Psychological Review
2012, Vol. 119, No. 3, 457–479

0033-295X/12/$12.00 DOI: 10.1037/a0027727

On the Measurement of Criterion Noise in Signal Detection Theory: The Case of Recognition Memory

David Kellen, Karl Christoph Klauer, and Henrik Singmann
Albert-Ludwigs-Universität Freiburg

Traditional approaches within the framework of signal detection theory (SDT; Green & Swets, 1966), especially in the field of recognition memory, assume that the positioning of response criteria is not a noisy process. Recent work (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008) has challenged this assumption, arguing not only for the existence of criterion noise but also for its large magnitude and substantive contribution to individuals' performance. A review of these recent approaches for the measurement of criterion noise in SDT identifies several shortcomings and confoundings. A reanalysis of Benjamin et al.'s (2009) data sets as well as the results from a new experimental method indicate that the different forms of criterion noise proposed in the recognition memory literature are of very low magnitudes, and they do not provide a significant improvement over the account already given by traditional SDT without criterion noise.

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

Supplemental materials: http://dx.doi.org/10.1037/a0027727.supp

Signal detection theory (SDT; Green & Swets, 1966) represents one of the most successful mathematical models in psychology, with implementations ranging from fields such as perception (e.g., Swets, Tanner, & Birdsall, 1961), recognition memory (e.g., Wixted, 2007), categorization (e.g., Maddox & Bohil, 1998), reasoning (e.g., Dube, Rotello, & Heit, 2010), and decision making (e.g., Pleskac, 2007). Comprehensive introductions to the SDT framework can be found in Macmillan and Creelman (2005) and Wickens (2002).

In very general terms, the SDT framework can be stated as follows: Consider a decision maker who observes enumerable events, defined as trials, in which a specific class of stimulus is presented (target trials) or no stimulus from that class is presented (distractor trials). For every trial, an evidence value concerning its nature (target trial vs. distractor trial) is available to the decision maker. Evidence is defined as a continuous random variable, with distinct probability distributions for the different kinds of trials. The decision maker translates the evidence values into discrete observable responses by establishing response criteria along the evidence scale. These response criteria define the different ranges of evidence values that are mapped onto each available response alternative (e.g., "target" vs. "distractor"). The characteristics of the evidence distributions in conjunction with the established response criteria result in a predicted probability distribution for the observed responses.

One of the most common assumptions in the SDT framework is that the positioning of response criteria along the evidence scale does not randomly vary across trials, that is, that response criteria are stationary unless consciously and purposely shifted by the decision maker. This invariance assumption contrasts with the variability that is assumed for stimulus-related (representational) processes and is somewhat implausible given the difficulties inherent in maintaining and updating response criteria positions (e.g., Verde & Rotello, 2004) and the response dependencies that have repeatedly been observed (e.g., Gilden & Wilson, 1995; Mueller & Weidemann, 2008; Treisman & Williams, 1984), suggesting a trial-by-trial adjustment of response criteria. Also, this invariance assumption contrasts with the assumptions made by more complex models, such as the Ratcliff diffusion model (Ratcliff, 1978), that simultaneously account for accuracy and response time performance and that allow random changes in response-bias parameters on a trial-by-trial basis as well (see also Ratcliff & Starns, 2009). Despite its plausibility, the variability of response criteria—usually referred to as *criterion noise*—has proved to be extremely difficult to assess empirically in the SDT framework given that its inclusion almost invariably results in an unidentified model (Mueller & Weidemann, 2008; Rosner & Kochanski, 2009). Figure 1 provides a generic depiction of the SDT model in the presence and absence of the criteria invariance assumption. The terms *criterion noise* and *criteria variability* are used interchangeably.

Note that the notion of criteria variability that is addressed here concerns random changes in criteria positioning that result from

This article was published Online First April 2, 2012.

David Kellen, Karl Christoph Klauer, and Henrik Singmann, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

The research reported in this article was supported by Deutsche Forschungsgemeinschaft Grant Kl 614/32-1 to Karl Christoph Klauer. We thank Aaron Benjamin for providing his original data. Additionally, we thank Aaron Benjamin, Larry DeCarlo, Ken Malmberg, and John Wixted for valuable comments that led to improvements in both substance and clarity.

Correspondence concerning this article should be addressed to David Kellen, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany. E-mail: david.kellen@psychologie.uni-freiburg.de

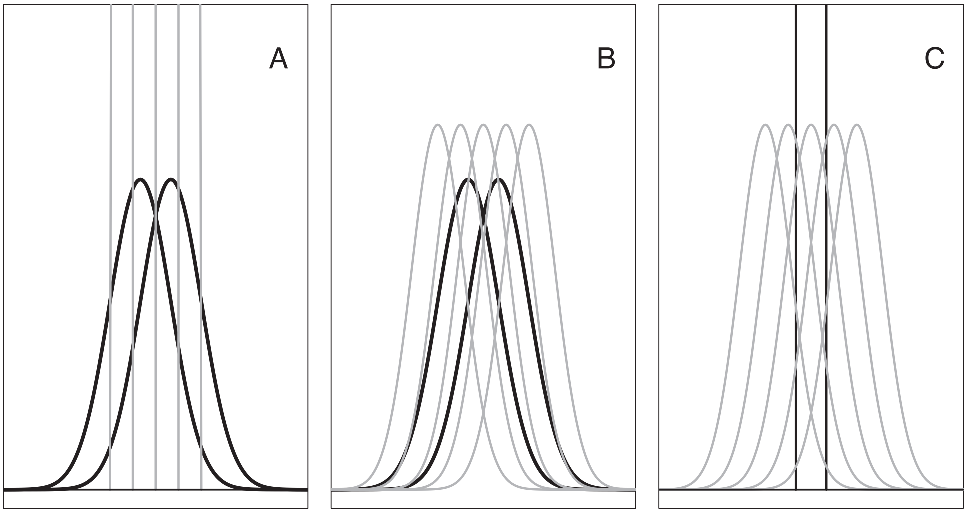

Figure 1. General illustration of the signal detection theory model under different assumptions regarding stimulus evidence and response criteria. The black lines depict the evidence for the different stimulus classes, whereas the gray lines depict the different response criteria. In Panels A and B, stimulus evidence is assumed to vary according to defined probability distributions, whereas in Panel C, it is assumed to have no variance. Regarding response criteria, in Panel A, they are assumed to be fixed to specific values, whereas in Panels B and C, they are assumed to vary according to defined probability distributions.

the latter being assumed to be a noisy process, so that discussions regarding its evaluation and plausibility do not encompass the changes in response criteria (or the absence of such changes) that are observed when participants are tested with different types of stimuli or in different test contexts (e.g., Benjamin & Bawa, 2004; S. D. Brown & Steyvers, 2005; Morrell, Gaitan, & Wixted, 2002; Singer & Wixted, 2006; Stretch & Wixted, 1998). These two kinds of criterion differences are related (for a detailed discussion, see Benjamin, Diaz, & Wee, 2009) but ultimately concern distinct phenomena (stimulus or context-based changes vs. random changes). To avoid confusion, changes in response criteria caused by the testing of different types of stimuli or different test contexts are referred throughout this article as *criteria shifts*.

The presence of criteria variability can severely distort the description of observed performance if left unaccounted (see Benjamin et al., 2009; Malmberg & Xu, 2006; Mueller & Weidemann, 2008; Ratcliff, McKoon, & Tindall, 1994; Treisman, 1987), a possibility that has serious implications for the various fields that rely on SDT but especially for the field of recognition memory in which most of the work produced relied and continues to rely on receiver operating characteristic (ROC) functions obtained through the use of a confidence rating scale (for reviews, see Wixted, 2007; Yonelinas & Parks, 2007). ROC functions consist of the sets of responses (hit and false alarm rates) that are predicted for differently positioned response criteria (see Wickens, 2002). Benjamin et al. (2009) and Mueller and Weidemann (2008) demonstrate how criteria variability can distort ROCs in different manners, leading to potentially erroneous interpretations of results.

The present article is organized as follows: First, the challenges for the SDT framework raised by the studies of Balakrishnan (1998, 1999) and Van Zandt (2000) are described, challenges that motivated a reassessment of the assumption of criteria invariance and the development of models with the specific goal of measuring criterion noise. Moreover, the two SDT accounts dedicated to the measurement of criterion noise in recognition memory, namely the models proposed by Mueller and Weidemann (2008) and Benjamin et al. (2009), are reviewed and evaluated. Their advantages and limitations are thoroughly discussed as they provide guidelines for what SDT accounts of criterion noise should achieve. An alternative account to the one by Benjamin et al. is then proposed along with an experiment that exemplifies its use. Furthermore, the sensitivity of both Benjamin et al.'s approach and the present one is evaluated, revealing the strong limitations that the measurement of criterion noise in the context of recognition memory is subjected to.

Relating Criterion Noise and Violations of SDT: The Decision Noise Model (DNM)

Although the issue of criteria variability in SDT was extensively studied in the past (e.g., Treisman, 1987; for a review, see Benjamin et al., 2009), it received very little attention in the field of recognition memory. The impact of criterion noise on individuals' performance became a concern in this field when it assumed a central role in the explanation of supposed violations of SDT assumptions reported in the field of visual perception by Balakrishnan (1998, 1999) and in recognition memory by Van Zandt (2000).

Balakrishnan (1998, 1999) investigated one of the fundamental assumptions of SDT, that individuals evaluate the evidence value of presented stimuli and can assess the likelihood of a stimulus belonging to a specific stimulus class given the perceived evidence. When responses are collected by means of a bipolar rating scale, the relative proportions of rating response *m* for stimulus classes A and B can be used to calculate the log-likelihood ratio of a stimulus belonging to either class given rating response "*m*," $LL = log\left(\frac{P(stimulus = A|\text{“}m\text{”})}{P(stimulus = B|\text{“}m\text{”})}\right)$. Values below 0 indicate that it is more likely that a stimulus belongs to class B, whereas values above 0 indicate a greater likelihood that the item belongs to class A. When $LL = 0$, it is equally likely that the stimulus belongs to either class. When considering an unbiased decision maker with an above-chance performance, *LL* should be above 0 for response ratings associated with classification "A" and should be below 0 for response ratings associated with classification "B." In consequence, for an unbiased decision maker, the two response ratings adjacent to the scale midpoint (e.g., in a 6-point scale, these would be the ratings "3" and "4") should define an interval that includes the point of equal likelihood ($LL = 0$). When responses are biased, this interval should not include the value 0, as for both response ratings adjacent to the scale midpoint $LL < 0$ or $LL > 0$, depending on the direction of the response bias. The latter prediction stems from the notion that when responses are biased, the classification criterion is attached to an evidence value for which one stimulus class is more likely than the other.[1]

Balakrishnan (1998, 1999) tested this prediction of SDT using a perceptual task in which participants had to correctly classify one of two possible stimuli (two horizontal lines of different lengths) as either "short" or "long" using a confidence scale. Response bias was manipulated by means of either a base rate or a payoff manipulation. Results indicated that despite the bias manipulation and the indication of biased responding by standard SDT measures, the interval defined by the least confidence ratings still included the point of equal likelihood, suggesting that the responses were still unbiased. Balakrishnan claimed that these results demonstrate that the classical SDT framework is fundamentally flawed and that alternative modeling approaches that contemplate the judgments' response times need to be considered to obtain an appropriate characterization of individuals' performance. Despite several criticisms of Balakrishnan's claims (Kornbrot, 2006; Treisman, 2002; but see Balakrishnan & MacDonald, 2002), no SDT model was proposed that could account for the observed data.

In the field of recognition memory, Van Zandt (2000) tested the closely related assumption that when manipulating response bias, the different response criteria are established on a common evidence axis. If this assumption holds, then the ROCs obtained for each bias condition should perfectly overlap. Participants were tested across five bias conditions, either by manipulating item base rates or payoffs. In each condition, participants gave their responses using a confidence scale. The obtained ROCs did not overlap as expected, and the parameter estimates indicated a consistent change in target variance. More specifically, as the bias to respond "Old" increased, the variance of evidence values for targets became significantly smaller. Similarly to Balakrishnan (1998, 1999); Van Zandt (2000) argued that these violations of SDT assumptions indicate the need to move away from this classical approach and toward more complex models that contemplate response times as well, namely sequential sampling models (e.g., Ratcliff, 1978). Overall, these results from the fields of perception and memory represented a major challenge for the SDT approach.

To account for these violations of SDT, Mueller and Weidemann (2008) proposed the DNM, a SDT model that distinguishes three sources of variability: stimulus, representational, and criteria variability.[2] Let *S* indicate stimulus class, with $S = t$ used to denote target stimuli and $S = d$ distractor stimuli.

In the DNM, the stimulus distributions are defined as Gaussian distributions whose parameters (means and standard-deviations) *exactly correspond* to the different stimulus classes' objective characteristics:

$$f_S \sim N(\mu_S, \sigma_S) \tag{1}$$

For example, Mueller and Weidemann (2008) reported an experiment in which participants had to indicate the category that presented stimuli (sets of asterisks) belonged to, with both stimulus categories being defined by Gaussian distributions with established parameter values (e.g., $\mu_t = 54$, $\sigma_t = 5$, and $\mu_d = 46$, $\sigma_d = 5$).

Representational variability describes the noisy processes that operate during stimulus encoding and introduce additional variability to the cognitive representation of the stimulus. By definition, this variability is assumed to be unbiased and normally distributed, so $f_{rep} \sim N(0, \sigma_{rep})$. The convolution of the stimulus and representational variability distributions results in the distributions of evidence for stimuli (e_S) that characterize the SDT model:

$$f_{es} \sim f_S \otimes f_{rep} \tag{2}$$

where ⊗ denotes the convolution operator.

Criteria variability describes changes of criterion positions across trials, which are assumed to be independent, each following a Gaussian distribution. This variability can be further subdivided into classification noise (the variability of the binary classification criterion, which sets the boundary between category A and category B responses) and confidence noise (the variability of confidence criteria that delimits the response regions on the confidence scale). Concerning the parameters for the response criteria distributions, two variance parameters are assumed, one assigned to the classification criterion (σ_{class}) and the other to the confidence criteria (σ_{conf}). Also, Mueller and Weidemann (2008) additionally assumed that confidence criterion means are symmetrically distributed around the classification criterion (see Mueller & Weidemann, 2008, Appendix A).

In the DNM, the perceived evidence values are compared to response criteria in a sequential and conditional manner: Let c_m with $m = -M, \ldots, -1, 0, 1, \ldots, M$ denote the $(2M + 1)$ response criteria that map perceived evidence values onto confidence responses (R) on a $(2M + 2)$-point rating scale. In this scale, $R = 1$ and $R = 2M + 2$ correspond to judgments with maximum confidence that an item is a distractor or a target, respectively.[3]

First, the evidence provided by a stimulus is compared to the classification criterion c_0. If the evidence value is smaller than c_0, then it is compared sequentially with nominally below criteria

[1] Additional conditions need to hold to make these predictions, such as that the interval between response criteria adjacent to the criterion responsible for the binary classification needs to be rather small (see Balakrishnan, 1999). The discussion of these conditions is beyond the scope of this article.

[2] Mueller and Weidemann (2008) used the terms "distal stimulus distribution" and "perceptual variability." Instead, we use the terms "stimulus distribution" and "representational variability," respectively.

[3] Although we only address the model for confidence scales with an even number of points, the model can be adapted to scales with an odd number of points (see Mueller & Weidemann, 2008, p. 491).

(e.g., c_{-1}, c_{-2}, and so forth) until a criterion that is smaller than the evidence is observed, producing the rating response corresponding to that criterion. If the evidence is lower than all sampled criteria, then response rating 1 is given. On the other hand, if the evidence is larger than c_0, then the evidence is sequentially compared with the nominally above criteria (e.g., c_1, c_2, and so forth) until a criterion value that is larger than the evidence is met, leading to the rating response associated with that criterion. When the evidence value is larger than all sampled criteria, response rating ($2M + 2$) is produced.

The probability of a response rating i given the presentation of a stimulus from class S is given by:

$$P(R = 1|S) = \int f_{eS}(x) \prod_{m \leq 0} (1 - F_{c_m}(x))\, dx$$

$$P(R = i|S) = \int f_{eS}(x) F_{c_{i-M-2}}(x) \prod_{m=i-M-1}^{0} (1 - F_{c_m}(x))\, dx \qquad \text{for } 2 \leq i \leq M + 1$$

$$P(R = i|S) = \int f_{eS}(x)(1 - F_{c_{i-M-1}}(x)) \prod_{m=0}^{i-M-2} F_{c_m}(x)\, dx \qquad \text{for } M + 2 \leq i \leq 2M + 1$$

$$P(R = 2M + 2|S) = \int f_{eS}(x) \prod_{m \geq 0} F_{c_m}(x)\, dx$$

where

$$F_{c_m}(x) = \int_{-\infty}^{x} f_{c_m}(z)\, dz.$$

The specification of these three sources of variability (stimulus, representational, and criterion variabilities) provides a comprehensive characterization of processes within the SDT framework that can be used to explain results that so far were considered to be incompatible with SDT. Unfortunately, this detailed account comes with a cost, as in its original form the model is not identified, leading to the need of imposing parameter restrictions. The model becomes identified by setting the representational variability parameter σ_{rep} equal to zero. A consequence of this restriction is that the evidence distributions directly correspond to the stimulus distributions, which are in turn specified by the experimental settings.

This state-of-affairs can be relatively inconsequential for some instances but can be problematic in others. For instance, for the experiment reported by Mueller and Weidemann (2008), as well as Balakrishnan's (1998, 1999) data sets, the stimulus distributions are fully specified a priori, meaning that restriction of the representational variability parameter to zero still does not compromise the identifiability of the ratio between classification and confidence noise, a central aspect in the account of Balakrishnan's results. A different picture emerges in the case of the recognition memory data of Van Zandt (2000). Items presented in research on recognition memory are merely classified as previously studied or not studied, meaning that there is no a priori information on within-category variability that can be assigned to the stimulus distributions given that the stimulus variability is a latent characteristic. Contrary to the case of perception, in which it is possible within certain limits to describe the relationship between stimuli and their evoked internal responses (e.g., Lu & Dosher, 2008), for higher-order cognitive domains such as recognition memory, there is no ready manner in which to establish a link between stimuli and internal evidence values. The solution found by Mueller and Weidemann for these cases was to fix the stimulus means to arbitrary means (e.g., $\mu_d = 45$ and $\mu_t = 54$) and to assume extremely low, virtually inexistent standard deviations (e.g., $\sigma_d = \sigma_t = .01$). These restrictions effectively eliminate any overlap between evidence distributions and concentrate most probability mass within a small interval, and approximate the use of Dirac delta functions as evidence distributions (δ_S). The Dirac delta function is a density function that places all probability mass on a single point, therefore not having any variance or spread (see Rosner & Kochanski, 2009). The restriction of the evidence distributions to single points leads to a simplified model:

$$P(R = 1|S) = \prod_{m \leq 0} \theta_{c_m},$$

$$P(R = i|S) = (1 - \theta_{c_{i-M-2}}) \prod_{m=i-M-1}^{0} \theta_{c_m} \text{ for } 2 \leq i \leq M + 1,$$

$$P(R = i|S) = (1 - \theta_{c_{i-M-1}})(1 - \theta_{c_0}) \prod_{m=1}^{i-M-2} \theta_{c_m} \qquad \text{for } M + 2 \leq i \leq 2M + 1,$$

and

$$P(R = 2M + 2|S) = (1 - \theta_{c_0}) \prod_{m \geq 1} \theta_{c_m}$$

where

$$\theta_{c_m} = \begin{cases} \Phi\left(\dfrac{\mu_{c_m} - \mu_S}{\sigma_{c_m}}\right) & m \leq 0 \\ \Phi\left(\dfrac{\mu_S - \mu_{c_m}}{\sigma_{c_m}}\right) & m > 0 \end{cases}$$

with $\Phi(\cdot)$ denoting the cumulative density function of the standard normal distribution. A depiction of the restricted DNM model is provided in Panel C of Figure 1. Parameter θ_{c_0} designates the probability that the sampled classification criterion is larger than a stimulus evidence value. Each parameter θ_{c_m} with $m > 0$ quantifies the probability that confidence criterion m is smaller than a stimulus evidence value. Conversely, parameters θ_{c_m} with $m < 0$ each correspond to the probability that confidence criterion m is larger than the stimulus evidence value.

Although it specifies a well differentiated set of processes that underlie individuals' performance, the particular DNM that is ultimately fitted to the data thereby assumes that performance is driven by evidence-response mapping processes. Whereas traditional SDT assumes that there is evidence variability within stimulus classes and no variability in evidence-response mapping processes, the restricted DNM adopts the completely opposite approach as the only parameters that are to be estimated concern response criteria. In fact, the restricted DNM becomes a discrete-state model that can be specified within the multinomial processing tree model class (Klauer, 2010).

The restricted DNM was fitted to data from the experiment reported by Mueller and Weidemann (2008), as well as from the previous studies by Balakrishnan (1998, 1999) and Van Zandt (2000), providing a satisfactory account. Whereas for the first two cases the evidence distributions were fixed to correspond to the respective stimulus distributions, in the latter case the evidence distributions were set to arbitrary values that effectively assumed no overlap, as discussed previously. As predicted by Mueller and Weidemann, when criteria variability exists and classification noise is much smaller than confidence noise, supposed violations of SDT principles such as the ones reported by both Balakrishnan and Van Zandt can effectively be described by the model. Although the model restrictions do not provide a joint assessment of the variability of both mnesic and response processes, the overall results demonstrate the important role that response processes can have on individuals' performance and the risks these processes pose in terms of model performance if unaccounted for.

Despite the merits of the DNM, it is important to highlight its shortcomings: First, the model cannot provide separate measures for the variability of stimulus and response processes, as one of them has to be fixed a priori. The account provided by Mueller and Weidemann (2008) precludes the measurement of mnesic processes, as the data are described solely by means of response variability processes. Although the original goal of the DNM was to demonstrate how response processes could account for apparent violations of SDT principles, the model is ultimately not able to provide an account of how the different memory and response processes contribute to the observed responses, limiting its future use.

Furthermore, the validity of resulting parameter estimates is not fully ascertained, as differences in classification and criterion noise estimates can be obtained even in the absence of any kind of criteria variability. Consider a recognition memory experiment in which individuals respond to 100 target trials and 100 distractor trials using a 6-point scale. Given the unidentifiability of DNM, we follow Mueller and Weidemann (2008) and fix target and distractor values, assuming that there is no stimulus-related variability (fix μ_t and μ_d to arbitrary values, with $\sigma_t = \sigma_d = 0$). In this context, consider an individual data set, whose response frequencies directly follow (no sampling variability introduced) an equal-variance signal detection model ($\mu_t = 1$, $\mu_d = -1$, $\sigma_t = \sigma_d = 1$) with symmetrical, equally spaced response criteria ($c = \{-1.50, -0.75, 0.00, 0.75, 1.50\}$) and without criterion noise.

For such data, the restricted DNM (with $\mu_t = 54$, $\mu_d = 45$) provides a virtually perfect fit ($G^2(5) = 0.08$), with maximum likelihood parameter estimates for classification and confidence noise standard deviations of 4.50 and 7.54, respectively. Furthermore, restricting the classification and confidence noise parameters to be equal results in a significant increase of badness of fit ($\Delta G^2(1) = 6.29$, $p < .05$).

If the variances for target and distractor evidence distributions differ, the differences between criterion parameter estimates become even more extreme. Consider the previous data-generating model with the only change that $\sigma_t = 1.25$, a value close to the ones typically encountered in the literature (e.g., Ratcliff et al., 1994). The goodness-of-fit deteriorates, with $G^2(5) = 1.52$ ($p = .91$), with estimates of classification and confidence noise standard deviations of 5.00 and 9.11, respectively. Again, fixing both criterion noise parameters to be equal leads to a statistically significant detriment of the model's goodness-of-fit ($\Delta G^2(1) = 7.71$, $p < .01$).

This means that for cases in which the data-generating models are as simple as they can be within the SDT framework, with no criteria variability being present, the restricted DNM parameter estimates indicate (statistically significant) differences between classification and confidence noise, differences that are similar to the ones reported by Mueller and Weidemann (2008). This result raises questions regarding the validity of the parameter estimates: If such differences between classification and confidence noise result from simple examples in which no criteria variability is present at all, then it is not clear how much trust one can place on parameter estimates obtained from real data sets whose generating processes are unknown and in all likelihood more complex. To make matters clear: The issue here is not the large values of criteria variability, as they result from model restrictions that force the observed results to be accounted by criteria variability processes, but the differences in the parameter estimates of classification and confidence noise that emerge even in cases where no criterion noise is present.

Overall, the account of recognition memory data provided by the restricted DNM has several shortcomings: Although it provides an explanation for the apparent violation of SDT assumptions and highlights the importance of response-related processes, the model cannot separate the contributions of representational (e.g., mnesic, perceptual) and response variabilities, and its parameter estimates do not seem to provide a valid account of data-generating processes defined within the SDT framework. Still, the model's merits and shortcomings are informative regarding the features that are expected from a SDT model that can account for the variability of response criteria.

Measuring Both Mnesic and Response Variability: The Ensemble Recognition Approach

In an extensive and careful review of the literature, Benjamin et al. (2009) discussed the theoretical importance of SDT models incorporating criterion noise and proposed a recognition memory paradigm that allows for the measurement of the variability in both mnesic and response processes. The proposed paradigm, termed the *ensemble recognition task*, consists of the presentation of test items in groups or ensembles of variable size, with items within one group being either all old or all new. Assuming that individuals integrate the evidence provided by the different items within an ensemble into a single evidence value (consequently affecting the shape of the target and distractor distributions), and assuming furthermore that criterion noise is unaffected by differences in ensemble size, it becomes possible to obtain separate estimates of these two sources of variability. Three models with different information integration rules were considered by Benjamin et al.: An averaging rule model, a summation rule model, and a maximum rule model (the OR model). The first two models represent possible ways of integrating evidence (thus permitting the estimation of criterion noise), whereas the third assumes that the items within an ensemble are evaluated separately. The measurement of the variability in both mnesic and response processes becomes possible because the information integration rules cause the evidence distributions to change across ensembles in a very specific manner (see Figure 2 for a depiction of the averaging model). These predicted changes permit the estimation of criterion noise, as the presence of criteria variability will affect them.

Restricted versions of these models were also considered, by assuming the absence of criterion noise, and/or that mean response criteria did not shift across the different ensemble sizes. The latter restriction stems from previous work that shows that individuals do not shift their response criteria when tested with different types of items or in different test conditions, despite the fact that those changes would be beneficial and that individuals are encouraged to shift them (e.g., Benjamin & Bawa, 2004; Morrell et al., 2002; Singer & Wixted, 2006; Stretch & Wixted, 1998).

The model with the best performance, namely a model that assumes an averaging rule and that mean response criteria do not shift across ensembles, suggests that criterion noise variability is extremely high when compared to the variability of target and distractor distributions. Accordingly, the descriptions of underlying processes that have been obtained so far by means of SDT models without criterion noise would be severely distorted.

The proposed paradigm and its associated models represent an important contribution to the field of recognition memory, as this is the first approach in the field that effectively allows for the variability of memory and response processes to be jointly estimated. The ability to account for possible distortions caused by criterion noise represents a major advantage, having the potential to provide new insights regarding the underlying processes and leading to reinterpretations of several phenomena that have been extensively studied in this field (for reviews, see Malmberg, 2008; Wixted, 2007; Yonelinas & Parks, 2007).

Nevertheless, there are several issues with the model analyses conducted by Benjamin et al. (2009) that question the reported conclusions. The issues are as follows: (1) The restriction of response criteria across ensembles is problematic given its implications in terms of model predictions, (2) the evidence for criterion noise provided by Benjamin et al. is heavily influenced by these response criteria restrictions, (3) the inspection of the parameter estimates for these restricted models indicate their implausibility, and (4) the data reported by Benjamin et al. are not undisputedly in favor of models that include criterion noise.

SDT Models for the Ensemble Recognition Task

Let h denote the ensemble size condition, n_h the corresponding ensemble size, and $c_{h,i}$ denote the ith response criteria (with $i = 1, \ldots, I$) in ensemble condition h. Note that in the absence of criterion noise, parameters $c_{h,i}$ denote the position of the response criteria, whereas in the presence of criterion noise, they denote their mean positions. Also, note that the response criteria are based on confidence rating responses from a $(I + 1)$-point scale, with confidence levels separated by response criteria. Furthermore, let (μ_t, σ_t) and (μ_d, σ_d) represent the mean and standard deviation for the target and distractor evidence distributions, respectively, and σ_c the standard deviation of the response criteria (the criterion noise). There are two important differences from the DNM: (a) The parameters of the evidence distributions refer to the representational processes and are not determined by objective character-

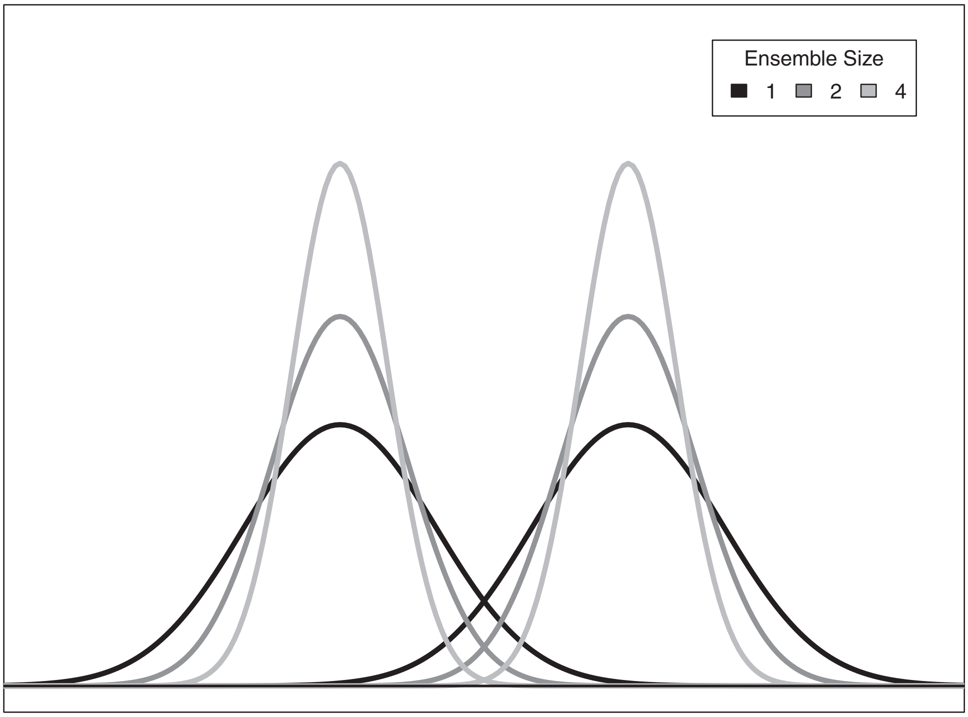

Figure 2. Depiction of the changes in the ensemble evidence distributions across ensemble sizes, as predicted by the averaging model.

istics of the stimuli; and (b) the response criteria are not evaluated sequentially, and only one criteria variability parameter is assumed. Without loss of generality, it is assumed that $\mu_d = 0$ and $\sigma_d = 1$. Note that μ_t, $\sigma_c \geq 0$, and $\sigma_t > 0$. Finally, let R denote the confidence rating response, with $R = 1$ and $R = I + 1$ denoting the maximum confidence responses that the stimulus is a distractor and a target, respectively.

Two kinds of parameter restrictions were considered: (1) that criterion noise is non-existent ($\sigma_c = 0$) and (2) that (mean) response criteria are fixed across ensemble size conditions ($c_{1,i} = c_{2,i} = \ldots = c_{H,i}$, for all i).

The averaging model is defined as follows:

$$P(R \geq i + 1|S) = \Phi\left(\frac{\mu_S - c_{h,i}}{\sqrt{\frac{\sigma_S^2}{n_h} + \sigma_c^2}}\right) \qquad (3)$$

The summation model:

$$P(R \geq i + 1|S) = \Phi\left(\frac{n_h\mu_S - c_{h,i}}{\sqrt{n_h\sigma_s^2 + \sigma_c^2}}\right) \qquad (4)$$

Unlike in Equation 3, when (mean) response criteria are not allowed to shift across ensemble sizes, the mean of the distractor distribution (μ_d) in Equation 4 cannot be fixed to 0 without loss of generality. The reason for this stems from the fact that for the summation model, increases in ensemble size lead to changes in the positioning of both target and distractor distributions relative to the response criteria. If one assumes that the response criteria are constant across ensembles, the value to which μ_d is fixed determines the model's predictions, as further explained below. This problematic issue led Benjamin et al. (2009) to not consider a summation model in which mean response criteria do not shift across ensemble sizes (although a model in which response criteria shift by a factor of n_h across ensemble sizes is considered in their Appendix C).

Note that if one assumes that criterion noise is non-existent ($\sigma_c = 0$) and that no restriction is imposed on response criteria, then the averaging and summation models become the same model (see Benjamin et al., 2009, p. 96).

Given that in this case, both averaging and summation rules are indistinguishable, this restricted model is referred to as the *integration model*.

The OR model precludes the existence of criterion noise and assumes that instead of integrating the evidence provided by the items in the ensembles, individuals evaluate items separately, responding according to the maximum evidence value provided by a single element:

$$P(R \geq i + 1|S) = 1 - \Phi\left(\frac{c_{h,i} - \mu_S}{\sigma_S}\right)^{n_h} \qquad (5)$$

Exploring and Adjusting the Parameterization of the Summation Model

As previously discussed, Benjamin et al. (2009) did not consider a summation model that assumes fixed mean response criteria across ensemble sizes. This means that among the candidate models, the averaging model was the only model with criterion noise in which the *response-criteria shift restriction* was evaluated. The reason for this omission rests on the fact that for the summation model increases in ensemble size lead to changes in the positioning of both target and distractor distributions relative to the response criteria, so that when response criteria are not allowed to shift across ensembles, the value to which μ_d is fixed has a critical role in determining the model's predictions.

If $\mu_d = 0$, then the model predicts positive shifts (by a factor of n_h) of the target distribution while the distractor distribution remains stationary. Also, both distributions' variances increase (which occurs for this and the other parameterizations). This leads to the prediction that an increase of ensemble size should result in an increase of hits and the extremity of their associated ratings, but also in a greater proportion of false alarms as well as more extreme confidence ratings in distractor rejection, given the increase in distractor variance.

On the other hand, if it is assumed that μ_d is negative, as it is when $\mu_t = \frac{\mu}{2}$ and $\mu_d = -\frac{\mu}{2}$ (where μ is the distance between the means of the distributions), then increases in ensemble size would result in a symmetrical shift of both distributions, leading to an increase of both hits and correct rejections, as well as a greater extremity of their associated confidence ratings.

One can also assume that μ_d is fixed to some arbitrary positive value, which would translate into an increase in the familiarity for both distributions as ensemble size increases, leading to greater proportions of both hits and false alarms, as if participants' responses became increasingly liberal.

Among these three possibilities for fixing μ_d, the use of a negative value, symmetrical to μ_t can be considered as the most plausible, as it predicts that larger ensembles lead to greater accuracy for both studied items and distractors as well as more extreme confidence ratings associated to the responses given, which is consistent with computational models of memory that assume that increases in the evidence available lead to a greater differentiation between targets and distractors (e.g., Criss & McClelland, 2006). We include the summation model with this specific parameterization of target and distractor distributions in the set of candidate models, as this allows the model to be tested (in a relatively plausible manner) when assuming fixed response criteria means across ensemble size. Again, note that this parameterization issue only exists when response criteria are restricted.

Same Versus Different Response Criteria Across Ensemble Sizes

As previously mentioned, in parallel to the question of whether criterion noise has an important role in adequately describing recognition memory performance, Benjamin et al. (2009) tested the possibility that response criteria do not *shift* across ensemble sizes, a possibility that stems from previous findings in the literature indicating the individuals' reluctance to shift response criteria positions in different test conditions or for different classes of test items (e.g., strong and weak targets), even when they are encouraged to do so (e.g., Benjamin & Bawa, 2004; Morrell et al., 2002; Singer & Wixted, 2006; Stretch & Wixted, 1998). It is important to note that all the SDT models considered that provide estimates for criterion noise can do so *even when* possible differences in response criteria are permitted. This means that imposing restrictions on response criteria is not necessary for the assessment of criterion noise values. The restriction of response criteria certainly leads to more parsimonious models, but the parsimony is not without costs, as explained next.

Restricting (mean) response criteria to not shift across ensembles leads to undesired consequences in terms of model predictions: Benjamin et al. (2009) pointed out the problems that this restriction raises for the summation model, but they missed the problems that it also entails for the averaging model. In the averaging model, changes in ensemble size only affect the variances of target and distractor distributions, decreasing them as ensemble size increases. Not allowing response-criteria shifts leads to the prediction that unless $c_{1,1} \geq \mu_d$ and $c_{1,I} \leq \mu_t$ (which would mean that for all ensemble sizes, at least 50% of the target and distractor trials are expected to be correctly classified with maximum confidence), the proportion of correct responses made with maximum confidence *decreases* as ensemble size increases. In other words, better evidence leads to less confident responses. For example, consider averaging model with parameters $\mu_t = \sigma_t = 1$, $\sigma_c = 0$, and $c_{h,I} = 1.50$ for all h. The expected proportions of maximum confidence hits (rating $I + 1$) for ensembles of size 1, 2, and 4 are .31, .24, and .16, respectively. This strong prediction is not only implausible, as it not only goes against previous findings in the literature that show increases in confidence along with increases in dicriminability (e.g., Ratcliff et al., 1994), but it is also inconsistent with the data reported by Benjamin et al. (see Benjamin et al., 2009, Table 1). Unfortunately, unlike in the case of the summation model, these predictions cannot be dealt with by an alternative parameterization of the averaging model. In addition, note that criterion noise counteracts the prediction of less confident responses. If $\sigma_c = 0.5$ for the example above, the expected proportions of maximum confidence hits are .33, .28, and .24 for ensemble sizes 1, 2, and 4, respectively. If $\sigma_c = 1$, then the expected proportions are .36, .34, and .33. When fitting data that are inconsistent with this model's prediction—which is the case with Benjamin et al.'s data sets—the criterion noise parameter might be inflated to attenuate the discrepancies between model predictions and observations, with the possibility of distortions in additional parameters not excluded. Given that this inflation would result from implausible model predictions enforced by the response-criteria shift restriction, it is perhaps better to focus the estimation of criterion noise on cases where no restriction is imposed, as the restriction will most likely result in the contamination of parameter estimates for the above reasons.

The inflation of the criterion noise parameter due to the response-criteria shift restriction is further exemplified with a simulation exercise: Consider the averaging model without criterion noise, with parameters $\mu_t = 1$ and $\sigma_t = 1$. Additionally, assume that participants have different response criteria across ensembles. To avoid choosing arbitrary changes in response criteria, let us assume that response criteria change in a principled manner, namely according to predefined likelihood ratios, a hypothesis that finds some support within the SDT literature (e.g., Glanzer, Hilford, & Maloney, 2009; Hautus, Macmillan, & Rotello, 2008). The log-likelihood ratio values adopted for the response criteria are $\{-0.80, -0.30, 0.00, 0.30, 0.80\}$.[4]

The predicted response proportions obtained with these parameter values were multiplied by the individuals' sample size in Benjamin et al.'s (2009) study and were fitted by the averaging models that restrict response criteria across ensembles, with and without criterion noise. The models' performance is assessed with corrected Akaike information criterion (AIC_c; Burnham & Anderson, 2002) and the Bayesian information criterion (BIC; Schwarz, 1978). For the data-generating model, parameter values are perfectly recovered (as no sampling variability is introduced) with $AIC_c = 37.78$ and $BIC = 88.28$, values that simply correspond to the penalty factors attributed to the model for its number of parameters and the sample size. Regarding the models with restricted response criteria, the model with criterion noise performs the best, with $AIC_c = 32.94$ and $BIC = 57.64$, but returns highly distorted (and numerically unstable) parameter estimates ($\mu_t = 28.10$, $\sigma_t = 0.01$, and $\sigma_c = 19.32$), exemplifying that criterion noise is inflated when response criteria are fixed across ensemble sizes. The model without criterion noise performs the worst, as $AIC_c = 43.00$ (but $BIC = 64.70$), although it provides a better parameter recovery ($\mu_t = 1.08$, $\sigma_t = 1.00$). In addition, this example also shows that the wrong model can be preferred by the model selection measures used. Note that other aspects ignored in this example—such as different generating parameter values, sampling variability, and model misspecification—are likely to lead to additional complications, which means that these results do not imply that AIC_c and BIC punishment factors are in general disproportionate for these models.

Overall, the restriction of (mean) response criteria across ensembles seems to be an unpromising option, as it leads to implausible model predictions that potentially distort parameter estimates. Note that this limitation could not be assessed in the results originally reported, as no parameter estimates for the different candidate models were provided by Benjamin et al. (2009).

A Reanalysis of Benjamin et al.'s (2009) Data Sets

In the study reported by Benjamin et al. (2009), 19 participants were tested with the ensemble recognition task, with ensembles of size 1, 2, and 4. There were 60 trials in each ensemble condition, equally divided between target and distractor trials. We fitted the several cases of the averaging, summation, and OR models, for a total of nine models, using the maximum likelihood method, to the aggregated and individual data sets. Note that analyses of the aggregated data were not originally reported by Benjamin et al., but we include them because despite the risks of data distortions (e.g., Estes & Maddox, 2005), they are useful when assessing parameter estimates for the different models—estimates that might be distorted for individual data sets given the small number of trials (see Cohen, Sanborn, & Shiffrin, 2008). The model fitting procedures were implemented in R (R Development Core Team, 2011), and the R scripts are made available in the supplemental materials. Model performance was assessed by means of AIC_c and BIC. Additionally, we include a summation model with the response-criteria shift restriction, using the above-described parameterization.

Regarding model performance results, summed over the 19 individual data sets (see Table 1), they are similar to the averaged ones reported by Benjamin et al. (2009). Model selection measures (AIC_c

[4] For the averaging model, if one assumes that $\sigma_t = 1$ and that response criteria are positioned according to predefined log-likelihood ratios (LL_i), it can be shown by means of simple algebraic manipulations that the positioning of response criteria across ensembles is given by $C_{h,i} = \frac{\mu_t}{2} + \frac{LL_i}{\mu_t n_h}$.

Table 1
Goodness-of-Fit and Model Selection Results for Benjamin et al.'s (2009) Data Sets

		Individual data sets				Aggregated data sets			
Restriction type	Model	G^2	*df*	AIC_c	*BIC*	G^2	*df*	AIC_c	*BIC*
Without mean response criteria restrictions	Averaging-σ_c	265.34*	228	1,030.06	2,041.33	20.25	12	56.45	166.72
	Summation-σ_c	266.43*	228	1,031.15	2,042.42	20.63	12	56.83	167.10
	Integration	273.38	247	991.16	1,950.70	20.63	13	54.81	158.97
	OR	286.60*	247	1,004.38	1,963.93	28.54*	13	62.72	166.88
With mean response criteria restrictions	Averaging-σ_c	604.88*	418	924.88	1,394.21	93.56*	22	109.60	158.66
	Averaging	879.27*	437	1,157.64	1,569.93	386.51*	23	400.54	443.47
	Summation-σ_c	603.88*	418	923.88	1,393.21	76.23*	22	92.27	141.33
	Summation	834.55*	437	1,112.92	1,525.22	268.52*	23	282.55	325.48
	OR	1,460.36*	437	1,738.735	2,151.03	957.29*	23	971.32	1,014.25

Note. The values under the label "individual data sets" are the sums of the individuals' values. These results for the individual data sets thereby represent a single model with different parameters for each individual. The upper half of the table shows the results when mean response criteria are allowed to vary between ensemble sizes; the lower half shows the results when mean response criteria are fixed to be equal across ensemble sizes. The affix "-σ_c" indicates the models that assume criterion noise. When mean response criteria are not restricted (upper half), the averaging and summation model without criterion noise are identical, referred to as the integration model. AIC_c = corrected Akaike information criterion; *BIC* = Bayesian information criterion; OR model = the maximum rule model for which criterion noise is per definition absent.
* $p < .05$.

and *BIC*) indicate a preference for models that allow for criterion noise but do permit response-criteria shifts. Despite being a small difference, the summation model performs better than the averaging model, which corroborates the plausibility of the chosen parameterization. However, note that all models with the exception of the integration model are rejected by the data when using the G^2 statistic ($p < .05$).

Response-criteria shift restriction. The evaluation of the response-criteria shift restriction indicates mixed results, depending on the method used and the assumptions regarding criterion noise: AIC_c and *BIC* tend to indicate a preference for the response-criteria shift restriction, although this preference is not consistent across models and levels of data analysis: For the averaging model, when criterion noise is assumed, the response-criteria shift restriction is preferred for the individual data sets (AIC_c: 15 individuals, *BIC*: 19 individuals) and summed individual results ($\Delta AIC_c = -105.17$, $\Delta BIC = -647.11$), but contradictory preferences are found in the aggregated data ($\Delta AIC_c = 53.15$, $\Delta BIC = -8.06$). A different pattern of results is found when criterion noise is assumed to be absent, as the response-criteria shift restriction is preferred less frequently for individual data sets (AIC_c: 7 individuals, *BIC*: 16 individuals), contradictory preferences are obtained in the summed individual results ($\Delta AIC_c = 166.48$, $\Delta BIC = -380.77$), and a rejection is found in the aggregate data $\Delta AIC_c = 345.72$, $\Delta BIC = 284.50$). A similar pattern of results is obtained with the summation model (see Table 1).

In contrast, evaluating the differences in goodness-of-fit (ΔG^2) via null-hypothesis testing results in an overall rejection of the response-criteria shift restriction, independently of the model: For the averaging model with and without criterion noise, the restriction was rejected for nine and 14 individuals, respectively, whereas for the summation model with and without criterion noise, significant results were obtained for six and 15 individuals, respectively. For both the summed individual results and the aggregated data set, the response-criteria shift restriction is rejected in all models (all $p < .01$).

The evaluation of the response-criteria shift restriction across ensembles strongly depends on the method used and the adopted criterion noise assumptions: With criterion noise in the model, AIC_c and *BIC* prefer the restricted models; without criterion noise, the pattern is mixed; null-hypothesis tests reject the restriction consistently. These results, in addition to the problems previously pointed out suggest that the response-criteria shift restriction should not be taken into consideration when attempting to measure criterion noise. As shown below, this restriction plays a fundamental role in the conclusions reached by Benjamin et al. (2009).

Evaluation of criterion noise. When testing for the presence of criterion noise, an interesting pattern of results emerges: The results favor the inclusion or exclusion of criterion noise, depending on whether the response-criteria shift restriction is imposed. This pattern is found independently of the method, model, or level of data analysis that is adopted. For the case of the averaging model, when response criteria are not allowed to shift across ensembles, the inclusion of criterion noise finds support in the individual data sets (AIC_c: 16 individuals, *BIC*: 15 individuals), summed individual results ($\Delta AIC_c = 232.76$ and $\Delta BIC = 175.72$), and the aggregate data ($\Delta AIC_c = 290.94$ and $\Delta BIC = 284.81$). In contrast, when no response criteria restriction is imposed, criterion noise does not find any support in the individual data sets (AIC_c: 1 individual, *BIC*: 0 individuals), summed individual results ($\Delta AIC_c = -38.90$ and $\Delta BIC = -90.63$), or the aggregate data ($\Delta AIC_c = -1.64$ and $\Delta BIC = -7.76$).

The same conclusions are reached when using null-hypothesis testing:[5] When imposing the response-criteria shift restriction, the

[5] In such testing, the null hypothesis ($\sigma_c = 0$) lies on the boundary of the alternative hypothesis ($\sigma_c > 0$). In these circumstances, the sampling distribution of the likelihood-ratio test statistic ΔG^2 no longer follows a χ^2 distribution with the appropriate number of degrees of freedom, but a $\bar{\chi}^2$ distribution, which consists of a weighted mixture of χ^2 distributions with different number of degrees of freedom (Self & Liang, 1987; Shapiro, 1985). For a single individual/aggregated data set, $\bar{\chi}^2 \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, whereas for the summed results of N data sets, $\bar{\chi}^2 \sim \sum_{i=0}^{N} \frac{1}{2}^{N} \binom{N}{i} \chi_i^2$. Note that χ_0^2 is a distribution that concentrates all probability mass on 0.

hypothesis ($\sigma_c = 0$) is rejected for 16 individuals, as well as for the summed results ($\Delta G^2 = 274.16$, $p < .01$) and the aggregate data ($\Delta G^2 = 276.10$, $p < .01$). When no such restriction is imposed, the hypothesis ($\sigma_c = 0$) is only rejected for one individual and is not rejected for the summed results ($\Delta G^2 = 8.04$, $p = .56$) and aggregate data ($\Delta G^2 = 0.38$, $p = .27$). A similar pattern of results, both with AIC_c/BIC and null-hypothesis testing, is obtained with the summation model (see Table 1).

From these analyses, it becomes clear that the contribution of criterion noise reported by Benjamin et al. (2009) is critically dependent on the presence of the response-criteria shift restriction, a restriction that is not essential for the effective estimation of criterion noise, despite being motivated by previous findings (e.g., Stretch & Wixted, 1998). To make matters worse, this restriction leads to a series of complications in both the averaging and summation models, making its implementation undesirable, as previously discussed.

Parameter estimates. So far, these models were discussed in terms of goodness-of-fit results and model parsimony (according to AIC_c and BIC), disregarding the estimated parameter values. Note that the purpose of this modeling enterprise is to be able to obtain separate measurements for the different processes operating. Given the theoretical meaning that can be attributed to the parameters (e.g., μ_t and σ_t; Wixted, 2007), it is important to assess their values across the different models and restrictions.

The results presented in Table 2 indicate that the response-criteria shift restriction has a critical role in determining the best-fitting parameter values. The parameter estimates for the OR model, as well as for the averaging and summation models without criterion noise and with response criteria restrictions, are omitted for the sake of brevity. As can be seen in Table 2, the response-criteria shift restriction frequently leads to extreme and numerically unstable values of μ_t, σ_t, and σ_c. Note that both σ_t and σ_c estimates are to be directly compared with the standard deviation of the distractor distribution (σ_d), which is set to 1 for scaling purposes. These parameter estimates are implausible in terms of the theoretical meaning given to SDT parameters in the recognition memory literature (e.g., Wixted, 2007), where the larger variability is sometimes attributed to variability in the encoding process itself. If the reported parameter estimates are taken at face value, then it would mean that a variable encoding process during study induces a variability in the evidence distribution that is several orders of magnitude larger than the variability present prior to study. For example, the σ_t estimate for participant 7 with the restricted averaging model is 56.87 times larger than σ_d. This possibility also fails to find any support in simulations done with more refined computational models of recognition memory, where reasonable values of encoding variability lead to much smaller differences in variability between the two evidence distributions (e.g., Shiffrin & Steyvers, 1997, p. 149). Additionally, the range of the parameter estimates is so broad that it is difficult to find any common scaling among participants. For the restricted averaging model, μ_t ranges from 0 to 115.24, and σ_t ranges from 0.01 to 131.78.

Table 2
Parameter Estimates for Benjamin et al.'s (2009) Individual and Aggregated Data Sets for Three Selected Models

	Averaging model						Summation model						Integration model	
	Not restricted			Restricted			Not restricted			Restricted				
Participant	μ_t	σ_t	σ_c	μ_t	σ_t	σ_c	μ_t	σ_t	σ_c	μ_t	σ_t	σ_c	μ_t	σ_t
1	0.00	1.47	0.00	0.00	2.04	1.09	0.00	6.96	9.61	0.00	54.81	100.55	0.00	1.47
2	0.36	1.46	0.00	9.06	12.81	20.71	9.29	32.62	41.10	17.86	56.24	74.18	0.36	1.46
3	0.20	1.21	0.00	0.29	1.29	0.68	2.39	9.24	18.65	7.09	21.78	74.68	0.20	1.21
4	0.59	1.42	0.13	0.74	1.65	0.53	0.62	1.46	0.54	25.44	44.04	77.16	0.58	1.40
5	1.20	1.28	0.00	40.81	24.12	24.37	1.73	1.66	1.40	1.47	1.13	1.72	1.20	1.28
6	2.24	2.28	0.84	56.26	41.05	30.77	1.43	1.80	0.00	1.25	1.48	1.48	1.42	1.80
7	37.99	30.27	18.97	96.00	56.87	53.39	1.35	1.58	0.00	1.10	1.26	0.44	1.35	1.58
8	0.95	1.10	0.00	4.11	0.07	3.12	1.31	1.29	1.36	36.37	24.13	62.98	0.95	1.10
9	0.76	1.38	0.00	30.44	32.08	28.89	0.90	1.51	0.95	12.52	10.63	30.92	0.76	1.38
10	1.53	0.01	2.85	21.12	0.01	42.22	0.33	0.98	0.00	0.34	0.99	0.68	0.33	0.98
11	0.94	1.74	1.31	2.91	4.39	4.48	0.41	1.21	0.00	12.93	13.56	65.34	0.41	1.21
12	0.58	1.15	0.04	3.14	0.01	3.94	0.73	1.33	1.13	10.59	14.85	27.86	0.58	1.15
13	1.53	1.18	0.17	4.52	1.84	2.00	1.48	1.18	0.00	43.57	13.02	57.25	1.48	1.18
14	1.18	1.17	0.77	23.01	7.00	21.77	0.73	1.07	0.00	1.22	1.39	2.33	0.73	1.08
15	1.43	2.43	0.90	56.99	71.57	50.44	0.82	1.70	0.00	1.15	2.30	2.29	0.82	1.70
16	1.38	1.92	0.00	115.24	131.78	59.31	5.70	7.04	5.33	14.44	14.69	35.30	1.38	1.92
17	2.22	7.00	0.00	1.83	4.00	0.37	20.20	62.68	14.06	1.86	3.50	1.45	2.22	7.00
18	1.14	1.33	0.42	31.25	15.47	24.21	0.94	1.24	0.00	2.30	2.23	3.84	0.94	1.24
19	0.23	0.85	0.06	0.35	0.80	0.58	0.25	0.73	0.94	0.46	0.01	3.42	0.23	0.85
Aggregated	0.66	1.35	0.24	52.69	54.18	59.42	0.62	1.32	0.00	2.56	4.08	7.22	0.62	1.32

Note. Given their extreme values, some estimates are numerically unstable. Therefore, we report the estimates that provided the best goodness-of-fit results out of 50 fitting runs. For the model referred to as *not restricted*, the mean response criteria are free to shift between ensemble sizes. Mean response criteria are fixed to be equal for the *restricted* models. When criterion noise was fixed to 0 and the response criteria were free to shift across ensemble sizes, averaging and summation models collapse into the integration model. Parameter estimates for the maximum rule model (the OR model), and for the averaging and summation models with both response criteria and criterion noise restricted, are omitted due to space restrictions.

The estimates for the aggregated data sets provide mixed results, as the estimates for the averaging model are still very large, although the ones for the summation model approach more commonly observed values. It may be that these extreme parameter estimates are in part caused by the small number of observations per individual (see Macmillan, Rotello, & Miller, 2004), which can lead to inaccurate estimates. When inspecting the data matrix resulting from Benjamin et al.'s (2009) experiment, it can be seen that out of 684 cells (19 participants × 36 response categories), 425 (62%) have frequencies smaller than 6, which is known to compromise parameter estimation and testing on data following a multinomial distribution (e.g., Bishop, Fienberg, & Holland, 1975, Chapter 12). Still, the averaging model estimates for the aggregated data suggest that the small number of trials is probably not the sole cause of these issues.

Establishing upper bounds on parameter values—such as μ_t, $\sigma_t \leq 4$, for example—could be seen as a way to deal with these extreme parameter estimates, although it does not represent a true solution to this problem as it would simply mask it. Upper bounds can be useful when dealing with occasional outliers, but in the present case, most of the parameter estimates are extreme: These upper bounds would be reached for 68% and 53% of the individuals with the restricted averaging and summation models, respectively. This would impose a ceiling that compromises the characterization of at least half of the participants in this data set and is likely to render the estimates useless as measures of the processes underlying participants' performance.

When the response-criteria shift restriction is not imposed, the number of cases with extreme parameter estimates drops dramatically, which suggests that in these cases, the extreme estimates were caused by the response-criteria shift restriction. Also, the estimates for σ_c become rather low for several individuals, which explains the non-rejection of the null hypothesis $\sigma_c = 0$. Regarding the integration model, which assumes no criterion noise, there are virtually no cases of extreme parameter estimates. The exception is participant 17, whose data vector has 21 empty cells out of 36, which compromises any reliable model fitting. If the extreme parameter estimates somehow represented an adequate description of the latent evidence distributions, then one would not expect the response criteria restriction to lead to such large differences in parameters μ_t and σ_t. The fact that large differences are observed strongly suggests that the extreme parameter estimates are the outcome of imposing response criteria restrictions that are not consistent with the processes generating the data.

Overall, the results confirm the suspicions that were raised regarding the impact of response criteria restrictions: Not only do they frequently lead the models to return unrealistic parameters estimates, they also have a decisive and undesired role in the assessment of criterion noise. A possible reason for the poor performance of the restricted models is that the remaining free parameters attempt to counteract the implausible predictions that the response-criteria shift restriction imposes, thereby compromising the validity of parameter estimates. Note that if mean response criteria do in fact shift as a function of ensemble size, a restricted model can compensate for this to some extent by inflating the parameter for criterion noise, as previously shown.

Discussion. Benjamin et al. (2009) claimed that criterion noise assumes high magnitudes, a result that could potentially lead to a reinterpretation of several findings so far reported in the literature (e.g., response conservatism; Thomas & Legge, 1970). Our reanalysis points out serious problems that compromise the original conclusions: Criterion noise only makes a significant contribution when response criteria are not allowed to shift across ensembles, a restriction that introduces problematic predictions in both the averaging and the summation model and that ultimately leads to extreme parameter estimates that are difficult to interpret in terms of the model's theoretical principles. Once this problematic restriction is lifted and response criteria are allowed to shift across ensemble sizes, criterion noise simply fails to make a significant contribution to the account already given by traditional SDT.

One of the most important points that this reanalysis makes is that descriptive parsimony in terms of AIC_c, BIC, or any other model selection measure cannot be a researcher's sole yardstick. Descriptions not only need to be as simple as possible, they also need to be interpretable, and the parameter estimates provided by the restricted models are somewhat challenging when it comes to attribute to them some sort of psychological meaning. One of the main problems with the parameter estimates is that they are based on very few trials, which combined with strong assumptions such as the integration rules or stability of criterion noise across ensembles[6] can lead to severely distorted results. Some of these problems can potentially be dealt with by means of a Bayesian hierarchical modeling approach (e.g., Pratte, Rouder, & Morey, 2010), although its implementation would require more participants than are available here.

Additional issues regarding Benjamin et al.'s (2009) model and conclusions were raised by DeCarlo (2010), who showed that the effects of criterion noise on SDT parameter estimates are heavily dependent on the manner in which the model is parameterized and how criterion noise is ultimately introduced, as a different parameterization from the one adopted by Benjamin et al. would only lead to reduced estimates of discriminability (DeCarlo, 2010, p. 309). In any case, it is possible a priori that Benjamin et al.'s specific parameterization of criterion noise would have provided a better account of the data than the SDT model without criterion noise. As it turns out, it does not, unless problematic restrictions are introduced.

It is important to note that although our results show that the claims made by Benjamin et al. (2009) regarding the contribution of criterion noise in recognition memory performance are ques-

[6] Benjamin et al. (2009) remarked that if one of the models assumes that criterion noise is affected by ensemble size *in the exact same manner as* the variability of targets and distractors (e.g., specifying criterion noise as $\frac{\sigma_c}{n_h}$ instead of σ_c for the averaging model), then this model would reduce to a model in which no criterion noise is assumed at all (see Benjamin et al., 2009, Appendix B). Given this equivalence, Benjamin et al. argued that the comparatively worse performance of models without criterion noise represents evidence in support of the assumption that criterion noise is constant across ensembles (Benjamin et al., 2009, p. 96). It is important to note that the proof in their Appendix B only concerns the case where criterion noise is forced to change across ensembles in a very specific manner and bears no weight regarding the general question of whether criterion noise varies across ensembles. If it did, it would result in the paradoxical case that relaxing a parameter restriction (e.g., allowing σ_c to vary across ensemble sizes) leads to a restriction in the model's predictions.

tioned by a careful assessment of model predictions, model performance, and parameter estimates, they do not necessarily imply that criterion noise is non-existent. Also, the models considered by Benjamin et al. only instantiated one of many possible models of criteria variability (see Klauer & Kellen, 2012), making any claims based on them dependent on the adequacy of the implemented criterion noise model.

A Model Generalization Approach for the Measurement of Criterion Noise

Although it can be argued that the reanalysis of Benjamin et al.'s (2009) results raise doubts regarding the presence of criterion noise, the evidence provided is paradigm-specific and therefore should not be generalized. Given this issue, it would be important to have an alternative method to estimate criterion noise, one that does not hinge on the same principles that underlie the ensemble recognition approach. If possible, this method should also attempt to go beyond Benjamin et al.'s approach and, in particular, allow for the assessment of additional implementations of criterion noise that have so far been proposed in the recognition memory literature—namely the DNM's response rule proposed by Mueller and Weidemann (2008), which assumes that response criteria are evaluated in a sequential and conditional manner. The possibility of considering different implementations of criterion noise is extremely important as it not only allows one to assess which of these is more adequate but also the robustness of conclusions across distinct criterion noise models.

An alternative method for the measurement of criterion noise is available by going beyond the traditional choice of a yes–no/rating task and capitalizing on the characteristics of SDT for less popular tasks. One of such tasks is the four-alternative forced choice with two responses task (4AFC-2R; Swets et al., 1961). In the 4AFC-2R task, individuals are to recognize the old item among four alternatives, knowing beforehand that three of them are distractors. In some of the trials, after giving their response, individuals are allowed to choose another item among the remaining three alternatives. This task was originally used by Swets et al. (1961) to test certain predictions of SDT and as an attempt to understand the relationship between the model's parameters (see also Solomon, 2007). More recently, Kellen and Klauer (2011) provided a full characterization of the SDT model for this task and exemplified its use in the field of recognition memory using data obtained by Parks and Yonelinas (2009).

The SDT model can be fully specified for the 4AFC-2R or, more generally, a *k*AFC-*n*R task (with $k \geq 3$ and $2 \leq n \leq k$) as easily as for the binary or rating judgments required to obtain ROCs: Let $F_{(\mu,\sigma)}$ and $f_{(\mu,\sigma)}$ be the distribution function and probability density, respectively, of the normal distribution with mean μ and standard deviation σ, with F and f being these functions for the standard normal distribution (i.e., $F = F_{(0,1)}$ and $f = f_{(0,1)}$). Additionally, let π_j be the unconditional probability that out of k alternatives − 1 correct and $(k - 1)$ incorrect − the jth choice ($j = 1, \ldots, n$) is the correct one. In a *k*AFC-*n*R task, the SDT model is specified by a simple rule: For each trial, the alternatives are ordered according to their associated evidence values and then are chosen in a decreasing order. Therefore π_1 is defined as the probability that the studied item has a larger evidence value than all $(k - 1)$ incorrect alternatives, whereas π_2 corresponds to the probability that the studied item will have a larger evidence value than just $(k - 2)$ incorrect alternatives, and so forth. According to SDT, π_j is given by:

$$\pi_j = \binom{k-1}{j-1} \int F^{k-j}(x) f_{(\mu_t,\sigma_t)}(x)(1 - F(x))^{j-1}\, dx \qquad (6)$$

Note that no response criteria are involved in Equation 6, meaning that parameters of interest such as μ_t and σ_t can be estimated in the absence of response criteria and thus of criterion noise by means of a task such as the 4AFC-2R. Kellen and Klauer (2011) estimated these parameters using such a task and obtained parameter estimates that closely resemble the ones obtained in the ROC literature, a result that suggests a relatively good match between the different methods. Although previous work in perception (e.g., Klein, 2001) has questioned the assumption of unbiased responding in *k*AFC tasks, note that in the perception literature, the alternatives presented on each trial normally correspond to observation intervals that occur sequentially, which can compromise the assumption that the observation and evaluation of the alternatives is independent. For the case of recognition, this issue does not apply given that all studied items are presented in a separate phase and all test alternatives are shown simultaneously.

The possibility of fitting the SDT model (and other related models as well; see Kellen & Klauer, 2011) within this type of multiple-alternative, multiple-response tasks allows one to adopt them as alternative methods to estimate model parameters. Still, this possibility does not exhaust its advantages, as this ability can be exploited for the study of criterion noise through the generalization of parameters across tasks, namely parameters μ_t and σ_t. If a SDT model is simultaneously fitted to data from a traditional yes–no/rating task, in which both stimulus and criteria variability are present, and data from the 4AFC-2R task, in which only the stimulus variability is present, then it becomes possible to obtain separate estimates for stimulus and criteria variability. This strategy follows previous approaches in cognitive modeling that have relied on the generalization of models across distinct tasks (e.g., Busemeyer & Wang, 2000; Chechile & Soraci, 1999; Jang, Wixted, & Huber, 2009). One of the benefits of this kind of approach is that it prevents one from falling prey to what has been called mono-operation bias (Shadish, Cook, & Campbell, 2002, Chapter 3), a bias that is incurred when conclusions rely excessively on a single method. In the case of recognition memory, the overwhelming majority of the work has relied on confidence-rating ROCs, which represent only one of many tasks that the models can be specified for, most of them completely overlooked (e.g., Murdock, 1963). Focusing one's efforts for model assessment and selection on a single task is likely to produce biased results, given that a "winning" model might be simply overfitting the data for that task, being in fact unable to provide accurate predictions and generalizations for different operationalizations.

Furthermore, it is important to take into account that the parameter generalization imposed by this strategy is by no means arbitrary. In fact, it is determined by the basic principles underlying SDT. As shown by Iverson and Bamber (1997), the famous area theorem (Green & Moses, 1966) can be extended to show that an individual's performance in *k*-alternative forced-choice tasks provides strong predictions regarding that individual's ROC function. According to the *generalized area theorem*, by specifying a random variable whose distribution function determines the ROC, it can be shown that the proportion of correct responses (π_1) in a kAFC task is an estimate of the ($k - 1$)th moment of that variable.[7] Given that the collection of all moments of a random variable on a finite interval is sufficient to characterize that same variable (Feller, 1966), the π_1 values for an infinite sequence of kAFC tasks provide a full characterization of the ROC. Iverson and Bamber also demonstrate that if a task allows for additional responses within a single set of alternatives, then lower moments can be estimated as well. For example, a 4AFC-4R task provides the same information as the set of π_1 values for 4AFC, 3AFC, and 2AFC tasks, meaning that the first three moments of the distribution function determining the ROC can be conveniently estimated through a single experimental condition.

Furthermore, a kAFC-kR task can be implemented as a k-alternative ranking task (J. Brown, 1965; Dalrymple-Alford, 1970; Iverson & Bamber, 1997). In the k-alternative ranking task participants are simply required to order the items according to the likelihood with which they are believed to have been previously studied. Implementing a kAFC-kR task by means of a k-alternative ranking task has the advantage that each trial provides information regarding the first ($k - 1$) moments of the function determining the ROC, instead of having to collect different types of trials with different number of choices. This means that more informative data and hence better parameter estimates can be obtained via the ranking task.

Two aspects of the generalized area theorem are especially important: First, the theorem is based on distribution-free assumptions, meaning that the predictions that emerge from it are independent of the specific distributions (e.g., Gaussian, Gamma, log-normal) that are adopted. Second, the ROC function considered by the theorem excludes the involvement of response criteria and therefore of any form of criterion noise, so that the comparison of rating and kAFC responses can in principle be capitalized upon for the estimation of criteria variability. The presence of criteria variability will lead to distortions in the parameter estimates (μ_t and σ_t) obtained with a confidence rating task, which will result in discrepancies between these parameter estimates and the corresponding ones obtained solely by means of kAFC responses. These discrepancies will consequently lead to worse goodness-of-fit results when μ_t and σ_t are assumed to be equal across tasks, an equality that is expected on the basis of the generalized area theorem. The introduction of criterion noise in the SDT model is expected to alleviate this misfit, as it would allow the model to account for the differences across tasks.

This model generalization approach (the assumption that μ_t and σ_t are equal across tasks) allows the estimation of criterion noise under the broader assumption that the differences across tasks correspond to the ones specified by SDT (Iverson & Bamber, 1997). This assumption can be seen as somewhat questionable given the body of evidence pointing out the existence of task-specific modes of processing (e.g., Benjamin, 2008; Malmberg, 2008), differences that might represent a violation of the predictions of SDT. Similarly to previous work (e.g., Benjamin et al., 2009; Morrell et al., 2002), the present approach is couched within the SDT framework, meaning that the SDT model is assumed to provide a suitable approximation to the cognitive processes operating during a recognition test and, by consequence, that its generalizations across tasks are assumed to hold.

The model generalization approach will be implemented through the use of a single study phase along with the mixing of tasks within the recognition test phase, namely 4-alternative ranking and 6-point confidence rating trials. The individual data resulting from both tasks can then be fitted simultaneously. Three models that implement criterion noise are considered: the DNM proposed by Mueller and Weidemann (2008); a restricted version of DNM designated as DNM$_r$, in which classification and criterion noise are fixed to be equal; and the restricted case of the law of categorical judgment (LCJ$_r$; see Klauer & Kellen, 2012; Rosner & Kochanski, 2009) that was considered by Benjamin et al. (2009). Note that the formal account of rating responses given by the LCJ$_r$ model corresponds to both the averaging and summation models (Equations 3 and 4) when n_h is fixed to 1.

The comparison of these alternative models provides a more global account of criterion noise as well as an assessment of which of these implementations better describes the data.

Method

Participants. Thirty individuals (29 students and 1 nonstudent) participated in this experiment. They were recruited via flyers and displays at the University of Freiburg and received €7 in exchange for participation. Mean age of participants was 23.4 years ranging from 18 to 33 years ($SD = 3.2$). Each experimental session lasted approximately 45 min.

Design. Experimental sessions were divided into a single study phase followed by a single test phase. In the test phase, two types of trials were presented in random order: rating trials and ranking trials. In the rating trials, participants saw a single word that could (or not) have been presented in the study phase and had to indicate—on a 6-point scale ranging from *sure new* to *sure old*—whether they believed this word had been presented in the study phase. In the ranking task, participants saw four words on the screen (one on each corner of the screen), only one of which had been presented in the study phase. The participants—who were aware of that only one word among the alternatives was previously studied—had to order the four words by clicking on them, from the one they believed most to have been presented in the study phase to the one they believed least to have been presented in the study phase. The position of the previously studied word in the ranking trials was randomly chosen. To prevent primacy and recency effects, the first and last five words of the study list were not presented in the test phase.

Materials. The word list contained 639 neutral German nouns taken from Lahl, Göritz, Pietrowsky, and Rosenberg (2009) ranging from 4 to 8 letters in length. According to the ratings obtained by Lahl et al., the words were all of medium valence

[7] In general terms, the ROC is a function $x \rightarrow g(x)$ from the closed interval [0, 1] into itself, increasing and continuous on (0, 1), with $g(0) = 0$ and $g(1) = 1$. Consider then two independent random variables V_t and V_d, both defined on [0, 1], with V_d uniformly distributed. Each pair of false alarm and hit proportions (p_{FA}, p_H) forming a point on a ROC is given by $p_{FA}(x) = P(V_d > x) = 1 - x$ and $p_H(x) = P(V_t > x) = g(1 - x)$. The generalized area theorem states that π_1 in a 2AFC task corresponds to an estimate of $E(V_t)$ and, more generally, that π_1 in a kAFC task corresponds to an estimate of $E(V_t^{k-1})$. For proofs and a detailed discussion, see Iverson and Bamber (1997).

(ranging from 3.50 to 6.50 on an 11-point scale) and low in arousal (ranging from 0.50 to 4.50 on an 11-point scale). Furthermore, all words were of approximately equal word frequency in common German, as indicated by log frequency ratings obtained for each word via WordGen (ranging from 0.30 to 2.90; Duyck, Desmet, Verberke, & Brysbaert, 2004).

In each experimental session, we used 610 randomly drawn words from the word list (the stimulus set). The study list consisted of 200 randomly drawn words from the stimulus set ("old" words). In the study phase, we presented five filler words at the beginning and five filler words at the end to prevent primacy and recency effects. These filler words did not appear in the test phase. One hundred of the words on the study list were randomly chosen for rating trials, the other 100 words were used in ranking trials. For the rating trials, another 100 words from the stimulus set were randomly chosen ("new" rating words). The remaining 300 words in the stimulus set were used as "new" words in the ranking task. In total, participants saw 200 rating trials (100 with "old" words, 100 with "new" words) and 100 ranking trials (each consisting of one "old" word and three "new" words) in the test phase.

Procedure. Participants were tested individually using PsychoPy (Peirce, 2007) for stimulus control. Prior to the study phase, participants gave their informed consent and were introduced to the task. They learned that they were to conduct a memory experiment consisting of a study and a test phase. We specifically explained to them that they would see a series of words in the study phase that they should try to memorize as well as possible.

Then the study phase started. Each word was presented for 1.5 s with a 0.5-s interstimulus interval. Directly after the study phase, participants read that the test phase would consist of two types of trials: rating trials and ranking trials. We thoroughly explained the nature of the ranking trials. Participants read that only one of the four presented words was old and the other three were new and that they should rank the words such that the order of the words reflects their feeling for the likelihood of the individual words being old. Also, they were informed that the position of the old word among the four was randomly chosen. The ranking was achieved by clicking the words with the mouse. The first word that was clicked received the highest rank (a "1" appeared next to the word), the second one that was clicked received the second highest rank (a "2" appeared next to the word), and so on. Furthermore, participants could deselect a word or several words by clicking on them again, in which case all other ranks were updated accordingly. Participants could only proceed to the next trial when all four words were ranked and they confirmed the shown ranking. To familiarize participants with the ranking task, they had to work on one test trial that could be repeated if participants wanted to. Next, participants read that in the rating trials they had to decide whether the item was old or new and that they should use a 6-point rating scale to indicate how confident they were in their decision. Then, the test phase started, with ranking and rating trials randomly alternated. After finishing the test phase, participants were thanked and debriefed.

Results and Discussion

Models were fitted using the maximum likelihood method. The model fitting routines were implemented in R (R Development Core Team, 2011). When fitting models in general, the question of whether to aggregate data across participants is posed: Although aggregating data is known to lead to severe distortions and erroneous inferences (e.g., Estes & Maddox, 2005), suggesting that fitting individual data is more suitable, aggregation can still be advantageous in situations with small number of trials per condition (Cohen et al., 2008). For this reason, the models were fitted to both aggregated and individual data.

Table 3 shows goodness-of-fit values and parameter estimates for the SDT model without criterion noise when fitting data from the 4-ranking and 6-point rating tasks separately and simultaneously. The goodness-of-fit tests indicated statistically significant deviation from the model for only three individual data sets for the ranking trials (smallest $G^2(1) = 4.33$, $p < .05$) and for only two individuals in the rating trials (smallest $G^2(3) = 8.99$, $p < .05$). Fitting both tasks jointly with equal parameters led to significant deviations for only four individuals (smallest $G^2(6) = 12.63$, $p < .05$). These results suggest that although not perfect, the account of individual data provided by the SDT model is adequate. Concerning the summed and aggregated data, the model is rejected on both tasks whether analyzed separately or jointly, which is to be expected given that the model represents only an approximation to the true data-generating process with probability of rejection approaching 1 as the data size increases. A SDT model that excludes criterion noise and allows mnesic evidence parameters (μ_t and σ_t) to differ across tasks—a model designated as SDT$_{sep}$—is later compared with the other candidate models.

Goodness-of-fit results as well as the respective parameter estimates for the models with criterion noise (with the exception of DNM$_r$) are reported in Table 4. For individual data sets, the models are rejected few times ($p < .05$; seven times for the DNM and four times for the LCJ$_r$). For the aggregate data, all models are rejected ($p < .01$), which is to be expected given the sample size. These results indicate that despite some discrepancies, the models provide in general a relatively good description of the data. Figure 3 shows the ROC and *z*ROC (ROC plot on a probit scale) plots for the aggregate data: Ratcliff et al. (1994) as well as Malmberg and Xu (2006) hypothesized that certain forms of criterion noise lead to inverse U-shaped *z*ROCs. The *z*ROC obtained is very slighty U-shaped, a result that indicates that the form of criterion noise proposed by Ratcliff et al. and by Malmberg and Xu is not present.

The parameter estimates concerning mnesic evidence (μ_t and σ_t) reported in Table 4 are consistent with the values normally found for the SDT model (without criterion noise) in this field (e.g., Yonelinas & Parks, 2007). The criterion noise parameter estimates (σ_c, σ_{class}, and σ_{conf}) assume rather low values, in many cases 0. This indicates that for these data sets, criterion noise as defined by these models is estimated to be very low or even non-existent, a result that runs counter to previous reports in the literature.

In terms of the model selection measures presented in Table 5, the results are clear: Both AIC$_c$ and BIC values strongly prefer the SDT model without criterion noise as the most adequate one, with the exception of AIC$_c$ for the aggregated data, which points the SDT$_{sep}$ as the most adequate model. These results reflect the low estimates of criterion noise presented in Table 4.

Table 3
Goodness-of-Fit Results and Parameter Estimates for the Reported Experiment for the Standard Signal Detection Theory Model Without Criterion Noise When the Four-Alternative Ranking and Six-Point Rating Tasks Are Fitted Separately and Jointly

	Four-alternative ranking task			Six-point rating task			Both tasks		
Participant	G^2	μ_t	σ_t	G^2	μ_t	σ_t	G^2	μ_t	σ_t
1	0.00	0.43	1.25	4.29	0.49	1.05	5.16	0.45	1.11
2	0.58	2.50	1.83	1.86	2.22	1.92	3.33	2.36	1.88
3	0.79	2.08	1.69	1.96	2.05	1.59	2.84	2.05	1.62
4	1.75	0.29	1.08	3.90	0.37	1.00	5.99	0.33	1.03
5	3.08	1.34	1.74	0.22	1.31	1.27	5.59	1.29	1.44
6	0.00	0.99	1.30	0.05	1.50	1.62	2.43	1.23	1.48
7	0.46	2.01	1.74	2.17	2.35	1.54	5.00	2.20	1.67
8	0.53	0.90	1.26	3.47	1.14	1.43	4.71	1.02	1.37
9	1.26	0.59	1.30	2.58	0.57	1.14	4.11	0.57	1.22
10	1.30	1.37	1.48	3.11	1.25	1.24	4.94	1.29	1.31
11	0.00	0.55	0.87	4.60	0.94	1.22	8.36	0.72	1.07
12	1.92	0.71	1.87	2.34	1.37	1.57	11.00	1.04	1.74
13	0.14	1.87	0.97	1.01	2.65	1.86	4.27	2.28	1.45
14	1.94	0.76	1.60	0.78	0.79	1.14	5.45	0.76	1.31
15	0.01	0.57	1.35	10.00*	1.21	1.50	15.16*	0.87	1.44
16	0.05	2.40	1.68	2.31	1.98	1.55	3.28	2.14	1.58
17	0.68	1.77	1.86	5.33	2.26	1.59	10.66	2.03	1.74
18	4.33*	0.91	1.29	8.99*	1.18	1.52	14.23*	1.05	1.44
19	0.12	2.08	1.19	3.54	1.67	1.23	6.30	1.89	1.24
20	0.94	1.28	1.50	5.02	1.67	1.51	7.73	1.47	1.52
21	1.26	1.15	1.06	0.62	1.60	1.28	3.86	1.32	1.13
22	7.02*	1.55	1.81	5.14	1.30	1.42	12.63*	1.41	1.68
23	7.30*	0.86	1.53	3.51	0.87	0.95	15.81*	0.84	1.14
24	2.13	1.08	1.11	3.27	1.43	1.33	6.72	1.24	1.23
25	0.74	1.72	1.64	1.94	1.43	1.71	3.78	1.58	1.68
26	3.03	2.13	1.80	2.57	1.54	0.95	10.86	1.68	1.18
27	0.11	1.79	1.55	7.09	2.39	2.17	8.71	2.13	1.92
28	1.65	0.67	1.64	1.85	1.07	1.44	6.67	0.86	1.53
29	0.34	0.63	1.05	7.55	0.42	1.11	9.09	0.54	1.10
30	0.34	2.83	1.76	1.55	2.43	0.97	6.42	2.58	1.29
Aggregated	13.05*	1.26	1.52	9.45*	1.32	1.42	30.69*	1.28	1.46

Note. The column labeled "Both tasks" shows the results with μ_t and σ_t restricted to be equal across the two tasks.
* $p < .05$.

Regarding the DNM, its difference in terms of goodness-of-fit from the SDT model without criteria variability can also be tested by means of null-hypothesis testing.[8] For the individual data sets, it was significant in only one case ($\Delta G^2 = 5.71$, $p < .05$), and no significant differences were found for both the sum of individual results ($\Delta G^2 = 23.99$, $p = .67$) and the aggregate data ($\Delta G^2 = 2.84$, $p = .11$).

Imposing the restriction $\sigma_{class} = \sigma_{conf}$ defining DNM$_r$ did not significantly deteriorate goodness-of-fit relative to the DNM: neither for individual data sets (largest $\Delta G^2(1) = 2.53$, $p = .11$), summed results ($\Delta G^2(30) = 15.75$, $p = .98$), nor aggregate data ($\Delta G^2(1) = 2.84$, $p = .09$).

Although the differences between DNM and DNM$_r$ were nonsignificant, it is interesting that in the majority of the cases the classification noise estimates are larger than the confidence noise estimates, a pattern that is at odds with the parameter estimates reported by Mueller and Weidemann (2008), who suggested the opposite difference. In consequence, it is also at odds with the explanation of Mueller and Weidemann's results proposed by Benjamin et al. (2009, p. 101), who considered that response criteria farther away from the classification criterion might be more variable due to a greater difficulty in maintaining them fixed. Regarding the results of Mueller and Weidemann, two aspects need to be taken into account: First, they used a restricted DNM that precludes representational variability and fits the data solely by means of variable response criteria, unlike the present approach that fits the data without such restrictions. Additionally, this restricted DNM can produce smaller estimates of classification noise compared to confidence noise even when no criteria variability is

[8] Note that, again, the null hypothesis ($\sigma_{class} = \sigma_{conf} = 0$) lies on the boundary of the alternative hypothesis ($\sigma_{class}, \sigma_{conf} > 0$), but in this case the test concerns the restriction of two parameters of interest. For a single data set (either an individual or the aggregated data set), the $\bar{\chi}^2 \sim \omega_0\chi_0^2 + \omega_1\chi_1^2 + \omega_2\chi_2^2$, where $\omega_0 = \frac{\cos^{-1}(\rho(\theta_1, \theta_2))}{2\pi}$, $\omega_1 = \frac{1}{2}$, and $\omega_2 = \frac{1}{2} - \frac{\cos^{-1}(\rho(\theta_1,\theta_2))}{2\pi}$, and $\rho(\theta_1, \theta_2)$ represents the correlation between the two parameters of interest (Self & Liang, 1987), correlations that were estimated by means of parametric bootstrap. For the summed results of N individual data sets, $\bar{\chi}^2 \sim \sum_{i=0}^{2N} \beta_i\chi_i^2$, where the β_i weights are defined as a function of the mixtures for the individual cases. Given the difficulty in obtaining an exact solution for the β_i weights, stable approximations were obtained via Monte Carlo simulation.

Table 4
Goodness-of-Fit Results and Parameter Estimates for the Reported Experiment for the DNM and the LCJ$_r$ Models

	DNM					LCJ$_r$			
Participant	G^2	μ_t	σ_t	σ_{class}	σ_{conf}	G^2	μ_t	σ_t	σ_c
1	4.84	0.48	1.14	0.00	0.59	5.16	0.45	1.11	0.00
2	2.47	2.69	2.08	0.91	0.88	2.90	2.61	2.04	0.62
3	2.61	2.05	1.60	0.52	0.00	2.84	2.05	1.62	0.00
4	4.54	0.35	1.01	1.02	0.00	5.99	0.33	1.03	0.00
5	5.52	1.32	1.46	0.00	0.30	5.59	1.29	1.44	0.00
6	2.42	1.23	1.48	0.36	0.00	2.43	1.23	1.48	0.00
7	4.28	2.28	1.72	0.00	0.42	5.00	2.20	1.67	0.00
8	4.71	1.02	1.37	0.01	0.00	4.71	1.02	1.37	0.00
9	3.94	0.58	1.23	0.08	0.40	4.06	0.60	1.25	0.46
10	3.16	1.38	1.36	0.88	0.34	4.88	1.33	1.34	0.36
11	7.77	0.74	1.07	0.73	0.00	8.36	0.72	1.07	0.00
12	10.84*	1.06	1.75	0.46	0.00	11.00	1.04	1.74	0.00
13	3.48	2.29	1.45	0.44	0.00	4.27	2.28	1.45	0.00
14	5.00	0.80	1.34	0.68	0.39	5.34	0.81	1.38	0.54
15	15.14*	0.89	1.47	0.98	0.00	15.16*	0.87	1.44	0.00
16	2.55	2.40	1.72	0.69	0.85	2.39	2.43	1.74	0.72
17	10.46*	2.05	1.75	0.39	0.00	10.66	2.03	1.74	0.00
18	12.41*	1.09	1.46	0.77	0.00	14.23*	1.05	1.44	0.00
19	0.59	2.09	1.21	1.19	0.83	3.81	2.16	1.29	0.86
20	5.20	1.56	1.58	0.79	0.00	7.73	1.47	1.52	0.00
21	3.78	1.32	1.13	0.20	0.00	3.86	1.32	1.13	0.00
22	12.24*	1.43	1.70	0.33	0.22	12.26*	1.55	1.78	0.64
23	15.55*	0.85	1.14	0.53	0.00	15.81*	0.84	1.14	0.00
24	6.72	1.24	1.23	0.01	0.01	6.72	1.24	1.23	0.00
25	1.79	1.67	1.73	0.88	0.20	3.20	1.75	1.77	0.72
26	10.82*	1.73	1.20	0.33	0.30	10.85	1.71	1.20	0.23
27	8.04	2.15	1.92	0.60	0.00	8.71	2.13	1.92	0.00
28	5.88	0.88	1.53	0.71	0.00	6.67	0.86	1.53	0.00
29	7.94	0.57	1.11	0.00	0.76	8.20	0.63	1.12	1.10
30	6.42	2.58	1.29	0.00	0.00	6.42	2.58	1.29	0.00
Aggregated	27.85*	1.29	1.46	0.33	0.00	30.69*	1.28	1.46	0.00

Note. DNM = decision noise model; LCJ$_r$ = restricted case of the law of categorical judgment.
* $p < .05$.

in fact present, which raises doubts regarding any claims based on the restricted DNM's parameter estimates for recognition memory data. Regarding Benjamin et al.'s explanation for the claimed difference, note that an equally plausible explanation for the opposite prediction can be entertained: In most circumstances, the mean position of the classification criterion is expected to be around the intersection of the target and distractor evidence distributions. This is the region where a greater uncertainty regarding the stimulus' class membership exists. It is plausible to expect that there is a higher response variability in situations of uncertainty than in cases where there is a great level of confidence regarding the stimulus's class.

Concerning the LCJ$_r$ model, no significant differences were found in the focused test of $\sigma_c = 0$ for the individual data sets (largest $\Delta G^2 = 2.49$, $p = .06$), for the summed results ($\Delta G^2 = 5.91$, $p = .96$), or for the aggregated data ($\Delta G^2 = 0.00$, $p = .50$).

Overall, the present results starkly contrast with previous claims that attributed a great magnitude to response criteria variability. Not only are the estimates of criterion noise small and in many cases close to zero, they also fail to provide a statistically significant contribution to account for the data. This failure is found for different implementations of criterion noise.[9] Also, note that a similar outcome is obtained with Benjamin et al.'s (2009) data as soon as problematic parameter restrictions are lifted, not only corroborating the present findings but also generalizing them across paradigms.

Given that the present results indicate an almost complete absence of criterion noise, it is important to test whether the generalization of parameters across tasks is adequate. The fact that criterion noise estimates are low might simply result from inadequate assumptions that lead to its underestimation. One form of reassurance comes from testing the generalization of parameters across tasks, as these should hold in the absence of criterion noise given the predictions stemming from the generalized area theorem. The parameter restriction that implements the model generalization (i.e., the restriction that sets μ_t and σ_t to be equal across tasks) is not rejected for the individual goodness-of-fit values (highest $\Delta G^2(2) = 5.26$, $p = .07$), with the exception of only one participant ($\Delta G^2(2) = 6.75$, $p < .05$), it is not rejected for the summed individual values ($\Delta G^2(60) = 68.73$, $p = .21$) but is rejected when considering the aggregated data sets ($\Delta G^2(2) = 8.19$, $p < .05$).

[9] Equivalent results were obtained with a model that implemented the response rule of the law of categorical judgment (symmetrically corrected), proposed by Klauer and Kellen (2012).

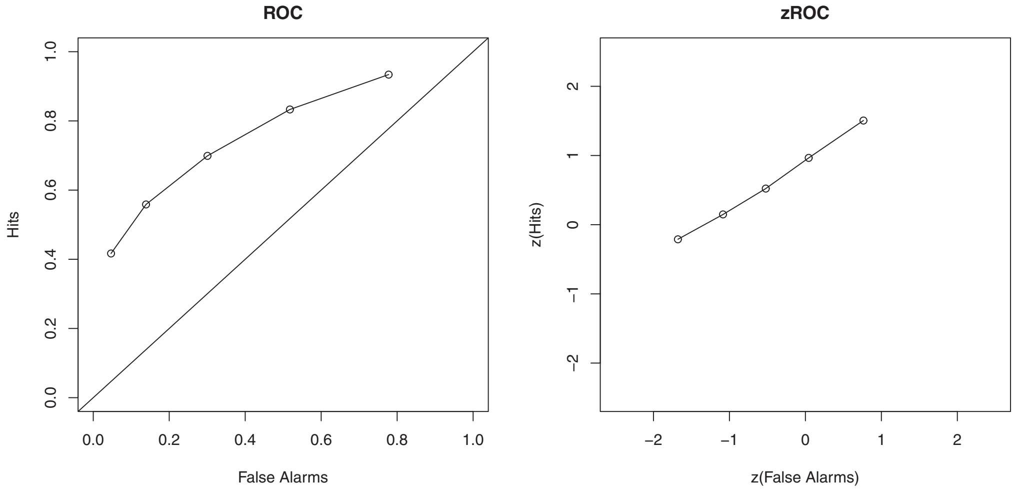

Figure 3. Receiver operating characteristic (ROC) and *z*ROC (ROC plot on a probit scale) plots for the aggregated data obtained in the reported experiment.

The latter rejection is not surprising given that the large sample size provides the statistical test with sufficient power to detect even tiny model violations, and that data aggregation might lead to distorted parameter estimates for each task. Similar results are found when comparing AIC_c and BIC values for the SDT and SDT_{sep}, as reported in Table 5. Taken together, the results show that the predictions made by the generalized area theorem (Iverson & Bamber, 1997) were not strongly violated, suggesting that the criterion noise estimates obtained are adequate.

Assessing the Sensitivity of Criterion Noise Measurement

Both the results from the reanalysis of Benjamin et al.'s (2009) experimental data and the present results converge to the same conclusion: The inclusion of a criterion noise parameter does not seem to provide any significant improvement to the models' account of the data. Nevertheless, before making any strong claims regarding the presence or absence of criterion noise based on these results it is important to take into account the ability of the two methods of detecting the presence of criterion noise of different magnitudes. As previously pointed out by Rosner and Kochanski (2009), a SDT model that assumes no criteria variability whatsoever can still account for data generated by a SDT model with criteria variability unless very large numbers of trials are used. This requirement can perhaps be easily fulfilled in other fields such as visual or auditory perception, but it is especially problematic for the case of recognition memory, as sample sizes in individual data sets are by necessity relatively small. The limitations in sample size are especially problematic for the two methods discussed here, as they require several items to be presented on each trial. A common way to overcome the limitation of individual data sets is to aggregate them and analyze the resulting group data (Cohen et al., 2008).

It might be the case that a model that precludes criterion noise might be preferred even when in fact there is criterion noise, and that the probability of this preference might be radically different between individual and aggregated data sets. Assessing the sensitivity of both methods to the presence of criterion noise for

Table 5
Goodness-of-Fit and Model Selection Results for the Reported Experiment

	Individual data sets				Aggregated data set			
Model	G^2	*df*	AIC_c	*BIC*	G^2	*df*	AIC_c	*BIC*
SDT	215.13	180	646.64	1,412.92	30.69	6	44.70	94.43
SDT_{sep}	146.42	120	705.04	1,686.44	22.50	4	40.52	104.44
LCJ_r	209.22	150	704.06	1,578.13	30.69	5	46.71	103.53
DNM_r	206.89	150	701.74	1,575.80	30.69	5	46.71	103.53
DNM	191.14	120	749.76	1,731.16	27.85	4	45.87	109.80

Note. The values under the label "individual data sets" are the sums of the individuals' values. These results for the individual data sets thereby represent a single model with different parameters for each individual. For all models, $p < .05$. AIC_c = corrected Akaike information criteria; *BIC* = Bayesian information criteria; SDT = signal detection theory; SDT_{sep} = a SDT model that excludes criterion noise and allows mnesic evidence parameters (μ_t and σ_t) to differ across tasks; LCJ_r = restricted case of the law of categorical judgment; DNM = decision noise model; DNM_r = a restricted version of the DNM.

different types of data sets is therefore important. Still, assessing the sensitivity of these methods is not straightforward as the impact of criterion noise on the individuals' performance is expected to be influenced by several additional factors beyond the nature and magnitude of criterion noise, such as the individuals' mnesic discriminability as defined by parameters μ_t and σ_t, as well as the positioning and distance between the mean response criteria (e.g., Macmillan et al., 2004). For the case in which there is more than one criterion noise parameter (i.e., classification and confidence noise parameters), the relative magnitude of these is also expected to play a role.

To test the sensitivity of these methods, individual-sized and group-sized data sets were generated using the parameter estimates obtained from the respective models without criterion noise, for the aggregate data sets. Besides the different approaches used, two important differences between the experiment reported and the ensemble recognition experiment reported by Benjamin et al. (2009) are the number of trials and the number of participants: Whereas in the ensemble recognition experiment, a total of 180 trials were collected per participant, for a total of 19 participants, in the experiment reported here, a total of 300 trials were collected per participant, for a total of 30 participants. These differences were taken into account in the data simulations. Note that the only difference between the individual and aggregated data sets is the total amount of trials, as the same parameter estimates are used to generate the artificial data sets.

The choice of using the parameter estimates obtained with the aggregate data to generate artificial data sets rests on the notion that they might be seen as representative of a "stereotypical" individual, instead of using arbitrary parameter values as done by Benjamin et al. (2009) and Rosner and Kochanski (2009). For the case of Benjamin et al.'s models, the parameters obtained with the original data sets did not assume any restriction of the mean response criteria across ensembles. Different values of criterion noise were then used along with the parameters obtained for the aggregated data to generate the data sets using different models. Criterion noise was varied between $\sigma_c = 0.50$ and $\sigma_c = 2$ in steps of 0.50.

For the simulations regarding the ensemble recognition experiment here reported, the averaging and summation models were used to generate the individual and group-sized data sets. The generated data were then fitted to the original data-generating model and the integration model. For the model generalization experiment here reported, data sets were generated by the LCJ$_r$ and DNM$_r$ models. In total, three factors were taken into account for the simulation of the two experimental methods available for measuring criterion noise: (1) the data-generating model, (2) the size of the data set, and (3) the level of criterion noise considered. For each combination of these factors, 1,000 data sets were generated and fitted.

The results presented in Table 6 indicate that for individual data sets, the rejection of the null hypothesis ($\sigma_c = 0$) in tests addressing the individual ΔG^2 values is unlikely unless large values of criteria variability are expected. Despite the differences in the model predictions, these are too small relative to the sampling variability of the data to be detected reliably. The picture is somewhat different for the aggregated data sets, in which the proportion of rejections of the false null hypothesis is quite higher. Unsurprisingly, the difference in the sample size between the present experiment and the ensemble recognition experiment reported by Benjamin et al. (2009) is reflected in the simulation results, with higher detection rates for the present experiment. Additionally, the results for the averaging model follow a nonmonotonic function, with detection rates increasing as criterion noise becomes larger but then decreasing for the largest values of criterion noise. One reason for this unexpected pattern is the manner in which criterion noise operates on the ensemble recognition task: For larger values of criterion noise, individuals' performance approximates chance level for all ensemble sizes, which compromises the discriminability between models, whether they incorporate criterion noise. Note that the absence of criterion noise in Benjamin et al.'s data does not reflect this "masking" effect of

Table 6
Simulation Results for the Sensitivity of the Different Models to Detect Criterion Noise of Various Sizes

Data	Model	$\sigma_c = 0.50$	$\sigma_c = 1$	$\sigma_c = 1.50$	$\sigma_c = 2$
		Benjamin et al.'s (2009) experiment			
Individual	Averaging-σ_c	84	58	32	23
	Summation-σ_c	53	67	64	74
Aggregated	Averaging-σ_c	472	679	617	539
	Summation-σ_c	136	426	611	658
		Present experiment			
Individual	LCJ$_r$	118	327	602	793
	DNM$_r$	199	431	673	831
Aggregated	LCJ$_r$	746	1,000	1,000	1,000
	DNM$_r$	745	1,000	1,000	1,000

Note. The values show the number of times in which the hypothesis restricting criterion noise to be 0 (H_0: $\sigma_c = 0$, $p < .05$) was rejected for 1,000 simulation runs per cell of the table. The parameter values used to generate the data sets (with the exception of the criterion noise parameters) correspond to the estimates obtained with these models for their respective experiments, when fitting the aggregated data sets and assuming the absence of criterion noise. LCJ$_r$ = restricted case of the law of categorical judgment; DNM$_r$ = a restricted version of the decision noise model.

extreme criterion noise values. If that would be the case, then the SDT model without criterion noise would give parameter estimates consonant with chance performance ($\mu_t \approx 0$ and $\sigma_t \approx 1$), which did not occur (see Table 2). Note that this pattern of results did not extend to the summation model.

The reason is that for the summation model, differences in ensemble size lead to a complete rescaling of the target and distractor evidence distributions, as previously discussed. As ensemble size increases, the variances of the evidence distributions increase as well, reducing the relative magnitude of criteria variability.

Overall, the results suggest that if criterion noise is present in the reported experiment, it has a quite low value, as criterion noise values of the magnitude suggested by the studies of Mueller and Weidemann (2008) and Benjamin et al. (2009) would have led to a frequent rejection of the null hypothesis for the individual data sets. Despite the low statistical power, the results here presented are far from being inconsequential, as they allow the dismissal of the notion that large values of criteria variability are normally present.

Furthermore, these results are in agreement with Rosner and Kochanski's (2009) results: Unless large sample sizes are used, a SDT model that precludes criterion noise can still provide an adequate account of the data. This issue highlights an important aspect in the implementation of SDT in different fields that should not be overlooked: Although the models and methods stemming from SDT that are used might be exactly the same, their effectiveness varies greatly among fields given their specific constraints. Although in perceptual tasks it is often easy to collect several hundreds of trials per participant, it is in most cases not possible to do this when studying recognition memory.

General Discussion

The failure of a model to account for experimental results usually leads to its reformulation and to the questioning of its basic assumptions. For the case of SDT, the results of Balakrishnan (1998, 1999) and Van Zandt (2000) motivated the development of solutions that focused on the notion that the assumption of criteria invariance is not only fundamentally wrong but is also undermining the model's account of the data. The present work indicates that criteria variability—irrespective of the way it was specified—seems to be quite low, to the point that it cannot provide any noticeable improvement to the account already given by traditional SDT. This result contrasts with the conclusions reached by previous research, which suggested that the magnitude of criteria variability is equivalent to or even greater than mnesic variability, and that its contribution could be easily detected with a model that allowed the estimation of mnesic and response processes. These previous conclusions have been shown to be unwarranted given the several shortcomings and confoundings discussed earlier.

The results obtained dispel the notion that the SDT model without criteria variability does not provide an adequate description of participants' performance. As discussed by Benjamin et al. (2009), disparate effects in various fields, such as response conservatism and probability matching (e.g., Thomas & Legge, 1970), or the effect of fatigue on performance (e.g., Galinsky, Rosa, Warm, & Dember, 1993), might result from unaccounted criteria variability. These possibilities hinged on the assumed existence of high criteria variability, which the present results do not support. The attribution of such effects to the presence of criterion noise is perhaps premature.

Still, one could argue that the overall results here reported are relatively inconsequential, as the forms of criterion noise here considered are only reliably detected when assuming large values. This would mean that criterion noise still represents a possible explanation for findings that are hard to interpret within the classic SDT framework. This reasoning fails to take into account that when criterion noise was estimated, it barely affected the estimates of μ_t and σ_t (compare parameter estimates when SDT is fitted to both tasks jointly, reported in Table 3, and the LCJ$_r$ and DNM's parameter estimates reported in Table 4). This means that low values of criterion noise produce small distortions in memory parameters and, hence, small divergences from traditional SDT predictions, whereas severe divergences from these predictions (e.g., Malmberg & Xu, 2006; Van Zandt, 2000) that have been associated to criterion noise require considerably larger values of criteria variability to take place.

It is important to note that the present results do not constitute an argument against the existence of criteria variability. The results constitute an argument against the claim that criterion noise (as currently modeled) has a major influence on recognition memory performance. Our results show either low values of criterion noise that seem to barely affect individuals' performance, or the absence of criterion noise. Additionally, the lack of statistical power for low levels of criterion noise, especially for the case of individual data sets, indicates the possibility of low values of criterion noise remaining undetected and highlights the limitations that recognition memory research is subjected to in comparison to other fields that use the same measurement models. Future work in recognition memory modeling needs to take into account that the development of more fine-grained models is limited by the quality of the data available. This notion has led to the argument that given the existence of such limitations, it is perhaps more fruitful that one's efforts are focused on more crude and basic aspects of the models (e.g., Rouder, Pratte, & Morey, 2010, p. 432).

The reanalysis of Benjamin et al.'s (2009) results as well as the results obtained with the model generalization approach encourage a global evaluation of the limitations of these two methods. The ensemble recognition task has several limitations: First, it requires that individuals integrate the evidence values provided by *each* element in the ensembles in a *specific* (sum or average) and *stable* manner. One can conceive of several idiosyncratic rules that take into account all the elements in an ensemble or only part of them, and the question of whether individuals use the same rule across ensemble sizes is still open. The literature on strategy identification in decision-making documents many cases where individuals only consider a fraction of the information available and the difficulty in identifying the specific manner in which individuals integrate (or not) the information available (e.g., Glöckner, 2009; Moshagen & Hilbig, 2011). Second, the use of information integration rules makes the models susceptible to several problems, as previously shown. In particular, restricting the (mean) response criteria to be the same across ensemble sizes results in the introduction of additional problems such as enforcing implausible predictions, as previously discussed. Third, the sensitivity analyses reported in Table 6 indicate that the probability that criterion noise provides a statistically significant contribution in the ensemble recognition task is very low.

The model generalization approach also has limitations: Despite having a greater sensitivity to presence of criterion noise than the ensemble recognition approach, this sensitivity is still quite low. In addition, one needs to assume that the generalization made by SDT across tasks holds. One possibility is that individuals use a different approach to the ranking task than the one specified, although this possibility would have wider implications for the SDT in general, as it would represent the existence of boundary cases for SDT. Such a scenario would seriously compromise the suitability of SDT to serve as a generalized cognitive measurement model and would indicate the need for adjustments. Overall, the ensemble recognition approach has a greater number of limitations than the model generalization approach, a difference that recommends the use of the latter method in further studies. Still, the complete dismissal of the ensemble recognition approach is perhaps too harsh a judgment given that the measurement of criterion noise in recognition memory is still in its infancy and that further refinements and adjustments might overcome the current problems.

As previously stated, the notion that the positioning of response criteria is a noise-free process is implausible, but the question that cognitive modelers need to ask is whether criteria variability has an influence that necessarily needs to be taken into account. What the present results suggest is that the SDT assumption of criterion invariance, although most likely wrong, still constitutes a valid approximation to individuals' performance, given the constraints of the experimental design. Questioning the existence of a process should not be confused with questioning its relevance, especially its relevance within a specific experimental context. Furthermore, it is interesting that the arguments for the relevance of criterion noise in recognition memory are mostly based on findings from other fields, findings that therefore have a limited value for the case of human memory. Too often findings related to SDT coming from different sources (e.g., perception) are used as evidence in rather distinct fields (e.g., reasoning; see Dube et al., 2010). Although SDT provides a general framework that can be used in various fields, the notion that the nature of the underlying processes might be radically different cannot be ignored (Bröder & Schütz, 2009, p. 599). The usage of SDT in different fields hinges more on the fact that the resulting data matrices from these fields have a similar structure, making SDT a convenient data analysis tool, than on the existence of a general unifying theory. The tendency of researchers to engage in such generalizations (and their dangers) has been discussed in the literature as the "tools-to-theories" heuristic (Gigerenzer, 1991). The discussion of hypotheses and findings reported in other fields is of course very important, as it inspires and informs different research approaches, but the mere assumption that the findings generalize across fields is rarely justified.

Although the criterion noise models proposed so far do not appear to provide a better description of the data, there are many issues related to criterion noise that remain to be solved: For example, the original results from Van Zandt (2000) still need to be explained by a SDT model that provides a description of both mnesic and response processes. Although the presents results indicate a quite low or absent criteria variability, Van Zandt's results are consistent with the predictions made by Treisman and Williams's (1984) model of criterion variability (see Treisman & Faulkner, 1984), a model proposed in the perception literature. Is there any way to reconcile these apparently contradictory results? One possibility is that criterion noise is related to response bias, more specifically experimentally induced response bias. In the experiments reported by Van Zandt, individuals provide confidence ratings under different base rate (ranging from 10% new items and 90% old items to 90% new items and 10% old items) or payoff conditions. It could be that memory testing in biased conditions—either toward "old" or "new" responses—creates a conflicting situation for participants that generates criterion noise. Wixted and colleagues (Mickes, Hwe, Wais, & Wixted, 2011; Wixted & Gaitan, 2002) have argued that individuals' positioning of response criteria is pre-conditioned by extensive error feedback that these individuals received throughout their life. When encouraged to produce biased responses, individuals would attempt to set the response criteria accordingly, but these positionings would be in conflict with the response criteria positions that previous experience has conditioned them to use. The conflict between these different positionings could effectively lead to the emergence of criterion noise, as these opposing tendencies would likely compromise the stability of response criteria in the recognition task, perhaps affecting the classification and confidence criteria differently. This would mean that criterion noise is mostly present in test conditions that encourage response bias, which would explain the findings reported by Van Zandt as well as the present results. Nevertheless, this explanation is speculative, and future research efforts should focus on the measurement of criterion noise in test conditions that induce response biases to test different hypotheses, although the use of the model generalization approach would face serious difficulties as different sets of parameters would have to be estimated for each bias condition, reducing even further the number of trials per test condition. One way of overcoming this problem would consist in the specification of hierarchical models, which so far has only been done for SDT models without criterion noise (e.g., Pratte et al., 2010).

Another issue that needs to be considered concerns the sequential dependencies found in the responses given on recognition tests (e.g., Malmberg & Annis, 2011; see also Criss, Malmberg, & Shiffrin, 2011; Malmberg, Criss, Gangwani, & Shiffrin, 2012). The existence of these sequential dependencies seems to be at odds with the results reported here, as it is not precisely clear how these dependencies could occur in the absence of criteria variability. One reason for this apparent paradox lies in the fact that the criterion noise processes here considered do not take into account previous trials. In fact, the positioning of response criteria is ensured to be independent across trials. This means that the criterion noise processes considered here fail to provide any direct description of sequential dependencies. In addition, the sequential dependencies so far observed in recognition memory data (Malmberg & Annis, 2011) seem to be rather at odds with some of the predictions made by Treisman and Williams's (1984) model of criterion variability (see Malmberg & Annis, 2011), a result that raises the question of whether or not the latter model can account for both Van Zandt's (2000) results as well as Malmberg and Annis's (2011) results. This situation should encourage the development of new models that include criterion variability processes that are influenced by prior responses. Still, the estimation of several forms of criterion noise (see Klauer & Kellen, 2012; Rosner & Kochanski, 2009) already represents a formidable challenge due to the difficulty of obtaining accurate numerical estimates. The development of more complex models will add a new set of difficulties that future research efforts will need to overcome. In addition, one needs to consider that the loci of the observed sequential dependencies does necessarily lie in the decision processes but could also emerge from fluctuations of the mnesic processes. The possibility of sequential dependencies arising from differences in discriminability across trials has been discussed in the perception literature (e.g., Atkinson, 1963; S. D. Brown, Marley, Donkin, & Heathcote, 2008) and should not be overlooked in future efforts in the field of recognition memory.

The latter point raises the question of whether future refinements of the SDT model necessarily imply the inclusion of criterion noise: The assumption that criteria positioning is a noise-free process is not the only assumption of SDT that is generally considered implausible and that can cause the documented inadequate descriptions of individuals' performance. For example, Turner, Van Zandt, and Brown (2011) proposed an alternative framework for SDT in which stimulus evidence distributions evolve across time, reflecting the statistical properties of the previously encountered experimental trials. Within this framework, decisions are solely based on likelihood ratios, estimated from the evidence available to the decision maker up to that moment, precluding the establishment of response criteria along the evidence axis. This alternative framework can successfully account for several findings in the literature, although it does not provide a measurement model like the traditional SDT approach does.

Furthermore, the traditional SDT model assumes that the information available as evidence is based on a fixed evidence sample, precluding the possibility of dynamical accumulation of information and thus making the model incapable of accounting for several effects such as speed–accuracy tradeoffs (e.g., Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009). Relaxing such assumptions toward a dynamical signal detection model (e.g., Balakrishnan & MacDonald, 2011) may lead to an alternative solution for the inconsistencies so far encountered in the literature.

References

Atkinson, R. C. (1963). A variable sensitivity theory of signal detection. *Psychological Review, 70,* 91–106. doi:10.1037/h0041428

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

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

Balakrishnan, J. D., & MacDonald, J. A. (2002). Decision criteria do not shift: Reply to Treisman (2002). *Psychonomic Bulletin & Review, 9,* 858–865. doi:10.3758/BF03196345

Balakrishnan, J. D., & MacDonald, J. A. (2011). Performance measures for dynamic signal detection. *Journal of Mathematical Psychology, 55,* 290–301. doi:10.1016/j.jmp.2011.05.001

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (p. 175–223). London, England: Academic Press.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51,* 159–172. doi:10.1016/j.jml.2004.04.001

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

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—Or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 587–606. doi:10.1037/a0015279

Brown, J. (1965). Multiple response evaluation of discrimination. *British Journal of Mathematical and Statistical Psychology, 18,* 125–137. doi:10.1111/j.2044-8317.1965.tb00696.x

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

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York, NY: Springer.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44,* 171–189. doi:10.1006/jmps.1999.1282

Chechile, R. A., & Soraci, S. A. (1999). Evidence for a multiple-process account of the generation effect. *Memory, 7,* 483–508. doi:10.1080/741944921

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review, 15,* 692–712. doi:10.3758/PBR.15.4.692

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316–326. doi:10.1016/j.jml.2011.02.003

Criss, A. H., & McClelland, J. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55,* 447–460.

Dalrymple-Alford, E. C. (1970). A model for assessing multiple-choice test performance. *British Journal of Mathematical and Statistical Psychology, 23,* 199–203. doi:10.1111/j.2044-8317.1970.tb00444.x

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54,* 304–313. doi:10.1016/j.jmp.2010.01.001

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117,* 831–863. doi:10.1037/a0019634

Duyck, W., Desmet, T., Verbeke, L., & Brysbaert, M. (2004). WordGen: A tool for word selection and non-word generation in Dutch, German, English, and French. *Behavior Research Methods, Instruments & Computers, 36,* 488–499. doi:10.3758/BF03195595

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12,* 403–408. doi:10.3758/BF03193784

Feller, W. (1966). *An introduction to probability theory and its applications.* New York, NY: Wiley.

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

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98,* 254–267. doi:10.1037/0033-295X.98.2.254

Gilden, D. L., & Wilson, S. G. (1995). On the nature of streaks in signal detection. *Cognitive Psychology, 28,* 17–64. doi:10.1006/cogp.1995.1002

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review, 16,* 431–455. doi:10.3758/PBR.16.3.431

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making, 4,* 186–199.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin, 66,* 228–234. doi:10.1037/h0023645

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

Hautus, M. J., Macmillan, N. A., & Rotello, C. B. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review, 15,* 889–905. doi:10.3758/PBR.15.5.889

Iverson, G., & Bamber, D. (1997). The generalized area theorem in signal detection theory. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (p. 301–318). Hillsdale, NJ: Erlbaum.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138,* 291–306. doi:10.1037/a0015525

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology, 55,* 251–266. doi:10.1016/j.jmp.2010.11.004

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75,* 70–98. doi:10.1007/s11336-009-9141-0

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. doi:10.1037/a0025824

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

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation. *Perception & Psychophysics, 68,* 393–414. doi:10.3758/BF03193685

Lahl, O., Göritz, A. S., Pietrowsky, R., & Rosenberg, J. (2009). Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods, 41,* 13–19. doi:10.3758/BRM.41.1.13

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

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

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics, 66,* 406–421. doi:10.3758/BF03194889

Maddox, W. T., & Bohil, C. J. (1998). Base-rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1459–1482. doi:10.1037/0278-7393.24.6.1459

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57,* 335–384. doi:10.1016/j.cogpsych.2008.02.004

Malmberg, K. J., & Annis, J. (2011). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General.* Advance online publication. doi:10.1037/a0025277

Malmberg, K. J., Criss, A. H., Gangwani, T., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference that results from recognition memory testing. *Psychological Science, 23,* 115–119. doi:10.1177/0956797611430692

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review, 13,* 99–105. doi:10.3758/BF03193819

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

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1095–1110. doi:10.1037/0278-7393.28.6.1095

Moshagen, M., & Hilbig, B. E. (2011). Methodological notes on model comparisons and strategy classification: A falsificationist proposition. *Judgment and Decision Making, 6,* 814–820.

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

Murdock, B. B. (1963). An analysis of the recognition process. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behavior and learning* (pp. 10–32). New York, NY: McGraw-Hill. doi:10.1037/11178-002

Parks, C. M., & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Sciences, USA, 106,* 11515–11519. doi:10.1073/pnas.0905505106

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8–13. doi:10.1016/j.jneumeth.2006.11.017

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin & Review, 14,* 379–391. doi:10.3758/BF03194081

Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of confidence, choice, and response time. *Psychological Review, 117,* 864–901. doi:10.1037/a0019737

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 224–232. doi:10.1037/a0017682

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785. doi:10.1037/0278-7393.20.4.763

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

R Development Core Team. (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

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. doi:10.1037/a0014463

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review, 17,* 427–435. doi:10.3758/PBR.17.3.427

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82,* 605–610. doi:10.2307/2289471

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika, 72,* 133–144. doi:10.1093/biomet/72.1.133

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

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34,* 125–137. doi:10.3758/BF03193392

Solomon, J. A. (2007). Intrinsic uncertainty explains second responses. *Spatial Vision, 20,* 45–60. doi:10.1163/156856807779369715

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379–1396. doi:10.1037/0278-7393.24.6.1379

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

Thomas, E. A., & Legge, D. (1970). Probability matching as a basis for detection and recognition decisions. *Psychological Review, 77,* 65–72. doi:10.1037/h0028579

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

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

Treisman, M., & Faulkner, A. (1984). The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical and Statistical Psychology, 37,* 199–215. doi:10.1111/j.2044-8317.1984.tb00800.x

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

Turner, B., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118,* 583–613.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. doi:10.1037/0278-7393.26.3.582

Verde, M. F., & Rotello, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review, 11,* 560–566. doi:10.3758/BF03196611

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

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

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

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133,* 800–832. doi:10.1037/0033-2909.133.5.800

Received September 23, 2011
Revision received December 19, 2011
Accepted February 7, 2012 ■