Psychological Review
2013, Vol. 120, No. 3, 727–730

0033-295X/13/$12.00 DOI: 10.1037/a0033141

REPLY

On the Measurement of Criterion Noise in Signal Detection Theory: Reply to Benjamin (2013)

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

Kellen, Klauer, and Singmann (2012) questioned whether possible criterion noise would contribute significantly to modeling recognition memory. Our arguments were based on a reanalysis of the data by Benjamin, Diaz, and Wee (2009) as well as on new experimental data. In a comment, Benjamin (2013) questioned some of Kellen et al.'s conclusions and raised important issues regarding the new experimental data. In this reply, we revisit our arguments and provide new analyses in response to Benjamin's questions and issues.

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

In order to measure criterion noise, Benjamin, Diaz, and Wee (2009) used the ensemble recognition task. By assuming certain information-integration rules (e.g., summation, averaging), criterion noise in the signal detection theory (SDT) model (Wickens, 2002) becomes identifiable.[1] In addition to the measurement of criterion noise, Benjamin et al. also evaluated whether or not (mean) response criteria could be restricted to be equal across ensemble size. Model-selection analyses preferred the averaging model with the response-criteria restriction and criterion noise, a result that led Benjamin et al. to conclude that criterion noise plays a major role in recognition judgments.

According to Benjamin (2013), our criticisms of Benjamin et al. (2009) critically hinge on the fact that the parameter estimates of the winning model take on implausible values, and that the above-mentioned response-criteria restriction has to be rejected in a test of statistical significance for Benjamin et al.'s data. In fact, our critique was somewhat more elaborate. We argued that the response-criteria restriction produces a series of problems at the level of qualitative predictions, and potentially distorting parameter estimates (e.g., inflating criterion noise estimates). Given these issues, we argued that this ultimately unnecessary restriction should be avoided. Furthermore, we pointed out that Benjamin et al.'s evidence for criterion noise was *only* found when imposing the problematic response-criteria restriction. The arguments comprising our criticism have been presented in detail by Kellen, Klauer, and Singmann (2012), so we do not address them further here. Instead, we focus on two particular issues that deserve further discussion.

Statistical Properties of the Models

A simulation analysis reported by Kellen et al. (2012) showed that the ensemble recognition task and the associated models have extremely low power for detecting the presence of criterion noise. When generating data with the (unrestricted) models (using Benjamin et al.'s, 2009, sample sizes) and introducing different levels of criterion noise, the detection of criterion noise using null-hypothesis testing in individual data sets never exceeded 8%.

A similar simulation can be done while imposing the response-criteria restrictions. The purpose of this simulation is to assess the impact of the response-criteria restriction on the probability of *falsely* rejecting the hypothesis that there is no criterion noise (i.e., $\sigma_c = 0$) when *criterion noise is in fact absent.* One-thousand artificial data sets were generated using the averaging model with no criterion noise ($\sigma_c = 0$) and without response-criteria restrictions, using the generating parameter values already employed by Kellen et al. (2012). The generated data sets were then fitted with the averaging model with (mean) response criteria restricted to be equal across ensembles. The *true* restriction $\sigma_c = 0$ was tested for each fitted data set via null-hypothesis testing and was falsely rejected in 98% of the cases. Similar rates of false positive results were obtained with the summation model (rejection in 92% of the cases).[2]

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. R scripts can be obtained from David Kellen upon request.

Correspondence concerning this article should be addressed to David Kellen, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Engelbergerstr. 41, D-79085 Freiburg, Germany. E-mail: david.kellen@psychologie.uni-freiburg.de

[1] Let μ_t and σ_t denote the mean and standard deviation of the evidence distribution for targets (studied items), respectively. Without loss of generality, the mean (μ_d) and standard deviation (σ_d) of the evidence distribution for distractors (new items) are restricted to 0 and 1. Parameter σ_c denotes criterion noise.

[2] Note that because $\sigma_c = 0$ is at the boundary of the parameter space, the sampling distribution of the likelihood-ratio test follows a $\bar{\chi}^2$ distribution,

Taken together, as shown by Kellen et al. (2012), when not imposing response-criteria restrictions, the restriction $\sigma_c = 0$ is *seldom rejected*, even when high values of criterion noise are in fact *present.* On the other hand, the new simulation results show that when imposing response-criteria restrictions (as Benjamin et al., 2009, did), the restriction $\sigma_c = 0$ is *almost invariably rejected* despite that criterion noise is in fact *absent.* The probability of rejecting the criterion-noise restriction is barely affected by the actual presence of criterion noise, and instead is critically dependent on problematic ancillary restrictions. This indicates that very little evidence regarding the presence or absence of criterion noise can be obtained from Benjamin et al.'s (2009) modeling of the ensemble task.

Parameter Plausibility

Concerning the extreme parameter estimates obtained, Benjamin (2013) pointed out that SDT parameters such as μ_t, σ_t, and σ_d are defined only up to a multiplicative factor that is conventionally determined by setting σ_d, the standard deviation of the evidence distribution for distractors, equal to one a priori. Extreme values of σ_c or μ_t are thus extreme relative to a standard deviation of $\sigma_d = 1$. It is of course possible to fix, say, μ_t to a reasonable value and to re-scale the previously fixed σ_d accordingly. In consequence, instead of μ_t assuming extremely large values, σ_d will now simply assume extremely small values (see Figure 1, Panel D, in Benjamin, 2013).

Values such as those obtained for Benjamin's (2013) data and model are extreme inasmuch as they are not consistent with the predictions of more fine-grained computational memory models that describe encoding, storage, and retrieval processes (e.g., Shiffrin & Steyvers, 1997). Moreover, the extreme parameter estimates are also not consistent with dynamic versions of SDT such as the diffusion model (Ratcliff, 1978), according to which drift rate variabilities and drift rates themselves are found to be roughly similar in size to the variabilities and means of signal and noise distributions in SDT (e.g., Starns, Ratcliff, & McKoon, 2012, p. 16).

Moreover, even when the "real" parameter values are quite similar to what is usually found in the literature (see Kellen et al., 2012, p. 464), imposing the problematic response-criteria restriction results in artifactually extreme estimates of the parameter values as corroborated in the new simulation reported above: The data-generating parameters for the averaging model were $\mu_t = 0.62$, $\sigma_t = 1.32$, and $\sigma_c = 0$; but for 65% of the artificial data sets, the estimates of μ_t, σ_t, and σ_c were all larger than 10. Benjamin (2013) acknowledged that the response-criteria restriction leads to gross mispredictions but overlooked the impact of these mispredictions in terms of parameter estimates when attempting to justify extreme parameter values.

Criticisms to Kellen et al. (2012)

Benjamin (2013) raised important questions regarding our approach to measure criterion noise, in particular regarding the tacit assumption that forced-choice tasks such as the ranking task are unbiased (see Klein, 2001). Unfortunately, we did not record the spatial position of the old and new items in the ranking task, so we cannot test the unbiasedness assumption for our data. We can nevertheless evaluate the questions using alternative sources of evidence. First, there is no evidence in the memory literature indicating that there is a spatial bias in forced-choice tasks (Kroll, Yonelinas, Dobbins, & Frederick, 2002) or a systematic misprediction of forced-choice data, as would be expected if there was a spatial bias (Smith & Duncan, 2004). For further checking, we fitted the 63 individual data sets reported by Jang, Wixted, and Huber (2009) and tested the unbiasedness assumption (by restricting the confidence criterion determining the binary judgment in the two-alternative forced choice [2AFC] receiver operating characteristics [ROC] functions to 0). In these data sets, 2AFC and YES–NO confidence-rating trials were intermixed, a case which according to Benjamin is likely to bias 2AFC judgments. The unbiased SDT model was preferred over the unrestricted model in 86% individual data sets in terms null-hypothesis testing, 76% in terms of Akaike information criterion $(\text{AIC})_c$, and 95% in terms of Bayesian information criterion (BIC).[3] This preference is not surprising given that the binary-response criterion was on average unbiased (*Mdn* = 0.04, *SD* = 0.35), $t(62) = 0.94$, $p = .35$, with the 2AFC criterion restriction ($c = 0$) producing marginal increases in misfit, *Mdn* $\Delta G^2(1) = 0.80$, $p = .37$.

Another issue raised by Benjamin (2013) is that the correlation of σ_t estimates across tasks was only .20 for our data, which, he suggested, could indicate violations of assumptions. The observed correlations of μ_t and σ_t across tasks were .81 and .20, respectively. Does the low correlation for σ_t necessarily imply a violation of the assumptions? Not necessarily, as the size of the correlation can be reduced due to the low precision of σ_t estimates (Macmillan, Rotello, & Miller, 2004), a restricted range of the observed σ_t estimates, as well as model overfitting (Jang et al., 2009). In order to check this, we assessed the sampling distributions of the correlation coefficients using a parametric-bootstrap simulation (see Efron & Tibshirani, 1993, Chapter 6) in which we generated data from a model in which both μ_t and σ_t are equal across tasks, and computed parameter correlations when estimating μ_t and σ_t separately for each task.[4] Simulation results show that the 95% bootstrap confidence intervals for the correlations of μ_t and σ_t across tasks were [0.70, 0.92] and [0.10, 0.68], respectively, which means that both observed correlations are within their respective 95% confidence intervals, and that a lower precision in σ_t esti-

with $\bar{\chi}^2 \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Self & Liang, 1987). If Akaike information criterion $(\text{AIC})_c$ and Bayesian information criterion (BIC) were used instead of null-hypothesis testing, the averaging model with criterion noise would have been preferred in 98% and 93% of cases, respectively. In the case of summation model, the model with criterion noise would have been preferred in 94% and 79% of cases, respectively.

[3] We thank Yoonhee Jang for providing the individual data sets. Model fits were done using R (R Development Core Team, 2012) and package MPTinR (Singmann & Kellen, 2013). Scripts can be obtained from the first author upon request.

[4] The bootstrap simulation consists of three steps: (1) One artificial data set is generated for each of the 30 sets of individual parameter estimates obtained from Kellen et al.'s (2012) experiment, while assuming that μ_t and σ_t are equal across tasks and that $\sigma_c = 0$. Each of these artificial data sets exactly follows the experimental design of Kellen et al. (2) Parameters μ_t and σ_t are estimated separately for ranking and confidence-rating trials from these artificial data sets, and their correlation computed. (3) Steps 1 and 2 are repeated many times (in the present case 5,000 times) in order to obtain stable distributions of correlation values.

mates is expected even when memory parameters are in fact the same across tasks.

In the light of these results, the notion that unaccounted response biases severely distorted our results seems somewhat implausible, although we acknowledge that further work focusing on direct tests is desirable.

The Explanatory and Predictive Power of Criterion Noise

Benjamin (2013) furthermore repeated Benjamin et al.'s (2009) point that a number of important empirical phenomena reported in the literature could in principle be accounted for by criterion noise, and he criticized us for not discussing these phenomena. In fact, most of these phenomena have alternative accounts that are at least as plausible as criterion noise. Let us briefly illustrate this by means of four of Benjamin's examples.

Sequential Dependencies and ROC Distortions

Benjamin (2013) himself acknowledged that sequential dependencies (Malmberg & Annis, 2012) can be explained by models assuming fluctuations in the memory representations without criterion noise (e.g., Atkinson, 1963). However, he holds that such models do not provide any account for the observed ROC distortions (Balakrishnan, 1999; Van Zandt, 2000). This statement is inaccurate: For example, representational accounts were already proposed by Balakrishnan (1999, p. 1201) in his report of critical results implying ROC distortions (the same kind of distortions reported by Van Zandt, 2000; see Mueller & Weidemann, 2008). More recently, Turner, Van Zandt, and Brown (2011) proposed a dynamical SDT model that can account for many problematic results and supposed violations of SDT assumptions (e.g., Van Zandt, 2000). This SDT model assumes that the evidence distributions change across test trials, without the traditional notion of response criteria that are placed along the evidence axis.

Response Conservatism

Response conservatism concerns the observation that individual's shifts in response bias are smaller than optimal. As shown by Maloney and Thomas (1991), response conservatism can occur simply because the "true" latent evidence distributions do not correspond to the assumed Gaussian distributions. In fact, distributional misspecification can lead to results suggesting response conservatism even when individuals optimally adjust response criteria. This is especially problematic as it is well known that one cannot directly single-out a particular distributional assumption as the "correct one" (Krantz, Luce, Suppes, & Tversky, 1971, Chapter 2; Rouder, Pratte, & Morey, 2010; but see Wixted & Mickes, 2010b), which means that there is no direct solution for this conundrum.

Remember–Know ROCs

Criterion noise has been used to account for joint Remember–Know and confidence judgments (e.g., Wixted & Stretch, 2004). A critical aspect of these SDT accounts is that they are unidimensional, meaning that they only assume a single evidence axis. Wixted and Mickes (2010a) recently proposed a two-dimensional SDT model that provides an equally good account without criterion noise. Furthermore, Wixted and Mickes tested and confirmed a set of hypotheses emerging from the model, giving further credence to this two-dimensional account. In consequence, the need for assuming criterion noise may arise out of erroneous assumptions about the dimensionality of the underlying evidence space.

Inverse Relationship Between Confidence-Scale Size and Performance

Benjamin et al. (2009) established the prediction that criterion noise assumes a greater magnitude in cases where several response criteria have to be simultaneously established. Given the impact of criterion noise on performance, it is predicted that ROC points obtained with simple binary responses are above ROC points obtained with confidence-rating scales.[5] Benjamin, Tullis, and Lee (2013) reported interesting findings confirming this prediction, although the observed differences in performance were relatively small. The notion that the effect of scale size in recognition-memory performance is small is corroborated by recent work (e.g., Mickes, Hwe, Wais, & Wixted, 2011; Mickes, Wixted, & Wais, 2007) where 20- and 99-point confidence-rating scales were used, leading to parameter estimates that are indistinct from the ones usually reported in the literature with 6-point scales (e.g., Ratcliff, McKoon, & Tindall, 1994). Although Benjamin et al.'s (2013) results are consistent with criterion noise, they do not suggest that criterion noise as such plays an important role in recognition judgments, and they are thus consistent with our conclusion that the contribution of criterion noise for the modeling of recognition memory is typically very small.

Conclusions

Taken together, we see little reason to modify our claim: Criterion noise (as currently modeled) does not have a major influence on recognition-memory performance. In any case, the study of criterion noise in SDT is a difficult problem that can be approached in many different ways (Klauer & Kellen, 2012), and new interesting results such as the ones reported by Benjamin et al. (2013) show that there is much to discover and understand. Like Benjamin, we look forward to the new data that this ongoing debate will stimulate.

[5] It should be mentioned that this prediction is not only made by SDT models with criterion noise. Actually, it has long been shown to be a prediction of certain discrete-state models (see Krantz, 1969, p. 322).

References

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

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

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

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. (2013, February 18). 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*. Advance online publication. doi:10.1037/a0031849

Efron, B., & Tibshirani, R. (1993). *Introduction to the bootstrap*. London, England: Chapman and Hall.

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

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

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76,* 308–324. doi:10.1037/h0027238

Krantz, D. H., Luce, R. D., Suppes, P., & Tverksy, A. (1971). *Foundations of measurement* (Vol. I). New York, NY: Academic Press.

Kroll, N. E., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes–no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131,* 241–254. doi:10.1037/0096-3445.131.2.241

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

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

Maloney, L. T., & Thomas, E. A. C. (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology, 35,* 443–470. doi:10.1016/0022-2496(91)90043-S

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

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865. doi:10.3758/BF03194112

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

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

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

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:10.1080/01621459.1987.10478472

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

Singmann, H., & Kellen, D. (2013, January 24). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-012-0259-0

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

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64,* 1–34. doi:10.1016/j.cogpsych.2011.10.002

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118,* 583–613. doi:10.1037/a0025191

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

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

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

Wixted, J. T., & Mickes, L. (2010b). Useful scientific theories are useful: A reply to Rouder, Pratte, and Morey (2010). *Psychonomic Bulletin & Review, 17,* 436–442. doi:10.3758/PBR.17.3.436

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 January 29, 2013
Revision received March 18, 2013
Accepted March 20, 2013 ■