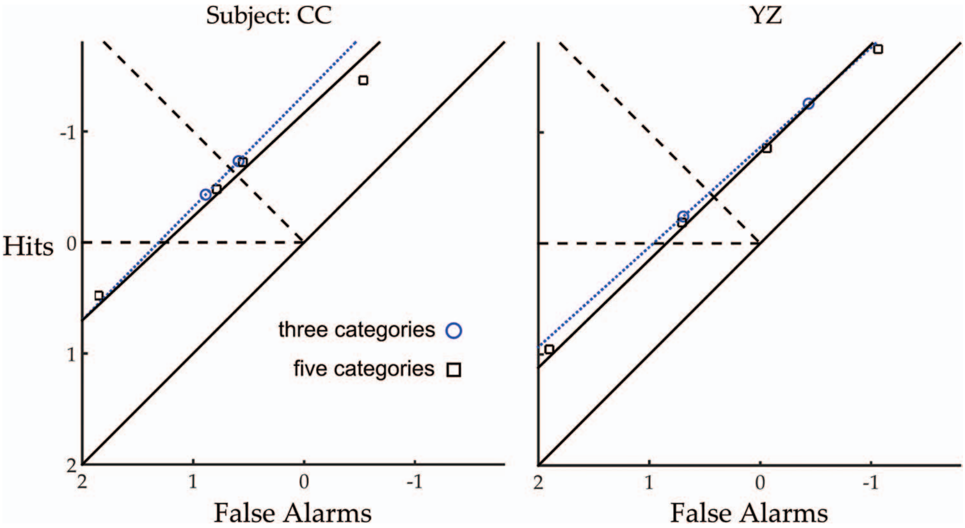


Figure 12. *z*-Score plots for subjects CC and YZ with three and five rating categories. Points on the *z*-transformed receiver operating characteristic from experiments with three rating categories lie above the best fitting line to points estimated from experiments with five rating categories. This result may reflect increasing decision noise with the use of additional response criteria. See the online article for the color version of this figure.

fitting decision noise models to simulated data from a mixture distribution. Despite assuming a 5% lapse rate well within the typical range assumed in attentional lapse studies, the decision noise models did not misattribute attentional lapse to a decision noise mechanism (see Appendix C for further details). These results suggest that the decision noise estimates from the decision noise models considered here are not mistakenly conflating decision noise with lapses in attention as an alternative mechanism of response variability.

General Discussion

In this article, we present a new framework for understanding performance in signal detection tasks that combines rating responses with multipass measurements. The framework resolves response variability arising from representation and decision processes, and can be applied to tasks with only two stimulus classes. Combined use of rating responses and multipass procedures provide stronger constraints on parameter estimation in extended SDT models with decision noise. A multipass procedure allows for a measure of total internal noise relative to consistent noise, but this technique by itself cannot achieve any further resolution of noise beyond this first-order partitioning. A rating response task with more than two stimulus classes may provide for separate estimates of decision and representation noise, but the efficacy of this approach does not extend to experiments with only two stimulus classes without significantly simplifying assumptions about the underlying noise levels. Our combination of these two approaches provides a set of observations rich enough to separate and measure contributions of noise components at the decision stage. The MCR procedure can be used whenever meaningful external noise manipulations can be defined for the stimulus set (see below).

We demonstrated the efficacy of our framework by simulating MCR experiments for observers with a number of underlying noise configurations. We modeled the data from each of these experiments and found that precision and accuracy of parameter fits improved by increasing the number of trials and passes. For each tested configuration, we found these measures improved when averaged over all parameters as well as when considering only the worst performing parameters. That each of these improvements depended on the number of trials and passes gives us strong evidence that response frequencies and response agreement estimates together constrained the extended SDT model with decision noise. More important, models with mismatched decision rules generally provided worse χ^2 with worsening results as the number of trials increased. This suggests that the framework is robust to model miss-specification and that methods of model selection could help identify underlying decision rules in addition to model parameters.

We also deployed this framework in a visual detection confidence-rating task with multiple passes. MCR procedures afforded estimates of response agreement in addition to response frequencies. For both subjects, the data were better explained by an extended SDT model with decision noise for tasks with five response categories. When only using three response categories, the decision noise model did not provide significantly better fits than the classical SDT model without decision noise. For many applications of SDT in which subjects may respond with a limited number of alternative categories, our result suggests the static criterion assumption of classical SDT remains valid and useful. However, ROCs for our subjects included features consistent with decision noise like peaked midpoints and lower performing operating characteristics for five but not three response categories. When a task structure offers a larger number of response categories, decision noise may become an important determinant in trial-by-trial response outcomes. Of course, the models we use to interpret our data affect what kinds of conclusions we may draw, and the classical signal detection model can be elaborated in a number of ways. A mixture model with static criteria (DeCarlo, 2002) provided very good fits as well when applied to our data. Moreover, the assumption of a latent distribution in the mixture model seems no less plausible than the assumption of fluctuating criteria in decision noise models. It may be the case that the decision noise models considered here misattribute an underlying latent distribution to greater variability in the criteria. To test this consideration we ran 250 additional simulated experiments of an MCR procedure to emulate an observer with static criteria. We assumed equal variance for signal-absent and signal-present distributions, sensitivity (d') equal to one, and a 5% rate of attention lapses modeled by sampling from a latent signal-present distribution with mean of zero. We simulated six passes of 500 trials each to match our experimental procedure and then fit these simulated data sets with each of our decision noise models. The median model fits showed recovered parameters quite close to the actual generative parameters used in the simulations. In particular, median fits for criterion variances were very nearly zero and median estimates of the positions of these criteria only slightly underestimated the true locations along the decision axis. The median fits for encoding parameters also closely matched the underlying generative parameters, although in this case the solutions converged with considerable variability and sometimes resulted in entirely unrealistic parameter values. Distinguishing between elaborated SDT models positing alternative mechanisms will require future experimental work and the developments presented in this article allow for the consideration of explanations involving decision noise that were not previously available.

Key features of ROC and zROC data do not depend on the static criterion assumption and in some cases contradict it. In the case of rating procedures, our framework now provides a way to identify and quantify the separate contributions of encoding and decision noise to these features. For example, some researchers have noted that the “peaks” in empirical zROCs could emerge with highly stable central criteria and highly variable criterion boundaries at more extreme positions (Mueller & Weidemann, 2008; Wickelgren, 1968). In the current study, one subject exhibited a peaked zROC and our model fits verified this prediction quantitatively. The framework introduced here may shed light on other anomalies observed in zROC data as well. Previous work has argued that decision noise is induced in rating tasks when task instructions require subjects to use the rating categories with equal frequency (Murray et al., 2002) or, more generally, when task instructions alter criterion placement from default positions that subjects would use absent any instruction (Kellen et al., 2012; Wixted & Gaitan, 2002). These authors suggest that decision noise emerges from the conflict between subject’s preconditioned preferences acquired over extensive lifetime experience, and instructions that bias subjects to adopt criterion positions conflicting with these default preferences. The subjects in our study had extensive practice in psychophysics experiments, so we expect that default preferences were moderated. Moreover, while we asked subjects to utilize the full scale, we did not request that subjects use each response category with equal frequency. Still, we remain agnostic as to whether decision noise results from conflicts between response instruction and predisposition, or whether this arises because of limitations on the resolution of a representation-response mapping, or for any other reason. The method we propose here may prove useful in determining the degree to which response instruction and subject expertise influence response variability.

Ours is not the first attempt to resolve decision and representational processes in signal detection tasks. For example, Wickelgren (1968) proposed a "criterion operating characteristic" that allowed for comparison of the variances of criteria adopted across different signal strengths. The method's validity, however, assumes equal noise *SD*s for all signal strengths. An alternative framework has been developed to separate decision and representational noise in the domain of perception with the decision noise model (DNM) of Mueller and Weidemann (2008). In memory recognition, Benjamin et al. (2009) developed an Ensemble Recognition task in which participants gave confidence ratings on whether stimulus ensembles of a variable number of words were previously observed on a study list. These authors compared fits from a number of models and reached the conclusion that decision noise played a significant role in subject performance. However, Kellen et al. (2012) introduced their own model generalization approach for memory recognition that interleaved trials of a 4AFC-ranking task with those of a confidence rating procedure. These authors found no evidence of decision noise in their study and offered a critique of the conclusions drawn by Benjamin et al. The merits and shortcomings of each of these frameworks are discussed in detail in Kellen et al. (2012) and Benjamin (2013).

In our view, both the Ensemble Recognition and the model generalization approach advance our understanding of response variability considerably, although they reach contradictory conclusions about the significance of decision noise in confidence rating tasks for recognition memory. One potential limitation with both of these approaches is the strong constraints imposed between different noise components. The Ensemble Recognition paradigm assumes that a single variance term applies to the noise at all criterion boundaries. Likewise, the model generalization approach assumes either a single variance for decision noise across all criteria (adopting the LCJ as a decision rule) or a single variance for the confidence boundaries (adopting the DNM decision rule). Our own experimental results suggest that criterion noise may vary considerably across criterion boundaries when decision noise is significant (see Appendix C for details). Further, the model generalization approach assumes that representational noise is constant across forced choice and rating-response paradigms, and that no decision bias (and by extension no decision noise) is present during the forced choice tasks. Though Kellen et al. argue that the decision bias observed in forced choice tasks only applies when trial stimuli are presented in sequence, the presence or absence of any such bias is ultimately unknown and is not precluded by their model. Bias has been shown to play a role in similar experimental paradigms that had previously assumed a bias free framework (Klein, 2001; Yeshurun, Carrasco, & Maloney, 2008). If decision noise contributes to response variability in n-alternative forced choice tasks, it may appear as inflated representational noise during model fitting; this inflated estimate of representation variability may then incorrectly discount the effects of any decision noise in the corresponding rating task. More generally, the constraints imposed by these models may lead to parameter estimates that do not accurately reflect underlying processes in representation and decision-making. Rosner and Kochanski's (2009) LCJ model allows independent parameter estimates for variance terms at the decision stage in paradigms with at least three stimulus intensities and at least four response categories. Although this model provides a powerful new tool to understand categorical judgment, it does not apply to the frequently used signal detection task with two stimulus classes without introducing constraints among the noise components. The framework presented here fills that gap for tasks with at least three response categories while allowing independence among noise components.

An essential feature of our approach requires the implementation of external noise. Research in recognition memory has not generally implemented this method, but the external noise method is not fundamentally incompatible with investigations of higher-level cognitive processes (Lu & Dosher, 2008, p. 71). For example, Tsetsos, Chater, and Usher (2012) used external noise to examine decision biases and preference reversals in the domain of economic value integration. With regard to the MCR method in particular, however, mnemonic representations of both studied and unstudied items will likely change with the number of times stimuli are presented during test trials. However, the MCR paradigm is only one of a number of methods that use multiple presentations to investigate levels of internal noise (Burgess & Colborne, 1988; Swets, Shipley, McKey, & Green, 1959). In particular, Nosofsky (1983) used multiple presentations without the use of external noise to estimate the representation and criterion noise in an auditory identification task. Nosofsky deployed this method to study noise contributions to the range effect, but this technique might offer a means of determining decision noise for tasks with only binary response alternatives. The Ensemble Recognition task of Benjamin et al. (2009) in the domain of recognition memory bears some resemblance to this approach insofar as additional presentations (or larger ensemble size) of stimulus samples lead to less variability in processes underlying representation.

Recent studies have brought to light the importance of a decision rule that resolves ambiguities that arise with noisy criterion boundaries in signal detection tasks with three or more response categories (Klauer & Kellen, 2012). When trial-sampled criteria overlap, category assignment becomes ambiguous without specific decision rules accounting for contingencies owing to positional relations among criteria and representations. However, any possible set of rules unambiguously resolving trial-sampled representations to category assignment may serve as a decision rule. Our experiments used either three or five response categories. The symmetry (or lack thereof) in the number of response categories may influence the choice of rule adopted by our subjects. Symmetric response structures have an odd number of category boundaries and an even number of response categories. These response structures might induce the adoption of an initial, central, and binary decision boundary with participants only subsequently utilizing the remaining criteria as a confidence rating on their antecedent choice. This is dubbed a sequential rule, along with any rule whereby subjects compare trial stimuli with trial-sampled criteria in a sequential manner. Asymmetric response structures have an even number of category boundaries. Because asymmetric response structures, like the one we examined in this study, do not naturally suggest any particular criterion as a central designation as in symmetric response structures, we restricted our examination to simultaneous rules in this article. However, rating category asymmetry may naturally allow for the emergence of a neutral category that subjects use as a preferred classification during trials with lapses of attention and so may not wholly reflect categorization based on representational determinants. Although we cannot determine a priori which decision rule a subject might adopt, specific data signatures may reflect idiosyncratic strategies to deal with significantly different processing constraints in the course of encoding information and making decisions about that information. Previous studies lend weight to the idea that task instructions (explicitly; Treisman & Faulkner, 1985), response structure (implicitly), and individual subject differences (Petrov, 2009) may all influence decision rule adoption. We hope to explore alternative decision rules and hybrid rules in future studies.

Klauer and Kellen (2012) showed that if an observer's criterion boundaries were centered and distributed evenly about the mean of an underlying representational distribution, the LCJ would yield asymmetric response distributions. They argued instead for a modified decision rule that determined response selection according to the proximity of an internal representation to the trial-sampled criteria and that would result in a symmetric distribution of response frequencies. We have instantiated that alternative rule here as LCJ_{sym}, but have found it underperformed relative to LCJ in our data sets for which decision noise was deemed significant. Given the limitations of our experimental study, we hesitate to make strong claims regarding the general validity of alternative decision rules in operation for specific tasks or individuals. Other tasks or experimental manipulations may very well induce subjects to adopt another decision rule such as LCJ_{sym} and the framework introduced here may allow us to identify that rule.

Experimental paradigms investigating perceptual and cognitive processes obtain information about these underlying processes by examining responses conditioned on input stimuli, task instructions, subject population, and so forth In the case of an MCR procedure, we collect additional information by conditioning subject responses on specific samples of external noise. By presenting these samples over multiple passes, we can estimate response agreement to test more nuanced hypotheses than would be feasible otherwise. Sequential dependence, for example, may offer a potential target for investigation insofar as the phenomenon of these dependencies introduce a form of systematic decision noise. Trial-by-trial dependencies certainly bear on estimates of agreement in multipass psychophysics tasks. Sequential dependencies influenced by stimulus schedule (Fernberger, 1920; Parducci, 1959), response choice (Howarth & Bulmer, 1956), or feedback (Carterette et al., 1966) could generate greater response agreement to the degree that these factors are preserved across passes. In this case, estimates of the internal to external noise ratio are at a lower bound. If response dependencies artificially increase agreement estimates, then removing these dependencies will reduce covariance estimates, which in turn leads to greater estimates of internal noise (Green, 1964). Levi et al. (2005) proposed randomizing the sequence of trials from pass to pass to mitigate agreement effects deriving from stimulus-response dependencies. The current study followed the prescription of Levi et al. by randomizing the stimulus schedule from pass to pass, but we did not examine response data for synchronized stimulus schedules across passes. Comparing internal to external noise ratios measured in multipass experiments with and without randomized trial ordering suggests itself as one way to begin teasing apart the purely stimulus related factors on trial outcomes from other contributions to response agreement.

Elaborated observer models makes more detailed claims regarding the functional mechanisms transforming stimulus inputs to overt responses (Lu & Dosher, 2008, 2013). Many of these models emphasize the account of representational processing, but use the simplified decision processes of standard SDT. When ignored, response variability arising from decision processes will redound to representational processes instead, potentially leading to erroneous model predictions. When task conditions call for increasing the number of response categories, decision boundaries may become more variable (Ratcliff & Starns, 2009). In these cases, observer models incorporating our framework may lead to a more detailed understanding of the transformation from stimulus to response.

The aim of analyzing noise contributions is a fundamental objective in cognitive psychology. Isolating component sources of noise helps us to characterize corresponding component processes in human behavior and decision making (Brunton, Botvinick, & Brody, 2013; Ratcliff & Starns, 2009). The MCR paradigm makes available new research directions involving noise analysis and decision strategy. The importance of the MCR procedure and analyses in future research will depend upon the amount of decision noise present for a given task, subject population, and experimental condition. If the decision noise is relatively negligible, a simpler SDT model will serve as a more parsimonious and efficient explanation for the observed outcomes. The experimental results presented here suggest that decision noise is not a significant determinant for tasks with few response alternatives, but may become more influential when the number of response alternatives increase.

Conclusion

In this article, we present a new framework that combines two well-established procedures in psychophysics: a confidence rating response procedure and a multipass experimental paradigm. In combination, these procedures allow estimation of response agreement as well as response frequency for each response category. We provide evidence that data collected with this framework sufficiently constrains extended SDT models with decision noise. Our simulation study showed that the parameters of a decision noise model fitted to responses from simulated experiments led to increasing accuracy and precision with increasing trials and passes. These simulations also demonstrated that decision noise models matched to the decision rule adopted by the subject will outperform mismatched models. We also conducted a visual detection rating experiment with multiple passes. Our results showed that decision noise was negligible when subjects responded with three confidence rating categories, but that it influenced trial responses with as few as five response categories. For tasks with few response alternatives, classical SDT may adequately account for the observed data. However, for tasks offering a large number of response alternatives or where decision noise is suspected, the framework presented here offers a more detailed description of the underlying processes.

References

Barlow, H. B. (1957). Increment thresholds at low intensities considered as signal/noise discriminations. *The Journal of Physiology, 136,* 469–488. http://dx.doi.org/10.1113/jphysiol.1957.sp005774

Benjamin, A. S. (2013). Where is the criterion noise in recognition? (Almost) everyplace you look: Comment on Kellen, Klauer, and Singmann (2012). *Psychological Review, 120,* 720–726. http://dx.doi.org/10.1037/a0031911

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

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1601–1608. http://dx.doi.org/10.1037/a0031849

Braida, L. D., & Durlach, N. I. (1972). Intensity perception: II. Resolution in one-interval paradigms. *The Journal of the Acoustical Society of America, 51,* 483–502. http://dx.doi.org/10.1121/1.1912868

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *The Journal of Neuroscience, 12,* 4745–4765.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science, 340,* 95–98. http://dx.doi.org/10.1126/science.1233912

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A, 5,* 617–627. http://dx.doi.org/10.1364/JOSAA.5.000617

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Atlanta, GA: Sage.

Carterette, E. C., Friedman, M. P., & Wyman, M. J. (1966). Feedback and psychophysical variables in signal detection. *Journal of the Acoustical Society of America, 39,* 1051–1055. http://dx.doi.org/10.1121/1.1909991

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

Dosher, B., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel selection. *Proceedings of National Academy, 95,* 13988–13993.

Dosher, B. A., & Lu, Z.-L. (1999). Mechanisms of perceptual learning. *Vision Research, 39,* 3197–3221. http://dx.doi.org/10.1016/S0042-6989(99)00059-0

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. http://dx.doi.org/10.1121/1.1911699

Egan, J. P., Shulman, 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. http://dx.doi.org/10.1121/1.1907783

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

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. http://dx.doi.org/10.3758/BF03212703

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature, 402,* 176–178. http://dx.doi.org/10.1038/46027

Graham, N. V. S. (1989). *Visual pattern analyzers*. New York, NY: Oxford University Press.

Gravetter, F., & Lockhead, G. R. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review, 80,* 203–216. http://dx.doi.org/10.1037/h0034281

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review, 71,* 392–407. http://dx.doi.org/10.1037/h0044520

Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *Journal of the Acoustical Society of America, 97,* 3749–3760. http://dx.doi.org/10.1121/1.412390

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

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

Hutchinson, T. P. (1981). A review of some unusual applications of signal detection theory. *Quality & Quantity: International Journal of Methodology, 15,* 71–98. http://dx.doi.org/10.1007/BF00144302

Kac, M. (1962). A note on learning signal detection. *I. R. E. Transactions on Information Theory, 8,* 126–128. http://dx.doi.org/10.1109/TIT.1962.1057687

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review, 119,* 457–479. http://dx.doi.org/10.1037/a0027727

Klauer, K. C., & Kellen, D. (2012). The law of categorical judgment (corrected) extended: A note on Rosner and Kochanski (2009). *Psychological Review, 119,* 216–220. http://dx.doi.org/10.1037/a0025824

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics, 63,* 1421–1455. http://dx.doi.org/10.3758/BF03194552

Lesmes, L. A., Jeon, S.-T., Lu, Z.-L., & Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. *Vision Research, 46,* 3160–3176.

Levi, D. M., & Klein, S. A. (2003). Noise provides some new signals about the spatial vision of amblyopes. *The Journal of Neuroscience, 23,* 2522–2526.

Levi, D. M., Klein, S. A., & Chen, I. (2005). What is the signal in noise? *Vision Research, 45,* 1835–1846.

Li, X., Lu, Z.-L., Xu, P., Jin, J., & Zhou, Y. (2003). Generating high gray-level resolution monochrome displays with conventional computer graphics cards and color monitors. *Journal of Neuroscience Methods, 130,* 9–18.

Lu, Z.-L., & Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A, 16,* 764–778. http://dx.doi.org/10.1364/JOSAA.16.000764

Lu, Z.-L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review, 115,* 44–82. http://dx.doi.org/10.1037/0033-295X.115.1.44

Lu, Z.-L., & Dosher, B. A. (2013). *Visual psychophysics: From laboratory to theory*. Cambridge, MA: The MIT Press.

Lu, Z.-L., & Sperling, G. (1999). Second-order reversed phi. *Perception & Psychophysics, 61,* 1075–1088. http://dx.doi.org/10.3758/BF03207615

Luce, R. D., & Nosofsky, R. M. (1984). Attention, stimulus range, and identification of loudness. In S. Kornblum & J. Raquin (Eds.), *Preparatory states and processes* (pp. 3–25). Hillsdale, NJ: Erlbaum.

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. http://dx.doi.org/10.3758/BF03202769

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A users guide*. Mahwah, NJ: Erlbaum.

Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review, 84,* 452–471. http://dx.doi.org/10.1037/0033-295X.84.5.452

McClelland, G. H. (2011). Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 83–100). Washington, DC: National Academies Press.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241. http://dx.doi.org/10.1146/annurev.psych.50.1.215

Miller, G. A. (1956, March). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. http://dx.doi.org/10.1037/h0043158

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

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2,* 79–104. http://dx.doi.org/10.1167/2.1.6

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

Parducci, A. (1959). An adaptation-level analysis of ordinal effects in judgment. *Journal of Experimental Psychology, 58,* 239–246.

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73,* 44–58. http://dx.doi.org/10.1037/h0022662

Pelli, D. G. (1990). The quantum efficiency of vision. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 3–24). Cambridge: Cambridge University Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. http://dx.doi.org/10.1163/156856897X00366

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *IRE Professional Group on Information Theory, 4,* 171–212. http://dx.doi.org/10.1109/TIT.1954.1057460

Petrov, A. A. (2009). Symmetry-based methodology for decision-rule identification in *same—Different* experiments. *Psychonomic Bulletin & Review, 16,* 1011–1025. http://dx.doi.org/10.3758/PBR.16.6.1011

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America, 24,* 745–749. http://dx.doi.org/10.1121/1.1906969

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

Rosner, B. S., & Kochanski, G. (2009). The law of categorical judgment (Corrected) and the interpretation of changes in psychophysical performance. *Psychological Review, 116,* 116–128. http://dx.doi.org/10.1037/a0014463

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes, 60,* 1–13. http://dx.doi.org/10.1006/obhd.1994.1072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review, 108,* 183–203. http://dx.doi.org/10.1037/0033-295X.108.1.183

Spiegel, M. F., & Green, D. M. (1981). Two procedures for estimating internal noise. *The Journal of the Acoustical Society of America, 70,* 69–73. http://dx.doi.org/10.1121/1.386583

Swets, J. A., Shipley, E. F., McKey, M. J., & Green, D. M. (1959). Multiple observations of signals in noise. *The Journal of the Acoustical Society of America, 31,* 514. http://dx.doi.org/10.1121/1.1907745

Tanner, W. P., Jr. (1961). Physiological implications of psychophysical data. *Annals of the New York Academy of Sciences, 89,* 752–765. http://dx.doi.org/10.1111/j.1749-6632.1961.tb20176.x

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409. http://dx.doi.org/10.1037/h0058700

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

Thomas, E. A. C. (1975). Criterion adjustment and probability matching. *Perception & Psychophysics, 18,* 158–162. http://dx.doi.org/10.3758/BF03204104

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

Treisman, M. (1984). A theory of criterion setting: An alternative to the attention band and response ratio hypotheses in magnitude estimation and cross-modality matching. *Journal of Experimental Psychology: General, 113,* 443–463. http://dx.doi.org/10.1037/0096-3445.113.3.443

Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 119–139. http://dx.doi.org/10.1037/0096-1523.10.1.119

Treisman, M., & Faulkner, A. (1985). Can decision criteria interchange locations - some positive evidence. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 187–208.

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

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35,* 2503–2522. http://dx.doi.org/10.1016/0042-6989(95)00016-X

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 9659–9664. http://dx.doi.org/10.1073/pnas.1119569109

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63,* 1293–1313. http://dx.doi.org/10.3758/BF03194544

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 positions in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316–347. http://dx.doi.org/10.1016/0022-2496(66)90018-6

Wixted, J. T., & Gaitan, S. C. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior, 30,* 289–305. http://dx.doi.org/10.3758/BF03195955

Wonnacot, T. H., & Wonnacot, R. J. (1981). *Regression: A second course in statistics*. New York, NY: Wiley.

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research, 48,* 1837–1851. http://dx.doi.org/10.1016/j.visres.2008.05.008

(*Appendices follow*)

General Appendix

Detailed Derivations and Experimental Procedures

Equation 1 illustrates a schema of noise components comprising the total variability of response, σ_{T_h}, to stimulus S_h at a single criterion, *C*. We now consider how three different task structures may constrain and identify these various components. These are *single criterion—single pass*, *single criterion—multi pass*, and *multiple criterion—multi pass*. In what follows, we denominate the means of criterion boundaries μ_C, as well as the means of the stimulus distributions, μ_S, in units of σ_{T_0}, σ_{S_0}, and σ_ext depending on convenience of task analysis.

Appendix A

Response Frequencies and Covariance

Single Criterion: Single Pass

In a typical signal detection task for which subjects provide a binary response to each trial event, we conceive the decision processes as a comparison of the internal representation of the stimulus to the position of the criterion boundary along the decision axis at some position μ_C. The traditional detection paradigm involves only two stimulus classes, “signal absent and “signal present”; in the following exposition, we let stimulus class $h = 0$ represent our “signal absent” stimulus and $h = 1$ represents our “signal present” stimulus. Then given some internal representation of a trial sample $s_{S_h}\sigma_{S_h} + \mu_{S_h}\sigma_{S_0}$ from stimulus $S_{h,}$subjects transform this internal response into an explicit response *R* according to the following decision rule.

$$R = \begin{cases} 1 & s_{S_h}\sigma_{S_h} + \mu_{S_h}\sigma_{S_0} > \mu_C\sigma_{S_0} \\ 0 & s_{S_h}\sigma_{S_h} + \mu_{S_h}\sigma_{S_0} \leq \mu_C\sigma_{S_0} \end{cases} \tag{A1}$$

Assuming a subject has sufficient knowledge of the probability density functions of the representational distribution for stimulus S_h, traditional SDT assumes that subjects maintain a static value of $\mu_C\sigma_{S_0}$ once they understand task instructions, payoff structure, and have adequate information to compute the likelihood ratio $\phi_{S_h}(\mu_C\sigma_{S_0})/\phi_{S_0}(\mu_C\sigma_{S_0})$. These affirmative responses to samples from the stimulus class *h* are given as,

$$P(R = 1 \mid S_h) = \int_{\mu_C\sigma_{S_0}}^{\infty} \phi_{S_h}(x)\, dx \tag{A2}$$

If there is some variability in the criterion, we represent a trial-sampled criterion offset as $c\sigma_\text{C}$. Then the decision rule is slightly modified for a given sample pair as follows.

$$R = \begin{cases} 1 & s_{S_h}\sigma_{S_h} + \mu_{S_h}\sigma_{S_0} > \mu_C\sigma_{S_0} + c\sigma_C \\ 0 & s_{S_h}\sigma_{S_h} + \mu_{S_h}\sigma_{S_0} \leq \mu_{C_0}\sigma_{S_0} + c\sigma_C \end{cases} \tag{A3}$$

Let Φ be the Gaussian cumulative distribution function and let Q $(x;\, \mu,\, \sigma) = 1 - \Phi\,(x;\, \mu,\, \sigma)$. When dealing with decision noise, the overall rate of affirmative responses for S_h trials is given as,

$$\begin{aligned} P(R = 1 \mid S_h) &= P(s_{S_h}\sigma_{S_h} + \mu_{S_h}\sigma_{T_0} > \mu_C\sigma_{T_0} + c\sigma_C) \\ &= P(s_{s_h}\sigma_{S_h} - c\sigma_C > \mu_C\sigma_{T_0} - \mu_{S_h}\sigma_{T_0}) \\ &= P(s_{T_h}\sigma_{T_h} > (\mu_C - \mu_{S_h})\sigma_{T_0}) \\ &= Q\left[(\mu_C - \mu_{S_h})\frac{\sigma_{T_0}}{\sigma_{T_h}}\right] \end{aligned} \tag{A4}$$

From Equation A4 we may recover the criterion position relative to each stimulus distribution *h* in units of the total variability of each distribution (we make the usual assignments of $\sigma_{T_0} = 1$ and $\mu_{S_0} = 1$). Additionally, assuming $\sigma_{T_h} = \sigma_{T_0}$, we may also recover the position μ_{S_h} in units of σ_{T_0}. We estimate these quantities with the following equations.

$$\begin{aligned} &A5.a \quad (\mu_C - \mu_{S_h})\frac{\sigma_{T_0}}{\sigma_{T_h}} = -z\big[P(R = 1 \mid S_h)\big] \\ &A5.b \qquad\qquad d' = \mu_{S_h} \\ &\qquad\qquad\qquad = z\big[P(R = 1 \mid S_h)\big] - z\big[P(R = 1 \mid S_0)\big] \end{aligned} \tag{A5}$$

When $\sigma_{T_h} \neq \sigma_{T_0}$, the position of μ_{S_h} can still be recovered in units of σ_{T_0} if we induce subjects to adopt different criteria *C* through experimental manipulation. In that case, we may recover the functional relationship $z[P\,(R = 1|S_h)] = f\,(z\,[P\,(R = 1|S_0)]$; assuming total noise remains constant), and estimate $\mu_{S_h} = \mu_C$ when $z[P(R = 1|S_\text{h})] = 0$. In addition, the slope of this functional form gives us $\alpha_h = \sigma_{T_0}/\sigma_{T_h}$.

(*Appendices continue*)

Table A1
Model Variables and Parameters[3]

Variables/parameters	Formal definition	Description
s_{ext}	$s_{ext} \in G(0, 1)$	Trial-sampled component of internal response induced by external noise.
σ^2_{ext}		Variance of *consistent noise* across all possible samples of external noise. We assume $\sigma_{\mathrm{ext}} = 1$.
s_{E_h}	$s_{E_h} \in G(0, 1)$	Trial-sampled component of internal response because of encoding processes on S_{h} trials.
μ_{S_h}		Mean of internal response to h^{th} stimulus category.
$\sigma^2_{E_h}$		Variance of *encoding noise* during S_{h} trials.
c_m	$c_m \in G(0, 1)$	Trial-sampled criterion of the m^{th} decision boundary.
μ_{C_m}		Mean of representation of m^{th} criterion boundary.
$\sigma^2_{C_m}$		Variance of *criterion noise* at the m^{th} decision boundary.
s_{U_h}	$s_{U_h} \in G(0, 1)$	Sample of *random noise* at the decision boundary reflecting variability of encoding and decision processes on S_{h} trials.
$\sigma^2_{U_h}$	$\sigma^2_{U_h} = \sigma^2_{E_h} + \sigma^2_C$	Variance of *random noise* during S_h trials at a single decision boundary.
s_{S_h}	$s_{S_h} \in (0, 1)$	Sample of *representational noise* reflecting encoding processes and external noise on S_{h} trials.
$\sigma^2_{S_h}$	$\sigma^2_{S_h} = \sigma^2_{ext} + \sigma^2_{E_h}$	Variance of *representational noise* during S_{h} trials
s_{T_h}	$s_{T_h} \in G(0, 1)$	Sample of *total noise* reflecting all stimulus, encoding, and decision factors on S_{h} trials.
$\sigma^2_{T_h}$	$\sigma^2_{T_h} = \sigma^2_{ext} + \sigma^2_{E_h} + \sigma^2_C$	Total variance of internal response on S_{h} trials.
$\boldsymbol{R} \equiv m$	$(\boldsymbol{R} \equiv m) \sim B(p)$	Bernoulli random variable signifying a match of subject response with a given stimulus category *m.*

Because the only relevant variance terms in Equation A5 are σ_{T_0} and $\sigma_{T_{\mathrm{h}}}$, the underlying variance components σ^2_C and $\sigma^2_{S_h}$ are constrained only by the relations

$$\begin{gathered} 0 \le \sigma^2_C \le \underset{\forall h}{min}\left[\sigma^2_{T_h}\right] \\ \underset{\forall h \neq h'}{max}\left[0, \sigma^2_{T_h'} - \sigma^2_{T_h}\right] \le \sigma^2_{S_h'} \le \sigma^2_{T_h'} \end{gathered} \tag{A6}$$

Therefore, any values σ_C, σ_{S_h}, satisfying the equations $\sigma^2_{S_0} + \sigma^2_C = \sigma^2_{T_0}$ and $\sigma^2_{S_h} + \sigma^2_C = \sigma^2_{T_h}$ will suffice to explain the observations $(\mu_{C_0} - \mu_{S_h})\dfrac{\sigma_{T_0}}{\sigma_{T_h}}$ and d'. In other words, we cannot separately estimate σ_C and σ_{S_h}

Single Criterion: Multiple Passes

Multi pass experiments ultimately provide more information from the data, providing not only estimates of response frequency for each stimulus class but also for each individual noise sample. Under the assumptions of the multi pass methodology, each external noise sample induces a representation comprised of a reproducible component, for example, s_{ext}, as well as a *random* component $s_{E_h}\sigma_{E_h}$. The consistent component is presumed to yield identical values for identical external noise samples, whereas the random component arising from encoding processes is presumed to deviate even for identical stimulus samples. Over multiple presentations across passes, we can estimate the probability that an observer will provide an affirmative response R given an external noise sample $s_{\mathrm{ext}}\sigma_{\mathrm{ext}}$. We will derive these probabilities and other relevant quantities from expected values over response outcomes EV_R and expected values over external noise samples $EV_{s_{ext}}$. The probability of an affirmative response, given sample s_{ext} is,

$$EV_R[R \equiv 1 \mid s_{ext}, S_h] = P(R = 1 \mid s_{ext}, S_h) \tag{A7}$$

[3] In addition to their use as samples of random variables, the terms s_{ext}, s_{E}, s_{U}, s_{S}, s_{T}, and c_{m} will sometimes be used as random variables themselves. In these cases they will be denoted in boldface.

(*Appendices continue*)

The factors contributing to these probability estimates conditioned on s_{ext} may be expanded as,

$$\begin{aligned} P\left(R=1 \mid s_{ext}, S_h\right) &= P\left(s_{ext} + \mu_{S_h} + s_{E_h}\sigma_{E_h} > \mu_C + c\sigma_C\right) \\ &= P\left(s_{E_h}\sigma_{E_h} - c\sigma_C > \mu_C - s_{ext} - \mu_{S_h}\right) \\ &= P\left(s_{U_h}\sigma_{U_h} > \mu_C - s_{ext} - \mu_{S_h}\right) \\ &= Q\left[\left(\mu_C - s_{ext} - \mu_{S_h}\right)/\sigma_{U_h}\right] \end{aligned} \tag{A8}$$

The probabilities expressed in Equation A8 are conditioned on the consistent component of the internal representation of a specific stimulus sample. Generally speaking, stimulus samples inducing greater values of s_{ext} tend to lead to higher probabilities that the subject will respond affirmatively to trial stimuli. The overall "yes" rate for a given stimulus class h is the expectation of a "yes" response with respect to s_{ext}.

$$\begin{aligned} P(R=1 \mid S_h) &= EV_{s_{ext}}\left[P(R=1 \mid s_{ext}, S_h)\right] \\ &= \int P(R=1 \mid s_{ext}, S_h)\phi(s_{ext})\, ds_{ext} \end{aligned} \tag{A9}$$

On the other hand, higher *consistency* between responses is conditioned on the total *random* noise of the internal representation. This random noise poses the limiting factor for consistent responses to a repeated stimulus with a given sample of external noise. When the ratio of total random to consistent noise is very low, response consistency will be high and the quantities expressed in Equation A8 will nearly equal zero or one for any given sample of external noise. On the other hand, when the internal to external noise ratio is very high, response consistency decreases and the probabilities in Equation A8 become less extreme. We may express the covariance of response between corresponding trials of two passes i and j as,

$$\begin{aligned} Cov[R_i, R_j \mid s_{ext}, S_h] &= EV_R\{[(R_i \mid s_{ext}, S_h) - P(R_i = 1 \mid S_h)][(R_j \mid s_{ext}, S_h) - P(R_j = 1 \mid S_h)] \\ &= P(R_i = 1, R_j = 1 \mid s_{ext}, S_h) - P(R_i = 1 \mid S_h), P(R_j = 1 \mid s_{ext}, S_h) \\ &\quad - P(R_j = 1 \mid S_h)P(R_i = 1 \mid s_{ext}, S_h) + P(R_i = 1 \mid S_h)P(R_j = 1 \mid S_h) \\ &= P(R = 1 \mid s_{ext}, S_h)^2 - 2P(R = 1 \mid s_{ext}, S_h)P(R = 1 \mid S_h) + P(R = 1 \mid S_h) \end{aligned} \tag{A10}$$

Under the multi pass procedure and using $\sigma_{ext} = 1$ as a unit of measure, Equation A5.a is restated as,

$$\frac{\mu_C - \mu_{S_h}}{\sigma_{T_h}} = -z\left[P\left(R = 1 \mid S_h\right)\right] \tag{A11}$$

Then the observed covariance estimates of responses R across corresponding trials between the i^{th} and j^{th} passes are computed as the expected values of the covariance with expectation taken with respect to s_{ext}.

$$\begin{aligned} Cov[R_i, R_j \mid S_h] &= EV_{s_{ext}}\{Cov\left[R_i, R_j \mid s_{ext}, S_h\right]\} \\ &= \int P(R = 1 \mid s_{ext}, S_h)^2\phi(s_{ext}) - P(R = 1 \mid S_h)^2 \\ &= \int Q\left[\left((\mu_C - \mu_{S_h}) - s_{ext}\right)\frac{1}{\sigma_{U_h}}\right]^2 \phi(s_{ext})ds_{ext} - P(R = 1 \mid S_h)^2 \\ &= \int Q\left[\left(-z\left[P(R = 1 \mid S_h)\right](1 + \sigma_{U_h}^2)^{1/2} - s_{ext}\right)\frac{1}{\sigma_{U_h}}\right]^2 \phi(s_{ext})ds_{ext} - P(R = 1 \mid S_h)^2 \end{aligned} \tag{A12}$$

For a given response frequency $P(R = 1 \mid S_h)$ the covariance is monotonically related to σ_{U_h}. That is, a given "yes" rate along with a covariance estimate corresponds to a specific ratio of *random* and *consistent* response variability. Therefore, by Equation A12, we may estimate σ_{U_h}. Squaring this term and using Equation 1, we can compute $\sigma_{T_h} = (1 + \sigma_{U_h}^2)^{1/2}$. Further, we may recover $\mu_C - \mu_{S_h} = -z[P(R = 1 \mid S_h)]\sigma_{T_h}$ as well as the mean of the signal distribution along the decision axis as, $\mu_{S_h} = z[P(R = 1 \mid S_h)]\sigma_{T_h} - z[P(R = 1 \mid S_0)]\sigma_{T_0}$.

When the internal noise is equal to zero, the covariance of response outcomes across i^{th} and j^{th} passes will equal the expected variance of the "yes" rate as calculated as a binomial random variable. That is, as $P(R = 1 \mid S_h) - P(R = 1 \mid S_h)^2$ (see Appendix B). For higher internal to external noise ratios, the covariance decreases.

The foregoing analysis shows that the multi pass procedure can recover the mean of the signal distribution along the decision axis (in units of σ_{ext}) without the equal variance assumption. Further, if

(*Appendices continue*)

internal noise does not change across bias manipulations we can predict the slope the zROC at a single criterion measurement. However, at this point, we have yet to isolate the quantity of response variability due to decision processes. With a single criterion, the components of random noise are only constrained by the following relations.

$$\begin{gathered} 0 \le \sigma_C^2 \le \underset{\forall h}{min}\left[\sigma_{U_h}^2\right] \\ \underset{\forall h \neq h'}{max}\left[0, \sigma_{U_h'}^2 - \sigma_{U_h}^2\right] \le \sigma_{E_h'}^2 \le \sigma_{U_h'}^2 \end{gathered} \tag{A13}$$

This implies that any values σ_C and σ_{E_h} consistent with $\sigma_{E_h}^2 + \sigma_{C_0}^2 = \sigma_{U_h}^2$ may generate the "yes" rates and covariance data of Equations A9 and A12. We cannot obtain unique solutions for the two terms from the data. We now attempt to resolve these components using multiple criteria.

Multiple Criterion: Decision Rules

As mentioned previously, the introduction of decision noise into signal detection models involving multiple criteria raises the issue of a decision rule. A decision rule is a strategy that allows an observer to assign a specific response to an internal representation. When decision noise is inconsequential for a task, different rules may prescribe the same decision for trial-by-trial responses. In these cases, the significance of utilizing any particular rule over another may be trivial. When decision noise grows significant enough to affect changes to the response outcomes for each trial, different rules may lead to distinctly different decision behavior. Over the course of an experiment, these decision rules may give rise to idiosyncratic data patterns associated with specific rules. In our research, we focus on three simultaneous decision rules: LCJ, LCJ$_c$, and LCJ$_{sym}$ (Klauer & Kellen, 2012; Rosner & Kochanski, 2009).

Multiple Criteria: Multiple Passes

With a simultaneous rule, an observer adopts a decision protocol with which the internal representation is compared with all criterion boundaries simultaneously. No criterion has any kind of priority with respect to the others, but we assume that the means of each criteria maintain their ordinal relation to each other throughout the duration of the experiment. For our development here, we consider $M + 1$ response categories, and we enumerate these categories according to their ordinal positions along the decision axis with the set [1, 2 . . . , M, $M + 1$].

The formal description of the overall response frequencies under this decision rule, as well as the LCJ$_c$ and LCJ$_{sym}$ decision rules, have been described elsewhere for single-pass procedures (Klauer and Kellen, 2012; Rosner & Kochanski, 2009). For an MCR procedure, the *consistent* noise component of the total response variability can be separately considered in describing the subject's rating response. The separate noise components will be given in units of the *SD* of this consistent noise component. Because the observed quantities of response rates and covariances are given relative to the level of consistent noise, we may consider the representational noise in terms of its component terms. For the LCJ decision rule, an observer subtracts the internal representation from the trial sampled criteria. The observer then classifies the representation from stimulus class S_h according to the response category corresponding to the criterion m with the least positive difference $[\mu_{C_m} + c_m\sigma_{C_m}] - [\mu_{S_h} + s_{E_h}\sigma_{E_h} + s_{ext}]$. If the strength of the internal representation exceeds all the trial-sampled criteria, then the observer responds with the highest level of confidence, $M + 1$.

We have already described the trial-by-trial response frequency for LCJ in Equation 3, with the overall response frequencies and the covariances described by Equations 4 and 5, respectively. For the LCJ$_c$ decision rule, observers subtract the trial sampled criteria from the internal representation and classify the internal representation in the category just above the decision boundary with the least positive distance. If all differences are negative, the observer classifies the internal representation with the lowest response category. For a given sample of external noise, the response probabilities are given as,

$$\begin{aligned} &P(R = m + 1 \mid s_{ext}, S_h) \\ &\quad = \int \phi(c_m; \mu_{C_m}, \sigma_{C_m}) \int_{\mu_{C_m}+c_m\sigma_{C_m}}^{\infty} \phi(s_{E_h}; \mu_{S_h} + s_{ext}, \sigma_{E_h}) \\ &\qquad \times \prod_{m' \neq m}\left[1 - \int_{\mu_{C_m}+c_m\sigma_{C_m}}^{\mu_{S_h}+s_{E_h}\sigma_{E_h}+s_{ext}} \phi(c_{m'}; \mu_{C_{m'}}, \sigma_{C_{m'}}) dc_{m'}\right] ds_{E_h} dc_m \end{aligned} \tag{A14}$$

Finally, for LCJ$_{sym}$, observers take the difference between the trial sampled criteria and the internal representation of the stimulus. The observer identifies the decision boundary corresponding to the least absolute value and classifies the trial in the category corresponding to the boundary index if the representation falls short of the boundary, or classifies the trial in the category just above the boundary index if the representation falls about the boundary. That is,

(*Appendices continue*)

$$\begin{aligned}P(R = m \mid s_{ext}, S_h) = {} & \int \phi\left(s_{E_h}; \mu_{S_h} + s_{ext}, \sigma_{E_h}\right) \int_{\mu_{S_h}+s_{E_h}\sigma_{E_h}+s_{ext}}^{\infty} \phi\left(c_m; \mu_{C_m}, \sigma_{C_m}\right) \\ & \times \prod_{m' \neq m} \left[1 - \int_{2\left(\mu_{S_h}+s_{E_h}\sigma_{E_h}+s_{ext}\right)-\left(\mu_{C_m}+c_m\sigma_{C_m}\right)}^{\mu_{C_m}+c_m\sigma_{C_m}} \phi\left(c_{m'}; \mu_{C_{m'}}, \sigma_{C_m}{'}\right) dc\right] dc_m ds_{E_h} \\ & + \int \varphi\left(s_{E_h}; \mu_{S_h} + s_{ext}, \sigma_{E_h}\right) \int_{-\infty}^{\mu_{S_h}+s_{E_h}\sigma_{E_h}+s_{ext}} \phi\left(c_{m-1}; \mu_{C_{m-1}}, \sigma_{C_{m-1}}\right) \\ & \times \prod_{m' \neq m-1} \left[1 - \int_{\mu_{C_{m-1}}+c_{m-1}\sigma_{C_{m-1}}}^{2\left(\mu_{S_h}+s_{E_h}\sigma_{E_h}+s_{ext}\right)-\left(\mu_{C_{m-1}}+c_{m-1}\sigma_{C_{m-1}}\right)} \phi\left(c_{m'}; \mu_{C_{m'}}, \sigma_{C_m}{'}\right) dc_{m'}\right] dc_m ds_{E_h}\end{aligned} \tag{A15}$$

For all the simultaneous rules we have discussed, the overall response frequencies across all trials are then computed as in Equation 4 and the covariance between any two response categories m and m' are described by Equation 5.

Decision noise changes the interpretation of the ROC for the decision models we have presented here. Ostensibly, the ROC intends to reflect the operating performance of the receiver during binary decision tasks for a single criterion positioned according to a specific likelihood ratio. When decision noise is not present, ROC analysis in rating tasks assumes that any stimulus inducing a response from a stricter response class should cumulatively redound to less strict response classes. In this case, the ROC accurately reflects the operating performance at the more lax criteria because every representation classified in the stricter categories would have been classified in the lower confidence categories if the stricter classifications had not been available for report. However, for simultaneous decision rules, classification in every response category depends on the trial-by-trial positions of all criteria, so the traditional interpretation of the ROC does not hold for any HR and FAR pairing.

Appendix B

Variance in Response Frequency

Multi pass methods have utilized measures of agreement (probability of agreement, correlation, and covariance) together with response frequency data to estimate total internal noise in psychophysical tasks (Burgess & Colborne, 1988; Gold, Bennett, & Sekuler, 1999; Green, 1964; Lu & Dosher, 2008). For the purposes of modeling, it may prove useful to state the expected variability of the observed response frequencies. However, in addition to modeling, the variability in response frequencies across multiple passes may also serve as an index of internal noise. The variability of a subject's response for a representation s_{ext}, induced by a given sample of external noise is formulated as a Bernoulli trial conditioned on s_{ext}. We express this quantity as,

$$\begin{aligned}Var[R \equiv m \mid s_{ext}, S_h] &= EV_R\{[(R \equiv m \mid s_{ext}, S_h) - EV_R[R \equiv m \mid s_{ext}, S_h]]^2\} \\ &= P(R = m \mid s_{ext}, S_h) - P(R = m \mid s_{ext}, S_h)^2\end{aligned} \tag{B1}$$

(*Appendices continue*)

On the other hand, the observed variability of response over an entire pass is the expected value of the variance as formulated in the equations above, with that expectation taken over all possible samples of s_{ext}. We express this overall variability as,

$$\begin{aligned} Var[P(R=m\,|\,S_h)] &= EV_{s_{ext}}\{Var[R \equiv m\,|\,s_{ext}, S_h]\} \\ &= \int [P(R=m\,|\,s_{ext}, S_h) - P(R=m\,|\,s_{ext}, S_h)^2]\phi(s_{ext})ds_{ext} \\ &= P(R=m\,|\,S_h) - \int P(R=m\,|\,s_{ext}, S_h)^2\phi(s_{ext})ds_{ext} \\ &= P(R=m\,|\,S_h) - P(R=m\,|\,S_h)^2 - Cov[R_i=m, R_j=m\,|\,S_h] \end{aligned} \tag{B2}$$

We may observe here that the overall variability of response frequencies for a given stimulus class within a multi pass paradigm generally differs from the overall variability of response frequencies when passes do not contain identical noise samples. This difference in variability increases with lower internal to external noise ratios, because the variability of response at a given sample s_{ext} approaches zero as the internal to external noise ratio approaches zero. Furthermore, the variability of this response at a given s_{ext} approaches $P(R = m\,|\,S_h) - P(R = m\,|\,S_h)^2$ as the internal noise increases and overwhelms the external noise. In the case of a high internal to external noise ratio, the index s_{ext} provides no consequential information and does not significantly influence subject performance.

As a consequence of the multi pass paradigm, we see from Equation A16 that $0 \leq Var[P(R = m\,|\,S_h)] \leq P(R = m\,|\,S_h) - P(R = m\,|\,S_h)^2$. Thus, the variance of response frequencies between passes also provides an index of the total internal to external noise ratio for each stimulus class.

Appendix C

Application of Method in Visual Detection Experiment

A psychophysical procedure designed to partition internal noise into representation and decision noise affords at least two meaningful contributions to the study of cognitive processes. First, separating different sources of noise can contribute to our understanding of the functional architecture underlying human decision behavior. Using the MCR procedure allows us to quantify the contributions of these two sources of internal variability to response variability. Second, this framework provides a test of the assumptions broadly used within the extensive literature on signal detection tasks. If decision noise is relatively low in a given signal detection task, then the simpler traditional SDT model suffices to explain the data. However, if decision noise is significant in a given task, then the estimates of sensitivity, response bias, and the conclusions derived from these estimates will be improved by the methods described here. We now describe a confidence rating visual detection experiment in which we analyzed both response frequencies and covariance across multiple passes to assess how decision and representational noise influence task performance.

Method

Procedure. We conducted a Gabor detection experiment in fovea with external noise over multiple passes. Subjects gave confidence ratings on the presence or absence of a Gabor temporally embedded in external noise. Two subjects completed experiments with both three and five rating categories in separate sessions during a day. Each subject completed two consecutive 40-min sessions per day over 5 days with, at minimum, a 15-min break between sessions. Each session consisted of six passes with 100 trials per pass. Corresponding trials across passes contained identical stimulus samples but with randomized stimulus schedules. In total, subjects responded to 500 trials per pass for both three and five response categories after concatenating corresponding sessions across all 5 days. We alternated the order of sessions so that if subjects started the previous day's session using three response categories, they would begin the next day's session with five response categories, and vice versa. In all conditions, the highest category rating corresponded to the highest degree of confidence in the presence of a signal stimulus for each trial and the lowest rating corresponded to the lowest degree of confidence in the presence of a signal stimulus. On each trial, the stimulus (either external noise alone or external noise with the Gabor signal) appeared at the center of the computer monitor with a signal probability of 0.5. A brief auditory cue sounded 133 ms before stimulus onset to minimize effects of temporal uncertainty (Spiegel & Green, 1981). The fixation cross and box disappeared

(*Appendices continue*)

after 664 ms, followed by the stimulus onset. Five stimulus frames consisting of two external noise frames, either a Gabor or blank frame, and two additional external noise frames appeared in sequence for 33 ms each, followed by a blank screen until subjects provided a rating response.

Following each trial response, a trial score (see Table C1) briefly appeared on the screen, followed by the subject's cumulative score for the session. Before the first pass on each day, subjects were instructed to utilize the full range of confidence ratings and to achieve the highest possible score over the course of the experiment. The scoring structure for both low and high response categories is given in Table C1. Subjects took short breaks after each pass of 100 trials.

To select a stimulus contrast, we used an Accelerated Stochastic Approximation Method (Treutwein, 1995) to estimate contrast thresholds before the MCR detection experiment. This adaptive procedure varied the contrast of the Gabor from trial to trial so as to converge on a threshold corresponding to a desired performance level of $d' = 1$ in high external noise using binary (yes/no) responses. Frames of external noise were completely independent across all trials.

Stimuli. We generated all stimuli with a G4 Macintosh computer utilizing Matlab programs with Psychtoolbox extensions (Brainard, 1997; Pelli, 1997). Stimuli appeared on a ViewSonic Professional Series P95f monitor with a refresh rate of 120 Hz and mean luminance of ~50 cd/m^2. A video attenuator modified gray level display by combining voltages of two graphic channels to produce 6,144 distinct gray levels for enhanced contrast (Li, Lu, Xu, Jin, & Zhou 2003). A psychophysical method (Lu & Sperling, 1999) was used to estimate and linearize luminance. Subjects placed their heads in a chin rest to minimize head movement and viewed the stimuli from ~1 m under scotopic lighting conditions.

Signal Gabor targets consisted of a 3.75 cpd sine wave grating oriented 12 degrees to the right of vertical and multiplied by a Gaussian spatial window with a *SD* of 0.44 degrees of visual angle. External noise frames consisted of individual pixels randomly sampled from a Gaussian distribution with 0 mean and a *SD* of 0.33 of the full contrast range. Both Gabors and external noise frames subtended 1.6 × 1.6 degrees of visual angle at the center of the screen. The box within which the stimuli appeared subtended the same visual angle as the target stimuli. The fixation cross subtended 0.12 × 0.12 degrees of visual angle.

Table C1
Payoff Matrix for Rating Detection Task

	Subject response				
	1	2[a]	3[a]	4[a]	5
Signal absent trial	2	1	0	−1	−3
Signal present trial	−3	−1	0	1	2

[a] The response alternatives indicated in gray comprised the payoff structure for the three response categories rating tasks, whereas the entire response range comprised the payoff structure for the five response categories rating task.

Observers. One University of Southern California graduate student as well as the first author participated in the study. Both subjects had normal or corrected to normal vision and both had significant previous experience as subjects in psychophysics experiments.

Data analysis. For each subject and for each rating structure, we collected response frequencies and covariance estimates across all 5 days. We computed both *within category* covariance (covariance between the same rating category across different passes) and *between category* covariance (covariance between different rating categories across different passes). For the purpose of fitting the model to the data, we also estimated the variances of all response rates and covariance estimates (Equations 7 and 8). We fit our data with a corrected law of categorical judgment (LCJ; Rosner & Kochanski, 2009), as well as with complimentary (LCJ$_c$) and symmetrically adjusted (LCJ$_{sym}$) modifications of the LCJ (Klauer & Kellen, 2012), and finally with the classical SDT model (cSDT) without decision noise. For all model fits we used a weighted least-squares cost function with a simplex optimization routine (Nelder-Mead) and assessed parameter fits with a χ^2 statistic. Our cost functions incurred significant penalty if ordinal positioning of candidate mean criterion positions became disordered, if variances fell below zero, or if the encoding noise for signal absent trials exceeded the encoding noise for signal present trials. We reorganized the response frequencies into standardized ROC plots according to the usual method of starting with the highest category rating and cumulatively adding response frequencies to the next strictest response category. We computed covariance estimates for signal absent trials and signal present trials and separately plotted them to more easily distinguish model fits to data. We also computed separate correlation statistics for response frequencies (r^2_{ROC}) and covariance estimates (r^2_{Cov}) because these data do not share a common scale.

Results

Parameter estimates and χ^2 results for all model fits to subject data are found in Table C2. For both subjects, the best fitting decision noise model did not provide significantly better fits to experimental data with three rating categories than the cSDT model without decision noise (subject YZ: $F(2, 3) = 5.6487$, $p = 0.096$; subject CC: $F(2, 3) = 0.7259$, $p > 0.1$). For subject YZ, we found the reduced (cSDT) model fits at $\chi^2 = 7.1811$, $r^2_{\text{ROC}} = 0.99$, $r^2_{\text{Cov}} = 0.96$. For subject CC, these fit statistics were $\chi^2 = 2.5716$, $r^2_{\text{ROC}} = 0.99$, $r^2_{\text{Cov}} = 0.98$.

For the paradigm with five response categories, we found the decision noise model LCJ fit the data better than any other model and significantly better than the cSDT model for both subjects (subject YZ: $F(4, 17) = 6.8171$, $p < 0.01$; subject CC: $F(4, 17) = 10.8981$, $p < 0.001$). Fits for subject YZ with this model were $\chi^2 = 8.5944$, $r^2_{\text{ROC}} = 0.99$, and $r^2_{\text{Cov}} = 0.95$. For subject CC, we found $\chi^2 = 6.6640$, $r^2_{\text{ROC}} = 0.99$, $r^2_{\text{Cov}} = 0.95$.

(*Appendices continue*)

Table C2
Parameter Estimates for Three and Five Response Categories (Joint Model Fits)

		Model parameters (in units of σ_{ext}) and χ^2																			
		2 criteria				Representation				4 criteria								Representation			
Sub	Mod	μ_{C1}	μ_{C2}	σ_{C1}	σ_{C2}	σ_{E0}	σ_{E1}	μ_{S1}	χ^2	μ_{C1}	μ_{C2}	μ_{C3}	μ_{C4}	σ_{C1}	σ_{C2}	σ_{C3}	σ_{C4}	σ_{E0}	σ_{E1}	μ_{S1}	χ^2
YZ	cSDT	−0.8	1.22	—	—	1.47	1.52	1.72	7.18	−1.85	−0.06	1.25	3.29	—	—	—	—	1.48	1.51	1.64	22.38
	LCJ	−0.37	1.03	1.28	1.25	1.02	1.02	1.66	1.51	−1.9	0.31	1.21	3.24	0.92	0.74	0.97	0.38	1.32	1.32	1.7	8.59
	LCJ_c	−0.93	1.22	1.03	0	1.49	1.49	1.7	2.67	−2.1	−0.36	1.11	3.22	1.36	1.09	0.45	0.04	1.36	1.36	1.68	10.88
	LCJ_{sym}	−0.87	1.21	0.9	0	1.48	1.49	1.69	3.03	−1.83	0.29	1.02	3.25	0.23	0.94	0.34	0.02	1.41	1.41	1.7	11.78
CC	cSDT	1.1	1.68	—	—	1.65	1.7	2.52	2.57	−1.19	1	1.57	3.79	—	—	—	—	1.74	2.03	2.77	19.08
	LCJ	1.1	1.7	0	0	1.66	1.75	2.57	2.27	−0.36	0.98	1.52	3.73	1.65	0.07	0.03	0.39	1.74	1.95	2.62	6.66
	LCJ_c	1.05	1.5	0.25	0	1.5	1.62	2.46	1.73	−1.22	0.91	1.41	3.56	1.29	0	0.02	0.72	1.5	1.86	2.51	9.41
	LCJ_{sym}	1.31	1.4	0.05	0.03	1.37	1.41	2.45	1.74	−0.99	0.88	1.36	3.44	1	0	0.02	0.8	1.41	1.8	2.41	10.77

Note. cSDT = classical signal detection theory model; LCJ = law of categorical judgment; c = complimentary; sym = symmetrically adjusted.

Winning model fits for all subjects and response categories are shown in Figure C1. To ensure these fit statistics accurately represented the predictive power of our model, we ran 100 cross validation checks on each of our data sets. For each subject, response condition, and iteration, we sampled (without replacement) 80% of trial stimuli and computed yes rates and covariances from subject responses to those stimuli across passes. After modeling each partial data set, we computed the expected values of yes rates and covariances of each fit to predict the yes rates and covariances of the complimentary portion of each data sample. For subject YZ with five response categories we used LCJ to determine the median $r^2_{\text{ROC}} = 0.97$ and median $r^2_{\text{Cov}} = 0.87$. For subject CC, $r^2_{\text{ROC}} = 0.99$ and median $r^2_{\text{Cov}} = 0.82$ for data with five response categories.

The LCJ dominated the classical SDT model for both subjects when using five rating categories and but not for three categories. We also investigated whether the change in response structure between low and high number of response categories could change the representational features of stimuli. To test this hypothesis, we fit the data from both the three- and five-category rating experiments together with the LCJ model under two distinct assumptions. In the first case, we allowed all parameters to vary independently; thus, permitting representation noise to vary with response structure; in the second case, we assumed that the representational parameters σ_{S_0}, σ_{S_1}, and μ_{S_1} remained identical across response structures. We used an *F* test for nested models to compare these results and found that the extended model did not significantly improve fits over the reduced model for either subject (subject CC: $F(3, 20) = 1.2443$, $p > 0.1$; subject YZ: $F(3, 20) = 0.8803$, $p > 0.1$). From this we conclude that the criterion variability but not representation variability was affected by the larger number of rating categories. We further fit our subject data using the classical SDT model (no decision noise) for the three-category response structure while jointly modeling data from the five-category response structure with the LCJ assuming either completely independent parameters or identical representational parameters. The results again showed no significant improvement using the full model relative to the restricted model (subject CC: $F(3, 22) = 1.1130$, $p > 0.1$; subject YZ: $F(3, 22) = 0.9047$, $p > 0.1$). These parameter fits are listed in Table C3. The fits imply that representational features do not significantly change with a change in response structure from three- to five-category rating tasks.

When we fit the data from our five-category rating task with the classical SDT model with no decision noise, we estimated encoding noise for each of our subjects. For subject CC we estimated encoding noise at 2.03 for signal present trials and 1.74 for signal absent trials (relative to σ_k). For subject YZ we estimated encoding noise for signal present and signal absent trials at 1.51 and 1.48, respectively. In the case of subject CC, estimates of representation parameters are quite similar between LCJ and cSDT. For subject YZ, however, cSDT overestimates encoding noise by about 14% for signal present trials and 12% for signal absent trials.

We also fit data for each subject using five response categories to the LCJ using only yes rates (i.e., without covariance data). For these fits, we retained the decision noise parameter results from the LCJ model and allowed the remaining parameters (criterion positions, representation noise on signal present trials, and the mean of the signal present distribution) to vary. When considering only ROC data, the model estimates each parameter in units of the representation noise of the signal-absent distribution (rather than merely the consistent noise). The results of these fits are shown in Table C4. We recomputed our original parameter estimates for the LCJ fits to full data sets (ROC and covariance data) for each subject in units of the entire representational noise for signal-absent trials. These are shown along with the ROC-only fits for comparison. The estimates for ROC-only and full data sets are nearly identical for both subjects.

The emphasis of this report was to illustrate the sufficiency of the framework to separately estimate contributions of decision and encoding noise in response data from signal detection tasks. Other models may explain this data as well. At the suggestion of one reviewer, we examined the performance of a mixture model (DeCarlo, 2002) according to which signal present trials are drawn

(*Appendices continue*)

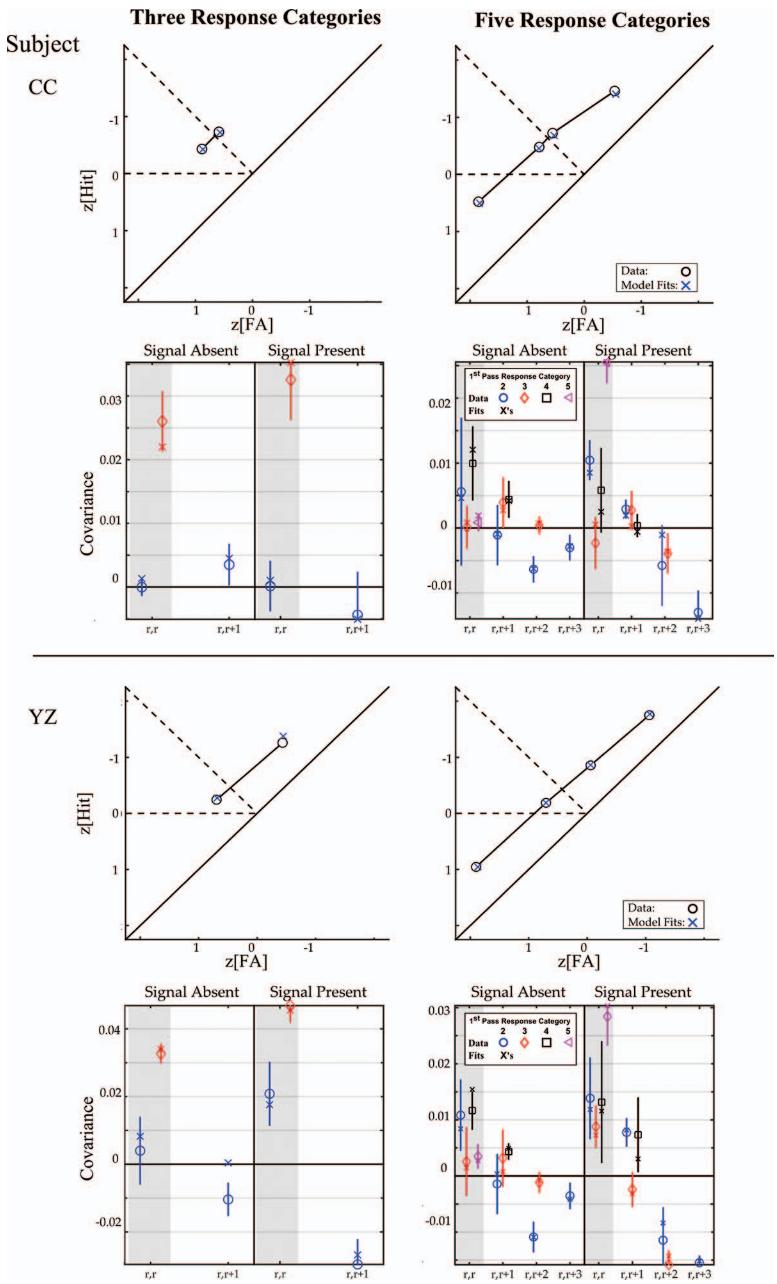


Figure C1. Classical signal detection theory (three response categories) and law of categorical judgment (five response categories) model fits for z-transformed receiver operating characteristic and covariance data for subjects CC and YZ. Covariance graphs: point [r,r] (gray bars) corresponds to within-category covariance, while all other points correspond to between-category covariance. See the online article for the color version of this figure.

(*Appendices continue*)

Table C3
Model Fits to Subject Data of Three and Five Rating Categories With Identical Representation Parameters

	Model parameters (in units of σ_{ext})														
	2 criteria				4 criteria								Representation		
Subject	μ_{C1}	μ_{C2}	σ_{C1}	σ_{C2}	μ_{C1}	μ_{C2}	μ_{C3}	μ_{C4}	σ_{C1}	σ_{C2}	σ_{C3}	σ_{C4}	σ_{E0}	σ_{E1}	μ_{S1}
YZ	−0.75	1.22	—	—	−1.9	0.26	1.24	3.31	0.43	0.71	0.76	0.08	1.43	1.43	1.73
CC	1.13	1.75	—	—	−0.36	0.99	1.52	3.73	1.64	0.07	0.03	0.39	1.73	1.95	2.62

from two underlying distributions depending on whether or not subjects gave an adequate allocation of attention while sampling from each trial (mean distribution of attended trials are given as μ_S, mean of nonattended trials assumed equal to zero[4]). The mixture model assumes that representational variance is equal for both signal present distributions as well as the signal absent distribution. For signal present trials, the portion of trials drawn from each distribution depends on a mixture parameter, λ. For our experiments with five response categories, fits for subject CC were $\chi^2 = 21.02$, $r^2_{\mathrm{ROC}} = 0.99$, $r^2_{\mathrm{Cov}} = 0.90$, whereas for subject YZ we found $\chi^2 = 22.02$, $r^2_{\mathrm{ROC}} = 0.99$, $r^2_{\mathrm{Cov}} = 0.94$. Differences in performance between three and five response category conditions were accounted for with a slight decrease in λ for subject YZ (7% for three vs 5% for five categories) and a more pronounced increase for subject CC (1% for three vs 9% for five categories). Applying the same cross validation testing described earlier, we found median $r^2_{\mathrm{ROC}} = 0.98$ and median $r^2_{\mathrm{Cov}} = 0.82$ for subject CC. For subject YZ, median $r^2_{\mathrm{ROC}} = 0.97$ and median $r^2_{\mathrm{Cov}} = 0.83$. Although these cross validation results for the mixture model are worse than those obtained for the LCJ decision noise model, the performance outcomes are still quite similar.

The reasonably good fits of the mixture model raises the question of whether a decision noise model might misattribute the effects of nondecision mechanisms to decision noise. To addressed this concern, we conducted an additional 250 simulations of an observer operating under the assumptions of a mixture model (assuming λ = 0.05 in line with estimates typical for psychophysical experiments; Green, 1995; Lesmes et al., 2006; Wichmann & Hill, 2001) and fit these data using the LCJ, LCJ_c, and $\mathrm{LCJ}_{\mathrm{sym}}$ decision noise models. Each simulated experiment consisted of six passes with 500 trials in each pass. We perturbed the true generative parameters of our mixture model by randomly sampling from a normal distribution with means matched to the true parameters and *SD* of $0.15\sigma_{\mathrm{ext}}$ to obtain initial guess parameters for our fitting algorithms. The parameters used in the generative mixture model as well as the recovered parameters from each decision noise model are shown in Table C5. Each of the decision noise models accurately estimated the influence of decision noise as nearly zero when fit to data generated from the mixture model. Furthermore, the median parameter estimates of the decision noise models all came very close to the true parameter values of the generative mixture distribution (excluding λ insofar as this parameter does not figure into our decision noise models). The 95% confidence intervals were quite large for the encoding noise parameters, with estimates sometimes reaching into nonsensical values, but this result might be expected when model assumptions fail to describe the mechanisms underlying the data-generative model.

Table C4
Parameter Estimates for LCJ Fit to ROC-Only vs. Full Data Sets for Five Response Categories

		Model parameters (in units of total representational noise on signal absent trials)					
		4 criteria				Representation	
Subject	Data modeled	μ_{C1}	μ_{C2}	μ_{C3}	μ_{C4}	σ_{R1}	μ_{S1}
YZ	ROC	−1.14	0.19	0.72	1.97	1	1.02
	ROC + Cov	−1.15	0.19	0.73	1.96	1	1.03
CC	ROC	−0.19	0.49	0.75	1.85	1.08	1.33
	ROC + Cov	−0.18	0.49	0.76	1.86	1.09	1.3

Note. LCJ = law of categorical judgment; ROC = receiver operating characteristic; Cov = covariance.

While we acknowledge the possibility that alternative elaborations of the SDT model may account for this data, we also noted that our data are consistent with the prediction issued by Benjamin et al. (2013) for ROCs generated from rating scales of different size: if additional criteria results in additional decision noise, then ROCs generated from larger rating scales should fall below ROCs measured with smaller rating scales. In our data, we plotted the best fitting line through zROC data when each subject used both three and five response categories. For both subjects, the yes rates from three-category experiments resulted in points lying above the best fitting line fitted to the data from five response categories (Figure 12).

[4] The mean distribution for unattended trials may be nonzero, but an *F* test for nested models showed no significant improvement over the reduced model. Therefore, we report results for the reduced model only.

(*Appendices continue*)

Table C5
Median Decision Noise Model Parameter Estimates for Simulated Data From Mixture Distributions

	Model parameters (in units of σ_{ext})							
Model type	μ_{C1}	μ_{C2}	σ_{C1}	σ_{C2}	σ_{E0}	σ_{E1}	μ_{S1}	λ
Mixture model	0.07	0.71	—	—	1	1	1.41	0.05
LCJ	0.07	0.66	0	0	1	1.05	1.29	—
LCJ$_c$	0.05	0.65	0	0	1	1.01	1.28	—
LCJ$_{sym}$	0.02	0.58	0	0	0.98	0.99	1.19	—

Note. LCJ = law of categorical judgment; c = complimentary; sym = symmetrically adjusted.

Discussion

The LCJ model fit the response rates and covariance estimates very well in the five category response experiments, and accounted for about 95% of the variability in the data for subject CC and 96% for subject YZ. Even though we computed separate estimates of r^2 for zROC and covariance data, the model still appears to capture the broad data trends. More particularly, the standardized ROC plots of subject CC exhibit a “bowing” shape suggesting greater sensitivity for zROC scores at the center than at peripheral criterion boundaries. Previous studies have predicted this shape for decision noise structures in rating tasks when criteria at extreme boundaries exhibit greater variance than at the more central boundaries (Mueller & Weidemann, 2008; Wickelgren, 1968). These predictions are borne out here.

Qualitative patterns in covariance data also provide some insight into the underlying representation at the decision stage. Greater encoding noise for a specific stimulus type has the effect of depressing the absolute value of covariances globally across all category boundaries but strictly within that stimulus type. On the other hand, greater criterion noise tends to lower the absolute value of covariance for both stimulus types. Additionally, both decision rules and boundary placement influence covariance outcomes. With internal noise at parity, covariance for *within category* estimates will reach a maximum as the response frequency for that stimulus category approaches 0.5 and will decrease with greater or lesser response rates.

Received March 18, 2014
Revision received February 24, 2015
Accepted February 24, 2015 ■