Psychonomic Bulletin & Review
2010, 17 (3), 436-442
doi:10.3758/PBR.17.3.436

Useful scientific theories are useful: A reply to Rouder, Pratte, and Morey (2010)

John T. Wixted and Laura Mickes
University of California, San Diego, La Jolla, California

In a recognition memory experiment, Mickes, Wixted, and Wais (2007) reported that distributional statistics computed from ratings made using a 20-point confidence scale (which showed that the standard deviation of the ratings made to lures was approximately 0.80 times that of the targets) essentially matched the distributional statistics estimated indirectly by fitting a Gaussian signal-detection model to the receiver-operating characteristic (ROC). We argued that the parallel results serve to increase confidence in the Gaussian unequal-variance model of recognition memory. Rouder, Pratte, and Morey (2010) argue that the results are instead uninformative. In their view, parametric models of latent memory strength are not empirically distinguishable. As such, they argue, our conclusions are arbitrary, and parametric ROC analysis should be abandoned. In an attempt to demonstrate the inherent untestability of parametric models, they describe a non-Gaussian equal-variance *model that purportedly accounts for our findings just as well as the Gaussian unequal-variance model does. However, we show that their new model—despite being contrived after the fact and in full view of the to-be-explained data—does not account for the results as well as the unequal-variance Gaussian model does. This outcome manifestly demonstrates that parametric models are, in fact, testable. Moreover, the results differentially favor the Gaussian account over the probit model and over several other reasonable distributional forms (such as the Weibull and the lognormal).*

Receiver-operating characteristic (ROC) data obtained from recognition memory experiments generally follow an asymmetrical curvilinear path when plotted in probability space and follow an approximately linear path with a slope of less than 1 when plotted in *z* space. These findings are interpreted by the Gaussian signal-detection model to mean that the memory strengths of the target items are more variable than the memory strengths of the lures. More specifically, if the Gaussian model is correct, then the slope of the linear *z*-ROC provides an estimate of the ratio of the standard deviation of the lure distribution (σ_{Lure}) divided by the standard deviation of the target distribution (σ_{Target}). Typically, the slope of the *z*-ROC is close to 0.80 (Ratcliff, Sheu, & Gronlund, 1992). Thus, the standard deviation of the lure distribution is estimated by the Gaussian model to be approximately 0.8 times that of the target distribution (i.e., $\sigma_{Lure} / \sigma_{Target} \approx 0.8$). The idea that the memory strengths of targets and lures are normally distributed, with the mean and variance of the target distribution exceeding the mean and variance of the lure distribution, is known as the *unequal-variance signal-detection model*. This model has guided thinking about recognition memory for over 50 years (beginning with Egan, 1958).

Mickes, Wixted, and Wais (2007) asked a simple question: Would the same result—namely, a higher mean and variance of the memory strengths for the targets as compared with the lures—be evident if one used a 20-point confidence scale and then simply computed the relevant distributional statistics from the ratings themselves instead of estimating them by fitting a Gaussian model to ROC data? And if an unequal-variance model were suggested by the ratings data, would the magnitude of the estimated ratio of the standard deviations based on the ratings (s_{Lure} / s_{Target}) be similar to the magnitude of the estimated ratio obtained by fitting a Gaussian model to ROC data ($\sigma_{Lure} / \sigma_{Target}$)?

A priori, agreement between the two ratio estimates seems unlikely, because there are many reasons why they might disagree. For example, if the Gaussian assumption is not valid, then disagreement between the two estimates seems more likely than agreement. In addition, if the rating scale does not approximate an interval scale, or if it covers only a limited range of the memory strength dimension, then, again, disagreement seems more likely than agreement. Somewhat surprisingly, Mickes et al. (2007) found that the two estimates showed good agreement, even at the level of the individual participant. Both methods suggested that the memory strengths of the targets were more variable than the memory strengths of the lures, and both further suggested that the standard deviation ratio was, on average, approximately 0.80. The group data from their Experiment 1 illustrate the basic finding. The slope of the group *z*-ROC, which provides an estimate of $\sigma_{Lure} / \sigma_{Target}$ if the Gaussian model is correct, was 0.833. The corresponding standard deviation ratio computed directly from the ratings (s_{Lure} / s_{Target}) was 4.14 divided by 5.01, or 0.826. The computation of this standard deviation ratio does not involve any distributional assumptions, yet it was nearly identical to the ratio estimate provided by the Gaussian interpretation of the ROC.

From these results, Mickes et al. (2007) drew two basic conclusions:

1. The two experiments reported here support a conclusion that is commonly drawn from ROC analysis—namely, that the memory strengths of the targets are more variable than the memory strengths of the lures. (p. 864)
2. The close agreement between the model-based ROC analysis and the model-free ratings method supports not only an unequal-variance model, but also the idea that the memory strengths are distributed in such a way that fitting a specifically Gaussian model to the data yields accurate conclusions (even if the true underlying distributions are not strictly Gaussian). (p. 864)

J. T. Wixted, jwixted@ucsd.edu

Rouder, Pratte, and Morey (2010) argue that these results are instead uninformative because parametric models of latent memory strength are simply not testable. More specifically, although the results are compatible with the Gaussian model and with the idea that the variance of the target distribution is greater than the variance of the lure distribution, they are also compatible with different parametric models that yield different conclusions about the relative variances of the two distributions. Thus, in their view, the scientific method cannot shed any light on the parametric properties of latent memory strength, so parametric ROC analysis should be abandoned in favor of a nonparametric approach.

In making their case, Rouder et al. (2010) overlook a key question: Why did the Gaussian ROC estimate of $\sigma_{\text{Lure}} / \sigma_{\text{Target}}$ and the direct rating estimate of $s_{\text{Lure}} / s_{\text{Target}}$ agree when there are so many reasons why they might have disagreed? As we show next, signal-detection models based on a variety of common distributional forms differ in the degree to which the two estimates agree. The fact that the agreement between these two estimates is higher for the Gaussian model than for other reasonable distributional forms shows that (1) parametric models are testable and (2) the empirical results differentially support the Gaussian account.

The Gaussian model not only outperforms a variety of non-Gaussian models that have been mentioned in connection with ROC analysis in the past, it also outperforms the new equal-variance signal-detection model that Rouder et al. (2010) contrived in an effort to demonstrate that parametric models of latent memory strength are not testable. That is, despite its being contrived after the fact and in full view of the to-be-explained data, we show that their new equal-variance model does not exhibit the degree of correspondence between ROC analysis and ratings data that the Gaussian model exhibits (a model that was not contrived after the fact). Thus, Rouder et al. have not provided a mathematical proof that an alternative equal-variance non-Gaussian model can account for the data as well as an unequal-variance Gaussian model can. Instead, they have introduced a new signal-detection model—one that, like any new model, will stand or fall on the basis of its ability to parsimoniously account for the empirical data. The fact that their new model is empirically outperformed by the Gaussian unequal-variance signal-detection model further underscores the point that parametric models are, indeed, testable, and it provides additional evidence in favor of, not evidence against, the longstanding Gaussian account.

An Unlikely Coincidence

Before the fact, the Gaussian model predicted the observed correspondence between ROC analysis and ratings data. The likelihood of that correspondence if the Gaussian model is wrong would seem to be fairly low, and the analyses that we present below reinforce that suspicion. A basic tenet of scientific reasoning is that when a low-probability prediction survives empirical scrutiny, confidence in the theory that made that prediction should increase.

Green and Swets (1966) noted long ago that although a Gaussian signal-detection model often adequately characterizes ROC data, any monotonic transformation of the decision axis variable predicts the same observed data (e.g., ROC data that correspond to a likelihood ratio model also correspond to a log likelihood ratio model). To take another example used by Rouder et al. (2010), exponentiating Gaussian random variables yields a highly skewed distribution described by the lognormal. Despite its dramatically non-Gaussian form, the lognormal signal-detection model fits ROC data *exactly* as well as the Gaussian model. Accordingly, the mere fact that the Gaussian model accurately describes ROC data cannot be taken to mean that the underlying memory strengths are Gaussian in form or that ratio statistics based on the Gaussian assumption, such as the estimate of $\sigma_{\text{Lure}} / \sigma_{\text{Target}}$, are necessarily valid (Egan, 1975). Similarly, as pointed out by others, distributions that are not even related to the Gaussian can often accurately describe ROC data, and such distributions invariably yield different estimates of the lure-to-target standard deviation ratio than the estimate provided by the Gaussian model. Lockhart and Murdock (1970) went so far as to suggest that virtually any unimodal distribution can yield ROC data that are largely indistinguishable from what the Gaussian model predicts.

Although these points have long been known, Rouder et al. (2010) reiterate them in their comment, and they go to considerable lengths to show that they are true. Their efforts serve as a useful reminder of the limitations of ROC analysis, but they do not bear on the analysis that we performed, which was concerned with the *congruence* between ROC analysis and direct ratings (not with ROC analysis per se). Would any reasonable signal-detection model yield the same degree of congruence that was found for the Gaussian unequal-variance signal-detection model? That is a key question, yet Rouder et al. do not consider it.

The most straightforward way to answer this question would be simply to fit various plausible signal-detection models to the ROC data and then separately ask how well those models describe the same ROC data when their parameters are constrained to equal the means and standard deviations computed directly from the ratings. We performed just such an analysis for five different distributional forms (the Gaussian, the Weibull, the logistic, the lognormal, and the exponential). These distributions are shown in the left column of Figure 1, and all of them have been mentioned in connection with ROC analysis over the years. It is visually apparent that two of these distributions are Gaussian-like (the logistic and the Weibull), whereas the other two are very non-Gaussian (the lognormal and the exponential).

In the middle and right columns of Figure 1, we show the results of two different analyses that serve to test these models. The middle column shows the least squares group ROC fits for each of the five models. The ROC data are from Experiment 1 of Mickes et al. (2007), and they were obtained using a 20-point confidence rating scale ranging from 1 (*sure new*) to 20 (*sure old*), which yields a 19-point ROC. For each model, the relevant parameters were iteratively adjusted to achieve the best fit (as is typically done in ROC analysis). The parameter of most interest here is $\sigma_{\text{Lure}} / \sigma_{\text{Target}}$, and its estimated value for each model is

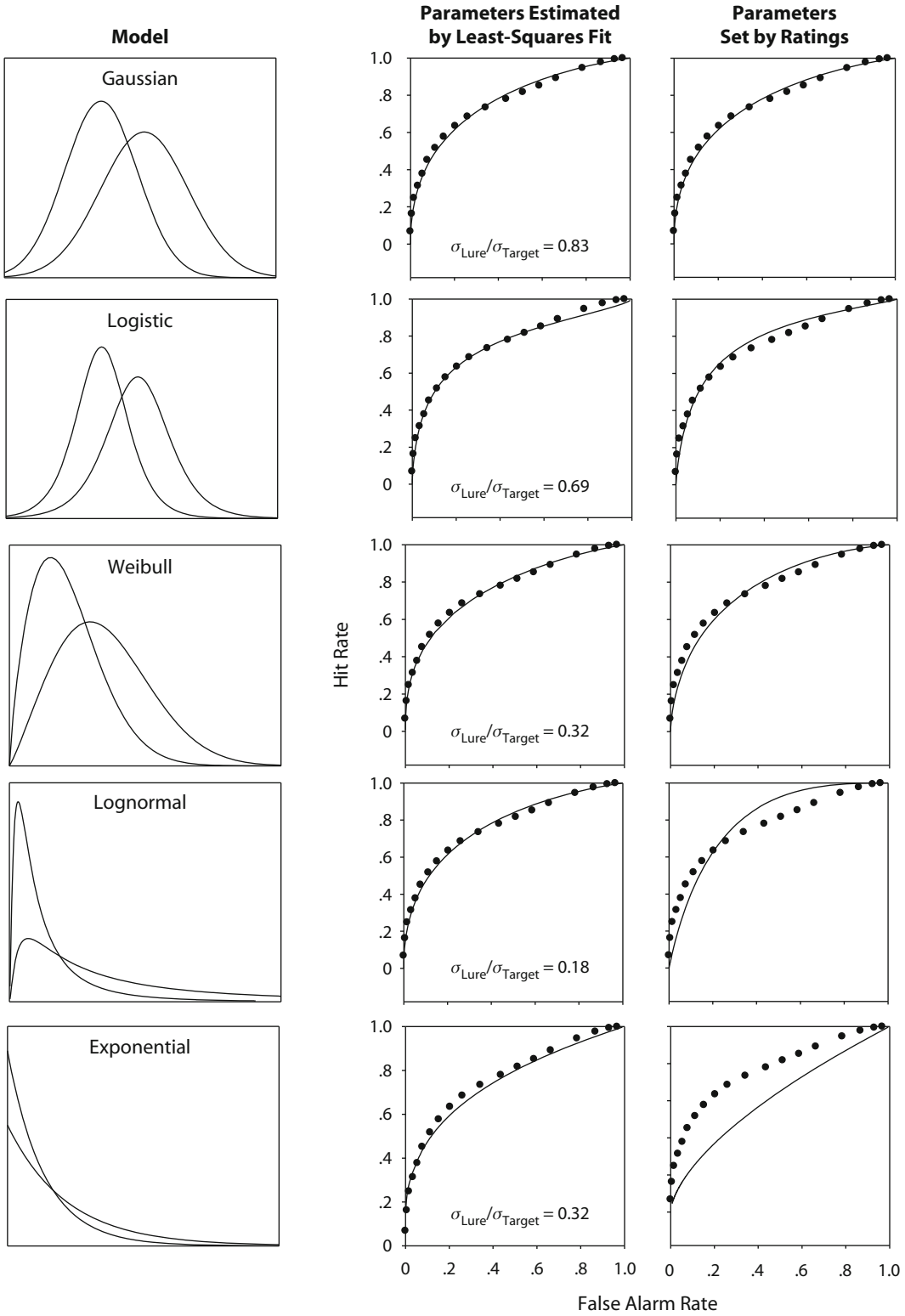

Figure 1. Left panels: An illustration of five signal-detection models based on different distributional forms. Middle panels: Least-squares fit (solid curve) of each signal-detection model to the 19-point group receiver-operating characteristic (ROC) data reported by Mickes, Wixted, and Wais (2007). Also shown is the estimate of $\sigma_{Lure} / \sigma_{Target}$ provided by each fitted model. Right panels: Predicted ROC (solid curve) from each signal-detection model with its parameters constrained to yield the mean and standard deviations computed directly from the ratings. That is, each model was constrained to yield $d = 1.19$ (the distance between the means of the target and lure distributions in lure standard deviations, according to the ratings) and $\sigma_{Lure} / \sigma_{Target} = 0.83$ (because, in the ratings, $s_{Lure} / s_{Target} = 0.83$). The data in the right panels are the same as the data in the middle panels.

shown in the figure. Obviously, despite their differences, all five models are capable of fitting these ROC data reasonably well. Moreover, the models provide widely varying estimates of $\sigma_{Lure} / \sigma_{Target}$, ranging from a high of 0.83 for the Gaussian model to a low of 0.18 for the lognormal model. Thus, this figure illustrates the well-known fact that ROC data do not uniquely support either the Gaussian interpretation or the idea that the standard deviation of the lure distribution is 0.80 times that of the target distribution.

The right column of Figure 1 shows the results of a different analysis. For this analysis, the parameters of each model were constrained to equal the corresponding values computed directly from the ratings (instead of being estimated by fitting the model to the ROC data). According to the ratings, $s_{Lure} / s_{Target} = 0.83$. Thus, the $\sigma_{Lure} / \sigma_{Target}$ parameter for each model was set to 0.83. In addition, according to the ratings, the mean of the target distribution was 1.19 standard deviations greater than the mean of the lure distribution, so the means of each model were also constrained to have that same relationship.[1] For each model, a comparison of the two plots (middle column vs. right column) illustrates the *congruence* between distributional estimates obtained from ROC analysis (which involves fitting the parametric model to the data) and distributional estimates computed directly from the ratings.

It is visually apparent that when the distributional statistics for each model are computed directly from the ratings, the performance of the Gaussian model is scarcely affected, whereas the other models now exhibit systematic deviations (deviations that are especially large for the lognormal and the exponential models). The same point can be made in a different way by noting that $\sigma_{Lure} / \sigma_{Target}$ estimated from fitting each model to the ROC data agrees with the computation of s_{Lure} / s_{Target} for the Gaussian model but not for the other models. Thus, agreement between ROC analysis and direct ratings is obviously not preordained, and, a priori, such agreement does not seem especially likely. These findings should increase confidence in the interpretation provided by the unequal-variance Gaussian model in comparison with the other four models shown in Figure 1.

The results shown in Figure 1 disconfirm the main argument made by Rouder et al. (2010), which is that parametric models of latent memory strength are not testable. Clearly, they are testable. In addition, the evidence weighs in favor of the unequal-variance Gaussian model. Indeed, this would appear to be the only evidence in the literature that specifically supports the Gaussian form over other distributional forms. Obviously, people may disagree about how compelling they find these results to be, but it does not seem reasonable to suggest that parametric models cannot be tested or that the results do not lend more support to the Gaussian model than to any of the other models that we considered.

Rouder et al. (2010) attempt to demonstrate that parametric models are not testable by comparing the performance of the unequal-variance Gaussian model with that of a newly contrived equal-variance non-Gaussian model (the probit model). As they put it:

> We demonstrate the arbitrariness of MWW's conclusion that studied-item latent strengths are more variable than new-item latent strengths by constructing two completely equivalent models of their data from Experiment 1. (p. 432)

According to them, the new equal-variance probit model can fit the asymmetrical ROC data and the frequency ratings data as well as the unequal-variance Gaussian model can. Thus, as they see it, neither the Gaussian model nor the probit model (nor any other parametric model of latent memory strength) is testable, because one can always find another model that fits equally well and that yields a different conclusion.

In making this claim, Rouder et al. (2010) appear to have mistakenly applied an argument that would have been valid had we restricted our analyses to fitting parametric models to confidence-based ROC data. Restricted to that domain of analysis, the unequal-variance Gaussian model and their new equal-variance probit model (not to mention all other models based on a monotone transform of the Gaussian variable) cannot be empirically distinguished, because they are mathematically constrained to provide equivalent fits. In that sense, these parametric models are not testable. However, we did not restrict our analysis to fitting parametric models to ROC data. Instead, the whole point of our article was to bring a *different* method of analysis to bear on the issue, one that involves measuring the congruence between parameter estimates based on ROC analysis and parameter estimates computed directly from the ratings (which involves no distributional assumptions). As we show next, the same method can be used to empirically distinguish the Gaussian model from the newly described probit model, thereby reinforcing the point that parametric models of latent memory strength are, in fact, testable.

The Unequal-Variance Gaussian Model Versus the Equal-Variance Probit Model

Rouder et al. (2010) worked out an equal-variance signal-detection model by passing random Gaussian variables (X) through a $\Phi(2X/3)$ filter, where Φ is the standard normal cumulative distribution function. This filtering process transformed the Gaussian unequal-variance model into the equal-variance signal-detection model (the probit model) shown in their Figure 5B. Because it involves a monotonic transformation of the Gaussian variable, this equal-variance model fits ROC data (including asymmetrical ROC data) exactly as well as does the unequal-variance Gaussian model. Next, they found that if the confidence criteria are arranged in a particular nonlinear fashion with respect to the memory strength axis (such that the criteria are relatively compressed on the high end of the scale as opposed to the low end of the scale), then this model can largely reproduce the ratings data reported by Mickes et al. (2007). Thus, both the ROC data (which suggest that the target distribution is more variable than the lure distribution when interpreted by a Gaussian model) and the ratings data (which yield standard deviations that also suggest that the target distribution is more variable than the lure distribution) can be largely reproduced by an equal-variance model. That being the case, Rouder et al. conclude that our data offer no support for unequal target

and lure variances, and they abandon hope that either ROC data or ratings data will ever shed any light on parametric models of latent memory strength.

As indicated earlier, when the parameters of the Gaussian model are set to equal the mean and standard deviation values computed directly from the ratings, the slope of the predicted *z*-ROC is 0.826, which is close to the observed value of 0.833. As a result, the ratings-based Gaussian model (i.e., the Gaussian model parameterized by distributional statistics computed directly from the ratings) fits the observed ROC data well. However, when the parameters of the probit model are set to equal the mean and standard deviation values associated with the ratings (i.e., when we performed the same test that we performed for the five models shown in Figure 1), it does not fit the ROC data as well, because it predicts a lower *z*-ROC slope of 0.751. Moreover, the underperformance of the probit model appears to be a consistent result. We recently collected data from two very similar ratings experiments as part of a different project, and we analyzed those data in the same way. When the parameters of the Gaussian and probit models were set to the mean and standard deviations computed directly from the ratings, the predicted *z*-ROC slopes were again closer to the observed *z*-ROC slopes for the Gaussian model in both experiments. Figure 2 shows the obtained *z*-ROC slopes for all three ratings experiments. Also shown are the predicted *z*-ROC slopes for both models. In all three experiments, the Gaussian model outperforms the probit model. Indeed, even with only three observations, the (very reliable) deviation between observed and predicted *z*-ROC slopes for the probit model is statistically significant [$t(2) = 7.50$, $p < .01$]. The small deviation between the observed and predicted *z*-ROC slopes for the Gaussian model does not approach significance.

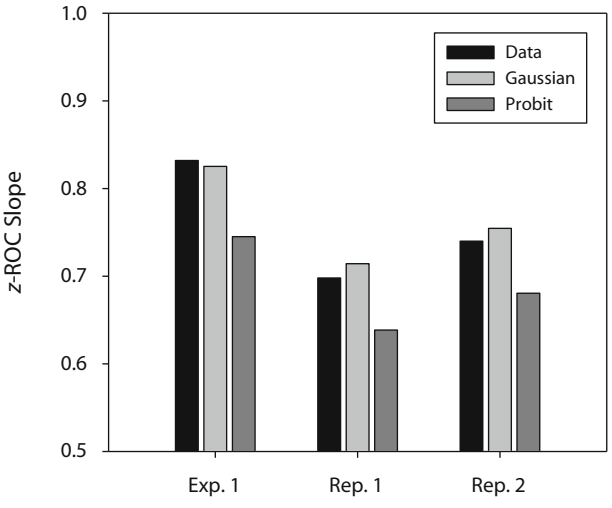

Figure 2. Obtained *z*-ROC slopes from Experiment 1 of Mickes, Wixted, and Wais (2007) and from two replications (Exp. 1, Rep. 1, and Rep. 2, respectively) and predicted *z*-ROC slopes from the Gaussian and probit signal-detection models after constraining both models to yield the parameters computed directly from the ratings.

The performance of the probit model could be improved by assuming a nonlinear mapping relationship between memory strength values and the ratings. In fact, this is exactly what Rouder et al. (2010) do in their efforts to show how well the probit model performed, but this is tantamount to adding additional free parameters to the model in order to rectify its lack of agreement with the observed data. Few would doubt that selectively adding free parameters to the probit model (or to any other model, for that matter) would improve its performance relative to the unequal-variance Gaussian model. The fact that models involving different numbers of free parameters can fit data equally well, which is what Rouder et al. essentially show, does not mean that the models are empirically indistinguishable. Instead, the ability of one model to more parsimoniously account for data than another model is a primary determinant of data-based model selection. Applying that rule here suggests that the probit model is less compelling than the Gaussian model. Thus, far from casting doubt on the unequal-variance Gaussian model, the efforts described by Rouder et al. lend further support to it.

The Limited Range and Interval Nature of the Confidence Scale

Our findings are consistent not only with the Gaussian signal-detection model but also with the idea that the ratings approximate an interval scale of measurement. That is, the tests that we performed suggest that the intervals between the confidence criteria are equal enough and the range of the confidence criteria on the memory strength axis is extensive enough to yield a useful estimate of the ratio of the standard deviation of the lure distribution to the standard deviation of the target distribution. This is not to suggest that the ratings provide a true equal-interval measurement scale. For example, scale biases, such as preferences for ratings of 5, 10, and 15, are often evident in ratings data. Also, although participants in our experiments were instructed not to overuse the endpoints of 1 and 20, they often did appear to overuse them (especially ratings of 20). However, because the confidence criteria are apparently arrayed in nonsystematic fashion on the memory strength axis across a wide enough range, the ratings can be used to compute relative distributional statistics that correspond to what would be obtained if the ratings provided a true equal-interval scale over the full range. If that were not the case, then the fact that distribution-free parameter estimates computed directly from the ratings closely correspond to Gaussian-based parameter estimates computed indirectly from parametric ROC analysis would be a surprising coincidence.

Rouder et al. (2010) argue that because the distribution of ratings to the target items is clustered at the high end of the scale and, to some extent, at the low end as well, the ratings cannot provide an interval scale. However, clustering at the endpoints shows that the range of the rating scale is limited, like an interval-scale fever thermometer that only ranges from 96ºF to 102ºF (and that therefore records every temperature less than 96º as 96º and every temperature greater than 102º as 102º). If the Gaussian model is correct, the ratings cover a limited range on the

memory strength dimension in a similar way. That limitation could have introduced disagreement between the Gaussian ROC fits and the ratings, but, in practice, this did not happen.

Intuitively, this might seem an odd result. The higher degree of clustering at the high end of the scale should cause the standard deviation estimate of the target distribution to be underestimated to a greater degree than the standard deviation estimate of the lure distribution. This, in turn, should cause the estimate of s_{Lure} / s_{Target} to be closer to 1.0 than the estimate of $\sigma_{Lure} / \sigma_{Target}$ estimated from fitting a Gaussian model to ROC data. However, a simulation analysis shows that this effect, although real, is small unless the clustering at the high end of the scale becomes more pronounced than it was in our experiment.

For this simulation, the mean of the lure distribution was set to 0 and its standard deviation was set to 1. The corresponding values for the target distribution were set to 1.2 and 1.25, respectively. Thus, the standard deviation of the lure distribution was 1 / 1.25 (or 0.80) times that of the target distribution. The 19 confidence criteria were placed beginning at 2 standard deviations below the mean of the lure distribution, with the distance between each criterion set to 0.28 standard deviations. Next, 10,000 strength values were randomly drawn from the lure distribution and another 10,000 strength values were randomly drawn from the target distribution, with each value assigned a rating of 1 to 20 depending on where the randomly selected strength value fell in relation to the 19 confidence criteria. These simulated ratings were then analyzed by fitting a Gaussian model to the *z*-ROC data (yielding an estimate of $\sigma_{Lure} / \sigma_{Target}$) and by computing s_{Lure} / s_{Target} directly from the ratings.

Figure 3 shows the resulting frequency distribution, which is clearly more truncated on the right than on the left (as was true of the real data). Nevertheless, the estimate of $\sigma_{Lure} / \sigma_{Target}$ obtained from fitting a Gaussian model to the simulated ROC data was 0.80, whereas the estimate of s_{Lure} / s_{Target} obtained directly from the simulated ratings was 0.82. Thus, the two measures showed good agreement despite the clustering of ratings at the high end of the scale. Similar results were obtained when the criteria were arranged in nonsystematic (instead of equal-interval) fashion on the memory strength dimension by adding Gaussian random error with a mean of 0 and a standard deviation of 0.04 to each criterion value. In one such run, the intervals between adjacent criterion values ranged from 0.18 to 0.40 instead of being fixed at 0.28. This yielded data with apparent scale biases in addition to clustering at the high end of the scale (much as one sees in real data). Even so, the two ratio estimates were both close to 0.80, and, in multiple runs of the simulation under these conditions, they usually differed by only 0.02 or 0.03. These results show that the data are compatible with a Gaussian model despite the non-Gaussian appearance of the target frequency distribution.

If one wished to quantitatively account for nonsystematic scale biases when fitting a Gaussian model to the ratings data, then additional parameters would be needed (i.e., an equal-interval scale Gaussian model such as that shown in Figure 3 would not fit well). However, because the deviations from a true interval scale are apparently nonsystematic, these additional parameters are not needed to achieve good agreement between Gaussian ROC estimates and estimates computed directly from the ratings. For the probit model, by contrast, such agreement can be achieved only by adding the assumption that the confidence criteria are arrayed on the memory strength axis in a particular way that systematically deviates from an equal-interval measurement scale. In effect, this means that a nonlinear transformation of the participant-supplied ratings would be needed in order to make the ROC-based

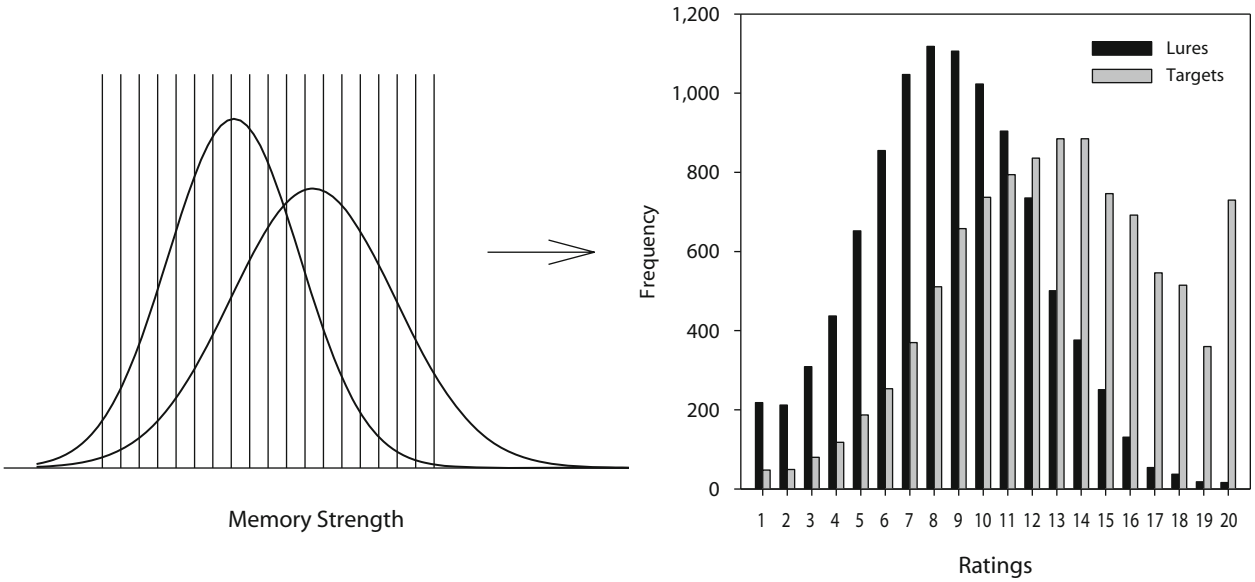

Figure 3. Unequal-variance Gaussian signal-detection model with 19 equally spaced confidence criteria (left) used to generate simulated frequency data (right). The true $\sigma_{Lure} / \sigma_{Target}$ value was 0.80. The value estimated from analyzing the simulated receiver-operating characteristic data was 0.80, and the value computed directly from the simulated confidence ratings was 0.82.

estimates and the ratings-based estimates agree for that model. Similarly, the performance of each of the non-Gaussian models shown in Figure 1 could be improved by adding a model-specific assumption about the nonlinear nature of the measurement scale. For example, exponentiating the ratings before computing the ratings-based means and standard deviations would bring those estimates into agreement with the estimates obtained from fitting the lognormal model to the ROC data. However, introducing a particular nonlinear transformation to rescue a model is tantamount to adding free parameters to that model. Ordinarily, models are penalized for depending on extra considerations like that. This is why we suggest that our findings differentially support the Gaussian model over the other models that we have considered. Moreover, as we see it, this outcome demonstrates that parametric models of latent memory strength are testable (which is the key point of contention).

The Class of Models in Question

The analyses described here bear on a class of signal-detection models consisting of two continuous distributions—one for targets and the other for lures. Among the six models that we considered that fall into that class (the five models shown in Figure 1, plus the probit model), our findings differentially support the unequal-variance Gaussian account. Many other models falling into that class, though not specifically tested here, would likely be outperformed by the Gaussian model as well. However, our findings do not differentially support the Gaussian account over models that do not fall into that class. Two such models often mentioned in connection with the analysis of asymmetrical ROCs are the mixture signal-detection model (DeCarlo, 2002) and the dual-process signal-detection/high-threshold model (Yonelinas, 1994). These models involve more than two continuous distributions. In particular, the former assumes a third (noise) distribution representing unattended items, and the latter assumes a separate memory process consisting of threshold recollection. Whether the direct rating methodology used here can help to differentiate the unequal-variance Gaussian model from those models remains to be seen.

Useful Scientific Theories Are Useful

The unequal-variance Gaussian model of latent memory strength is both a plausible and a useful model. In their classic signal-detection text, Green and Swets (1966) provided a thoughtful discussion about the Gaussian assumption of signal-detection theory (pp. 54–69). One of their points was that, a priori, there is reason to take the Gaussian assumption seriously because of the central limit theorem, which states that the sum of independent, identically distributed random variables (even non-Gaussian variables) approaches a Gaussian distribution as the number of such random variables increases. Conceivably, sensory processes (and, one might imagine, memory processes) are composed of the sum of many such variables. If so, then a Gaussian memory strength model would be expected. Green and Swets also noted that the Gaussian assumption has pragmatic utility in that it allows one to derive mathematical results that would be difficult or impossible to derive using other distributional assumptions. About this, they say, "Ultimately, the justification of any scientific assumption is pragmatic, and we shall attempt no further a priori rationalizations of this assumption" (p. 58).

The probit model would be hard-pressed to rival the Gaussian model in these respects even if it had rivaled the Gaussian model in its ability to fit the data. Finally, in terms of parsimoniously accounting for both empirical ROC data and ratings data, we have additionally shown that the Gaussian model outperforms the probit model as well as a variety of other models (at least at the group level of analysis). It is, in fact, the most useful model in that respect as well. If another theory comes along that happens to rival its ability to do that, then, instead of abandoning parametric ROC analysis in the face of such competition, the scientific method should be used to tease them apart. That, after all, is our business.

AUTHOR NOTE

This work was supported by Award R01MH082892 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. Correspondence concerning this article should be sent to J. T. Wixted, Department of Psychology, 0109, University of California, San Diego, La Jolla, CA 92093 (e-mail: jwixted@ucsd.edu).

REFERENCES

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

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Note AFCRC-TN-58–51). Bloomington: Indiana University, Hearing and Communication Laboratory.

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. New York: Academic Press.

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

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

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.

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

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.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1341-1354.

NOTE

1. For the exponential, this entailed a choice because the mean and standard deviation of an exponential distribution are the same, whereas they were not the same in the ratings (either for the targets or for the lures). We used the standard deviations from the ratings to set the exponential parameters, but the same basic story obtains if the mean values are used instead.

(Manuscript received December 15, 2009;
revision accepted for publication April 27, 2010.)