Psychological Review
2013, Vol. 120, No. 3, 720–726

0033-295X/13/$12.00 DOI: 10.1037/a0031911

COMMENT

Where Is the Criterion Noise in Recognition? (Almost) Everyplace You Look: Comment on Kellen, Klauer, and Singmann (2012)

Aaron S. Benjamin
University of Illinois at Urbana-Champaign

Recent articles, including Benjamin, Diaz, and Wee (2009), have argued that recognition memory may be better understood if consideration is given to sources of noise in the *decisions*, as well as to those in the *representations*, underlying recognition judgments. They based that conclusion on a wide consideration of persisting mysteries in recognition research as well as a new experimental paradigm involving *ensemble recognition*. Kellen, Klauer, and Singmann (2012) reanalyzed Benjamin et al.'s data and introduced their own new experimental paradigm to this debate. They concluded that criteria do not vary much from trial to trial in recognition testing and, thus, that decision noise in recognition is small or nonexistent. However, their alternative interpretation of Benjamin et al.'s data relies on a questionable conclusion to reject all models in which the locations of criteria are restricted to be the same across ensembles and a meta-assumption that a model should be rejected as false if it yields unconventional parameters. In addition, their experimental logic relies on the assumption that ranking tasks are always bias-free. Here, I question these assumptions and suggest avenues for reconciliation between these contrasting claims.

Keywords: recognition, criterion noise, decision noise, signal detection, recognition memory

Ever since a set of landmark articles by Egan (1958), Parks (1966), and Banks (1970), the theory of signal detectability (TSD; Green & Swets, 1966; Macmillan & Creelman, 2005) has been a dominant framework for understanding the processes involved in the *recognition memory* task. In that task, subjects evaluate whether individual test stimuli were experienced previously in a particular delimited context. The most prominent contribution of that framework is the explicit superposition of a statistical point of view onto the cognitive task of recognition. That is, the theory is guided in part by a consideration of sources of *noise* in memory.

TSD postulates specifically that trials vary randomly in the amount of evidence they yield to the decision maker. In the case of perceptual or attention tasks, this noise is presumed to arise from physical and perceptual fluctuations in signal transmission. In the case of recognition memory, in which repeated trials invariably involve different stimuli, the amount of noise also reflects the idiosyncratic history that each unique stimulus and an individual's history with that stimulus bring to bear.

More recently, it has been suggested that the decision process—in which evidence is evaluated by reference to a *criterion*—may itself be a source of noise. Theories of recognition (Benjamin, Diaz, & Wee, 2009; Wickelgren, 1968; Wixted & Stretch, 2004) and perception (Bonnel & Miller, 1994; Durlach & Braida, 1969; Mueller & Weidemann, 2008; Nosofsky, 1983) that incorporate a role for criterion noise have been developed and applied to theoretical problems in those domains.

Benjamin et al. (2009) reviewed a large set of persisting theoretical puzzles in recognition memory research and suggested that a consideration of criterion noise might aid theoretical development. In addition, they gathered data from a new experimental task—the *ensemble recognition* task—in which it was possible to decompose the separate contributions of representational noise and criterion noise. They concluded that criteria exhibited approximately the same variability across trials as did stimuli and, thus, that decision noise contributed nontrivially to recognition judgments.

Kellen, Klauer, and Singmann (2012) re-analyzed Benjamin et al.'s (2009) data and reached the opposite conclusion—that criterion noise made no meaningful contribution to performance. They also conducted an experiment using another new task—the *k-alternative ranking task*—that supported their conclusion. In this commentary, I raise questions about the logic they applied to their re-analysis and to their new experimental procedure. In addition, I briefly review evidence from the literature indicating that criterion noise is ample and meaningfully influences recognition performance. However, aspects of Kellen et al.'s conclusions and results clearly pose difficulties for the *noisy decision theory of signal detection* (ND-TSD) proposed by Benjamin et al., and I conclude by suggesting empirical avenues for reconciliation between these opposing views.

Criterion Noise in Ensemble Recognition

The keystone results on which Benjamin et al.'s (2009) and Kellen et al.'s (2012) groups differ are the ensemble recognition data reported by Benjamin et al. In that task, subjects were asked to endorse ensembles of one, two, or four previously studied items

Correspondence concerning this article should be addressed to Aaron S. Benjamin, Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820. E-mail: asbenjam@illinois.edu

and reject ensembles composed of new items. If each ensemble invites comparison to only one criterion (or set of criteria, in the rating task), then ensemble size should be related to stimulus noise but not to criterion noise.

In our original report, we compared the performance of a number of models, some of which varied in whether (a) they permitted the presence of criterion noise and (b) the locations of decision criteria were allowed to vary freely across ensemble conditions. The first variable provided the critical test of the assumption of criterion noise—the central hypothesis of the article. The second variable was included because the conditions under which recognizers shift criteria across conditions are unclear and an object of much current research (e.g., Benjamin, 2001, 2005; Benjamin & Bawa, 2004; Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998; Verde & Rotello, 2007), and we wished to remain agnostic with respect to the criterion-setting process in the ensemble recognition task.

Kellen et al. (2012) evaluated the same models (and an additional one in which an alternative assumption about the combination of evidence across an ensemble was implemented). Both Benjamin et al. (2009) and Kellen et al. found that the best fitting model was one in which criterion noise was present and in which criteria were not free to vary across the ensemble conditions. It was this result that was the basis for the conclusion by Benjamin et al. that criterion noise was present and ample in recognition memory.

This result notwithstanding, Kellen et al. (2012) expressed skepticism over the winning criterion-noise models. They noted several concerning aspects of the fits of the model. First, the superior fit of the criterion-noise models did not generalize to an analysis in which the aggregated data were modeled. Second, the fits to individual subjects revealed a larger number of anomalous parameter estimates in the conditions in which criteria were fixed across ensemble size than conditions in which they were not, which they took to indicate instability in the model fits.

Kellen et al. (2012) noted that, among the subset of models in which the criterion-shift restriction was not implemented, the best fitting model did not include criterion noise. On the basis of further analysis of the criterion-shift restriction assumption, and of the unconventional parameter values of stimulus and criterion noise that were apparent in the best models, they concluded that the criterion-shift restriction should be rejected. They thus concluded from the subset of models still under consideration that criterion noise is nonexistent or trivial, and they argued that Benjamin et al. (2009) reached the incorrect conclusion because they considered a set of models that were fundamentally flawed.

Fits to Aggregated Data

Benjamin et al. (2009) did not consider how the models fared with respect to fitting the aggregated data, because they considered those data misleading with respect to the central question of the presence of criterion noise. At the heart of the analyses here is the question of whether criteria vary randomly from trial to trial. This determination is difficult and is rendered only possible, in fact, because the ensemble recognition task provides a means for separating decision-based and representation-based sources of noise. Detecting the presence of decision noise against a background that additionally includes individual-difference sources of variation makes accurate estimation more, not less, difficult. Though there are good reasons why aggregate data should be considered under some circumstances (Cohen, Sanborn, & Shiffrin, 2008), cases in which the group data clearly introduce variation in a parameter (e.g., the locations of criteria) that may or may not be present in individual data—*and are the central variable under investigation*—seem poor candidates for such an approach (Estes & Maddox, 2005).

Fits to Individual Subjects

The anomalous results from individual fits of the models with restricted criteria are a potential source of concern. In fact, one common approach in model fitting is to require the model-fitting algorithm to limit its search to acceptable parameter values, typically by placing boundaries on those values. Benjamin et al. (2009) chose not to pursue this approach because it was not clear to us exactly what would constitute a reasonable value. Allowing criterion noise in detection-theoretic models is a sufficiently major theoretical change that it seemed shortsighted to bring our intuitions to bear on what the appropriate range of parameters should be.

Kellen et al. (2012) made the decision to not consider models that employed the criterion-shift restriction for three reasons. First, those models make the prediction that the proportion of high-confidence response should increase with ensemble size. This prediction is clearly wrong in our data and, thus, does indeed speak against the plausibility of the criterion-shift restriction as implemented by Benjamin et al. (2009). Below, I consider alternatives to the strong version of this assumption that might allow the model to circumvent this problem.

Second, they reported a likelihood-ratio test in which models with and without the restriction were compared. Here, the evidence is less convincing. To start with, the restriction was rejected for less than one third of the subjects for the winning model of Kellen et al. (2012)—hardly a basis for rejecting the assumption out of hand for the entire sample. The only other evidence brought to bear here is the fact that the criterion-shift restriction was rejected by likelihood-ratio test in the aggregate data, which is unconvincing for the reasons reviewed earlier. Even if this evidence were convincing, there should be some concern about the contrasting results of hypothesis-testing approaches and goodness-of-fit approaches to quantifying a model's performance. I certainly agree with the authors that fit statistics, such as Akaike information criterion $(AIC)_c$ and Bayesian information criterion (BIC), should not be taken as the ultimate arbiter of a model's merit, but trading the occasional vagaries of fit statistics for the Pandora's box of conceptual difficulties associated with null hypothesis significance testing seems risky. This seems particularly true in this case, where the fit statistics tell a different story than G^2: for the criterion-noise models, AIC_c and BIC are lower for the models with the criterion-shift restriction than the ones without. In fact, the magnitude of the advantage is such that only a small portion of the effect can be attributed to the lesser penalty due to extra parameters paid by the models without the restriction. This result indicates that the restricted models are doing a better job of fitting the data, even if that restriction is rejected by the standards of null-hypothesis significance testing within a subset of subjects.

Finally, Kellen et al. (2012) appealed to the values of the parameters yielded by the winning model as evidence for its

failure. At the heart of the final conclusion by Kellen et al. to not consider those models with restricted criteria is the following logic: If a model yields unusual parameters, even if it fits the data well, it should be rejected. That is, the anomalous parameter values are taken to indicate a failing of the model, rather than the difficulty of fitting that model to a limited set of data. Such an action seems unwarranted. Again, I agree that the superior fit of the criterion noise models with restricted criteria should not be taken as the final word on that model's correctness. However, it seems equally shortsighted to take the unconventional fits of some individuals as ironclad evidence for its falsity.

Certainly, we can imagine cases in which illogical or uninterpretable parameter values could be used to veto the outcome of a comparison of fit statistics, like when those parameters violate known capacities of biology or physics. However, this is not the case here. Estimates of σ_C (the parameter that measures criterion noise, scaled in standard deviation units of the noise representation) range in Benjamin et al.'s (2009) results from 0.6 to 7.2.[1] Interestingly, the parameters governing the distance between the evidence distributions (μ) and the standard deviation of the signal representation (σ_1) are highly correlated with σ_C ($r = .86$; .58), indicating some parameter mimicry.

Because the parameters in TSD are measured by reference to variance in the noise distribution, which is fixed to an arbitrary value, all scaling is also completely arbitrary. One could just as easily fix a different parameter, like σ_1, and estimate the other parameters with reference to it. One way of appreciating the common factor underlying values of μ, σ_1, and σ_C is to fix one of those values and then rescale the other parameters relative to it. This is done in the middle panels of Figure 1 for two groups of subjects: those with lower (and thus more "traditional") values of σ_C (see Panel C), and those with higher values of σ_C (see Panel D).[2] These same parameters are shown without rescaling (i.e., by using the original estimates with fixed σ_0) in Panels A and B.

What is apparent here is that the "anomalous" values of σ_C evident in Panels A and B are not actually revealing of anything strange about criterion noise. Because μ, σ_1, and σ_C all scale together, another interpretation is that the σ_0 estimate is the value that exhibits the most variability between these two groups of subjects. That is, those subjects that yielded high estimates of criterion noise are actually ones that exhibit low variability in strength values for unseen items on the recognition test. Note that the other parameters look roughly the same between Panels C and D.

This reconceptualization—one that would not have been possible had the fitting algorithm used a constrained search of σ_C values or if the criterion-shift restriction been rejected out of hand—suggests an alternative understanding of the unexpected variability in the σ_C parameter. For some subjects, the range of evidence values for unstudied items is small, yielding a response profile that is more threshold-like in form. This can be seen in Panel E, which plots the isodiscriminability functions for both groups when the role of criterion noise is ignored. The low-σ_C group evidences a traditional function that is clearly curvilinear and with greater mass on the low end. The high-σ_C group exhibits a function that is roughly linear throughout much of its range, much like a function based on thresholds would look like (Egan, 1958; Krantz, 1969; Luce, 1963). Of course, because we generated the function, we know that there are no thresholds—that is, there is nonzero likelihood for both noise and signal throughout the evidence range.

When criterion noise is added to the mix, as shown in Panel F, the threshold-like behavior of the high-σ_C group is no longer apparent. The isosensitivity functions shown there are ones that would be easily accommodated in the extant literature without concern. So there is nothing particularly concerning about the parameter values that yielded this behavior, even though the higher function represents a subset of subjects with the most extreme values of the very parameter presumed by Kellen et al. (2012) to be anomalous. We should not dismiss a model simply because of unconventional parameter values without a thorough consideration of how those estimates might reveal entrenched biases in our conceptualization of the task.

Evidence From the *k*-Alternative *n*-Response Task

Kellen et al. (2012) introduced to this debate a very clever new task in which subjects evaluate k items and rank n of them with respect to evidence for oldness (Kellen & Klauer, 2011). Using the four-alternative forced-choice with two rankings (4AFC-2R) version of this task, they were able to estimate traditional detection-theoretic parameters under forced-choice responding conditions. Forced-choice response conditions are thought to be free of the need to establish and maintain criteria and, consequently, provide a "base rate" of variability against which estimates from the yes/no procedure could be compared.

Using this task, Kellen et al. (2012) found that criterion noise was not necessary to explain the discrepancies in performance between the rating task and the forced-choice task—that is, the forced-choice task did not yield substantively lower estimates of stimulus variance than the yes–no task (in which criterion noise would be included.) However, there are two major concerns about how their task might not yield entirely accurate parameter estimates.

Bias in Forced-Choice Responding

Green (1964) was the first to note a strong theoretical relationship between performance in forced-choice tasks and performance on yes–no tasks. This relationship is based on the assumption that forced-choice responding is *criterion free*—that is, that recognizers choose the alternative with the higher amount of evidence without comparing that amount of evidence to a criterion value. One consequence of such criterion-free decision making is that response biases should not be evident.

Evidence in support of this relationship has been spotty at best, however, within studies of recognition memory (Green & Moses, 1966). Of the subset of studies that used the unequal-variance version of signal detection theory now favored by most researchers (Wixted, 2007), the only articles that have confirmed the relationship (Jang, Wixted, & Huber, 2009; Smith & Duncan, 2004) used confidence ratings to fit the competing models. Although forced-choice recognition might plausibly not involve criteria, confidence

[1] The variance in these parameters is larger in the fits to the same data reported by Kellen et al. (2012, Table 2).

[2] This division was created by a median split within the subjects on σ_C, omitting the single subject in the middle of the range.

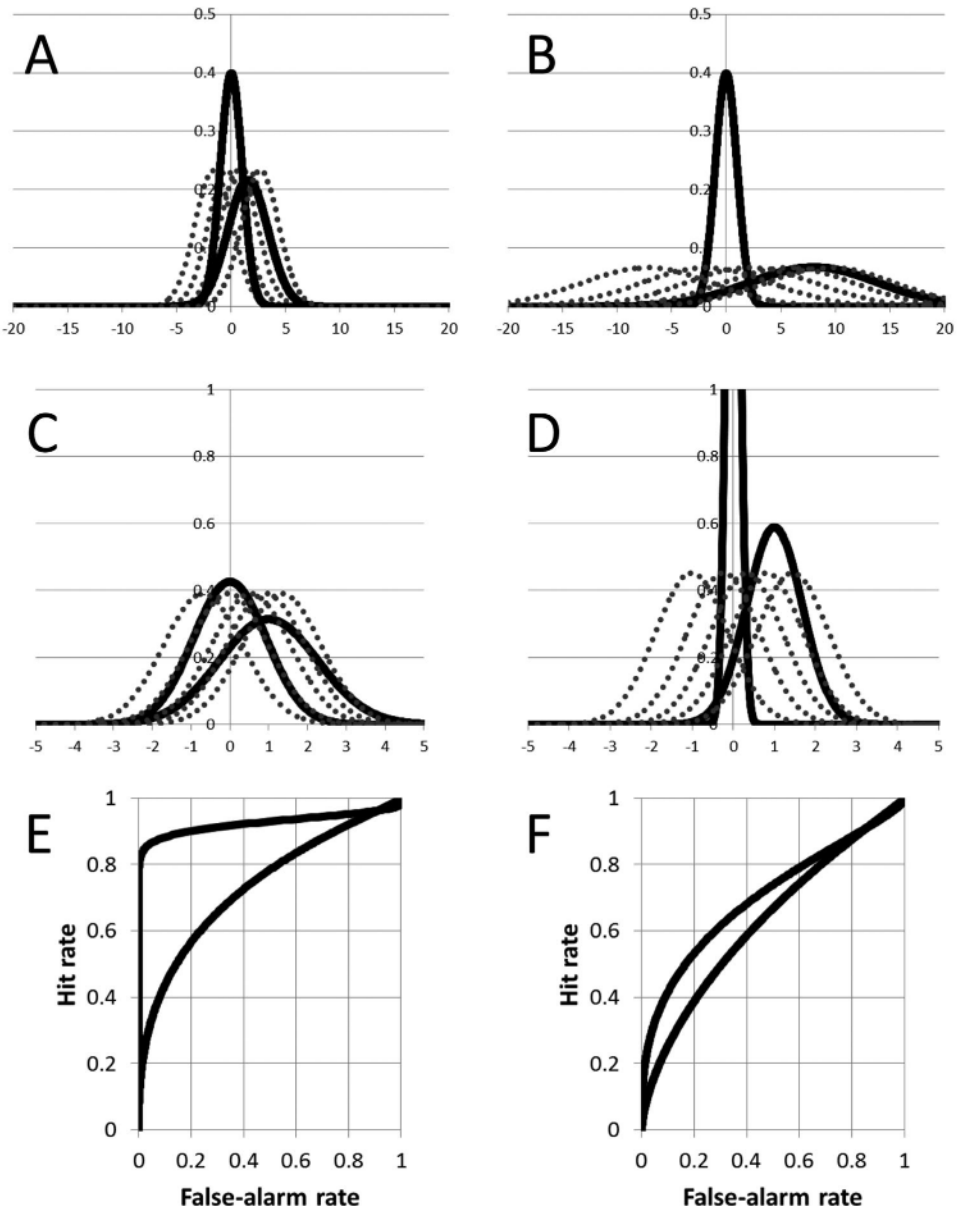

Figure 1. Panels A and B: Estimated probability distributions of evidence and criteria for the low-σ_c and high-σ_c group, respectively. Panels C and D: The distributions from Panels A and B rescaled to equivalent μ. Panels E and F: The implied isosensitivity functions for the two groups with and without criterion noise, respectively.

ratings almost certainly do—and there is no reason to think that those ratings will be any less noisy in the forced-choice task than in the yes–no task. Thus, the confirmed relationship between these tasks in those articles might not reflect the absence, but rather the presence, of criterion noise in both cases.

In addition, it has been noted that there can actually be considerable bias in forced-choice responding. Wickelgren (1968) described numerous influences that could serve to violate the criterion-free assumption of forced-choice responding, including position bias, correlated evidence values across intervals, variable or fluctuating attention, and forgetting across intervals. As Kellen et al. (2012) noted, some of these sources of bias are ones that are of greater concern in perceptual experiments, in which a temporal dimension is used to present the competing options in forced-choice tasks. Recognition testing usually does not use temporal intervals, so differential forgetting across the to-be-evaluated stimuli, for example, is less of a concern. However, there is no reason to assume that the perceptual dimensions (such as left–right display) in recognition testing do not induce bias, or to assume that stimuli are compared directly to each other rather than individually to a criterion. On the few occasions in which bias has been investigated within multi-alternative forced-choice testing, bias

has been apparent (e.g., Nisbett & Wilson, 1977; as cited in Ennis & O'Mahony, 1995; Macmillan & Creelman, 2005; as re-analyzed by DeCarlo, 2012; see also Klein, 1991). The logic of the comparison between yes–no and forced-choice tasks—and, thus, the soundness of Kellen et al.'s experimental procedure—relies critically upon unbiased responding.

In addition, there are numerous decision models that assume a probabilistic response rule in which the probability of choosing a specific alternative is a (positive) function of the likelihood of that alternative with respect to the competing options (akin to the Luce choice rule). Such a mechanism has been proposed for forced-choice responding in theories of categorization (Maddox & Ashby, 1993; Nosofsky & Zaki, 1998), perceptual decision making (Stüttgen, Yildiz, & Güntürkün, 2011), and memory for faces (e.g., Busey & Arici, 2009). The ubiquity of such a decision rule and the widespread evidence of probability matching, in numerous domains, suggest that forced-choice responding may not be as deterministic and criterion-free as TSD would suggest.

Task Strategy and Interleaved-Trial Designs

Even if we are not concerned about the effects of bias on forced-choice responding, there are aspects of Kellen et al.'s (2012) experimental procedure that have the potential to introduce violations of the criterion-free assumption in forced-choice responding. Specifically, trials of different types—yes/no and forced-choice response—were interleaved in their test. This choice to mix trials has the advantage of removing time-based confounds, such as differential recency across conditions, from the test. However, it has the negative consequence of affecting the hypothesized strategic approaches to the two tasks via a *carryover* effect.

Criterion-free responding is not the only potential approach to solving a ranking task. In a two-alternative task, for example, subjects can evaluate the first stimulus via the same criterion as would be used in a yes/no task and then respond "first" if the criterion is surpassed and "second" if not. In fact, within the eyewitness memory literature, such an approach has been argued to yield superior performance in lineup tasks than criterion-free relative comparison (though it should be noted, of course, that these eyewitness tasks have the additional complexity that the perpetrator, or "old item," might not be present in the lineup; Gronlund, 2004). Kellen et al.'s (2012) task used four, rather than only two, alternatives, but the logic is the same: Subjects may engage in the same processes as yes–no responding for a subset of the stimuli, rather than compare them directly to one another.

Tasks in which different trial types are interleaved exacerbate carryover effects. In recognition memory, for example, it is evident that subjects are reluctant to vary decision criteria across stimuli with different strengths when those stimuli are interleaved within a list (Morrell et al., 2002; Stretch & Wixted, 1998; Verde & Rotello, 2007) but are apt to do so when they are manipulated across lists (Hirshman, 1995). The concern here is that the interleaved design has the potential to induce a strategic approach to the forced-choice task in which the criterion-free assumption is violated. Under such conditions, the estimates between the two tasks will correspond not because criterion noise is absent in yes–no recognition, but rather because it is present in forced-choice recognition. It is also worth pointing out that Benjamin et al. (2009) also used an interleaved design, in which ensemble sizes were randomly intermixed at test. That design also has the possibility of introducing strategic homogeneity across trials that might not be evident in a between-list or between-subject design. For both tasks, replication using alternative designs will be important for adjudicating between the theoretical possibilities.

Finally, it is worth noting that the correlation between estimates of variability across subjects for the two tasks is only $r = .20$. (The values are taken from the third and sixth columns in Table 3 of Kellen et al., 2012.) Though they concluded that a common model for both tasks provided a superior fit to their data, the variability estimates for the rating task and the ranking task have less than 5% of their variance in common. If these tasks did indeed induce ideal criterion-free responding in ranking and not in rating, and if criterion noise was not present, one might expect this correspondence to be more impressive in magnitude.

Criterion Variability in the Larger Literature on Recognition

In cases like this where there is dispute over the appropriate interpretation of a set of results, certainly the nearly 100 years of literature on recognition and discrimination tasks more generally must provide some guide. Benjamin et al. (2009; see also Benjamin, Tullis, & Lee, in press) reviewed a large number of results and concluded that criterion noise provided a simple and parsimonious way of understanding enigmatic phenomena like nonstationarity (Treisman & Williams, 1984), failures to confirm the theoretical relationship between forced-choice and yes–no recognition procedures (Green & Moses, 1966), variability in the slope of the isosensitivity function across learning conditions when plotted in normal-deviate coordinates (i.e., the zROC; Glanzer, Kim, Hilford, & Adams, 1999), lack of equivalence between confidence-rating and bias-induction recognition procedures (Van Zandt, 2000), probability matching (Lee, 1963), conservatism in response to base rate manipulations (Healy & Kubovy, 1978), effects of aging on the slope of the zROC (Kapucu, Rotello, Ready, & Seidl, 2008), and variation in the slope of the zROC for "remembered" items (Wixted & Stretch, 2004). In addition, criterion variability has been convincingly demonstrated in perceptual tasks (Bonnel & Miller, 1994; Nosofsky, 1983) and plays a critical role in sampling models of recognition (Ratcliff & Rouder, 1998).

Kellen et al. (2012) noted that "the present results do not constitute an argument against the existence of criterion variability" but do "constitute an argument against the claim that criterion noise (as currently modeled) has a major influence on recognition memory performance" (p. 475). The overwhelming majority of the data reviewed by Benjamin et al. (2009), only a small portion of which are cited here, come from experiments on recognition memory. To these results, Kellen et al. offered few alternatives and addressed only one domain specifically. With respect to sequential dependencies, they pointed out that sequential fluctuations in the memory representation, rather than the criterion, could account for such results (see also Malmberg & Annis, 2012). But how could such a mechanism explain the results, for example, of Van Zandt (2000), who found a difference between rating-scale and response-bias procedures for measuring ROC functions? Perhaps criterion noise is not the common thread that can tie these disparate results together, but Kellen et al. provided no alternative and no expla-

nation for the many mysteries that led to a search for criterion noise in the first place.

One final result is worth noting. The theory of Benjamin et al. (2009) makes the claim that each criterion introduces noise to the recognition process and, thus, suggests that rating scales of greater length should yield poorer estimates of recognition performance than shorter scales. We have recently reported this result (Benjamin, et al., in press)—the isosensitivity function for responses made on an 8-option scale lies below the one for responses made on a 4-options scale, which lies below the single point yielded by yes–no recognition (a 2-option rating scale). This result seems difficult to reconcile with any theoretical view in which individual criteria do not add noise to the recognition decision.

Summary

The argument by Kellen et al. (2012) that criterion noise is not present in recognition hinges upon two claims. First, although they replicated the result of Benjamin et al. (2009) that the best fitting model of ensemble recognition was one that included a role for criterion noise, they argued that that model should be rejected because it implemented a flawed assumption regarding criterion shifts across ensembles, and because it yielded anomalous fits to some individual subjects. I have argued here that the statistical evidence underlying the decision to reject the criterion-shift restriction is unconvincing, and there is no reason to take the reported fits as anomalous. Moreover, untraditional parameter values do not unequivocally indicate the invalidity of a model, but may instead reveal the difficulty of fitting that model with limited data and parameter mimicry within models that incorporate multiple sources of noise.

Second, the comparison of forced-choice ranking and yes–no rating tasks relies critically upon the claim that ranking is criterion free but rating is not. The extant literature comparing these procedures has not supported this claim convincingly, and the interleaved test trials in Kellen et al.'s (2012) experiment make it more vulnerable to such violations than most.

Finally, the search for criterion noise was motivated not by an abstract concept but rather by the mass of results that have accumulated that suggest its omnipresence in recognition. Taken together, I think that the results lean very strongly in favor of the contribution of criterion noise to recognition. However, the points of contention between Benjamin et al. (2009) and Kellen et al. (2012) do suggest some new avenues for empirical reconciliation.

To start with, the prediction of the criterion-shift restriction that high-confidence response should increase in frequency with ensemble size is provably wrong in our data. An alternative would be to restrict the likelihood ratios rather than the location of criteria (cf. Hirshman, 1995; Stretch & Wixted, 1998) across ensemble conditions. Such a restriction would preserve the spirit of the idea that recognizers are reluctant to shift criteria and still avoid the flexibility associated with allowing the locations to vary freely across conditions.

With respect to the experimental issues, the concerns expressed here over the interleaved test trials in both experiments are quite easily addressed with experiments that reduce the potential for carryover effects. The equivalence between forced-choice and yes–no procedures should survive a generalization to a between-list manipulation of task, as should the results from the ensemble recognition task.

Considering the sources of noise that influence memory judgments is an important project that has both theoretical (Wixted & Mickes, 2010) and applied (Benjamin, 2010) applications. We are thankful to Kellen et al. (2012) for their sophisticated work on the problem and look forward to the new data that this ongoing debate will stimulate.

References

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99. doi:10.1037/h0029531

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 941–947. doi:10.1037/0278-7393.27.4.941

Benjamin, A. S. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about purportedly nonparametric measures. *Memory & Cognition, 33,* 261–269. doi:10.3758/BF03195315

Benjamin, A. S. (2010). Representational explanations of "process" dissociations in recognition: The DRYAD theory of aging and memory judgments. *Psychological Review, 117,* 1055–1079. doi:10.1037/a0020810

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

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (in press). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Bonnel, A.-M., & Miller, J. (1994). Attentional effects on concurrent psychophysical discriminations: Investigations of a sample-size model. *Perception & Psychophysics, 55,* 162–179. doi:10.3758/BF03211664

Busey, T. A., & Arici, A. (2009). On the role of individual items in recognition memory and metacognition: Challenges for signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1123–1136. doi:10.1037/a0016646

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

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology, 56,* 196–207. doi:10.1016/j.jmp.2012.02.004

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America, 46,* 372–383. doi:10.1121/1.1911699

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note, No.* 58-51(3), 32. doi:10.1117/1.1646178

Ennis, D. M., & O'Mahony, M. (1995). Probabilistic models for sequential taste effects in triadic choice. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1088–1097. doi:10.1037/0096-1523.21.5.1088

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

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Ex-*

perimental Psychology: Learning, Memory, and Cognition, 25, 500–513. doi:10.1037/0278-7393.25.2.500

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

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

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

Gronlund, S. D. (2004). Sequential lineups: Shift in criterion or decision strategy? *Journal of Applied Psychology, 89,* 362–368. doi:10.1037/0021-9010.89.2.362

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6,* 544–553. doi:10.3758/BF03198243

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313. doi:10.1037/0278-7393.21.2.302

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.

Kapucu, A., Rotello, C. M., Ready, R. E., & Seidl, K. N. (2008). Response bias in "remembering" emotional stimuli: A new perspective on age differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 703–711. doi:10.1037/0278-7393.34.3.703

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

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

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

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

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

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

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

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53,* 49–70. doi:10.3758/BF03211715

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141,* 233–259. doi:10.1037/a0025277

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

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

Nisbett, R. E., & Wilson, T. C. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. doi:10.1037/0033-295X.84.3.231

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

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science, 9,* 247–255. doi:10.1111/1467-9280.00051

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

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

Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 615–625. doi:10.1037/0278-7393.30.3.615

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1397–1410. doi:10.1037/0278-7393.24.6.1397

Stüttgen, M. C., Yildiz, A., & Güntürkün, O. (2011). Adaptive criterion setting in perceptual decision making. *Journal of the Experimental Analysis of Behavior, 96,* 155–176. doi:10.1901/jeab.2011.96-155

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

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. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254–262. doi:10.3758/BF03193446

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5,* 102–122. doi:10.1016/0022-2496(68)90059-X

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., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054. doi:10.1037/a0020874

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

Received April 16, 2012
Revision received January 9, 2013
Accepted January 11, 2013 ■