Journal of Experimental Psychology:
Learning, Memory, and Cognition
2012, Vol. 38, No. 3, 793–801

© 2012 American Psychological Association
0278-7393/12/$12.00 DOI: 10.1037/a0027040

COMMENTARY

Mixing Strong and Weak Targets Provides No Evidence Against the Unequal-Variance Explanation of *z*ROC Slope: A Comment on Koen and Yonelinas (2010)

Jeffrey J. Starns and Caren M. Rotello
University of Massachusetts Amherst

Roger Ratcliff
Ohio State University

Koen and Yonelinas (2010; K&Y) reported that mixing classes of targets that had short (weak) or long (strong) study times had no impact on *z*ROC slope, contradicting the predictions of the encoding variability hypothesis. We show that they actually derived their predictions from a mixture unequal-variance signal detection (UVSD) model, which assumes 2 discrete levels of strength instead of the continuous variation in learning effectiveness proposed by the encoding variability hypothesis. We demonstrated that the mixture UVSD model predicts an effect of strength mixing only when there is a large performance difference between strong and weak targets, and the strength effect observed by K&Y was too small to produce a mixing effect. Moreover, we re-analyzed their experiment along with another experiment that manipulated the strength of target items. The mixture UVSD model closely predicted the empirical mixed slopes from both experiments. The apparent misfits reported by K&Y arose because they calculated the observed slopes using the actual range of *z*-transformed false-alarm rates in the data, but they computed the predicted slopes using an extended range from −5 to 5. Because the mixed predictions follow a slightly curved *z*ROC function, different ranges of scores have different linear slopes. We used the actual range in the data to compute both the observed and predicted slopes, and this eliminated the apparent deviation between them.

Keywords: encoding variability, receiver operating characteristic (ROC), mixture models, signal detection models

Koen and Yonelinas (2010; hereafter "K&Y") recently reported a test of two theoretical accounts of one of the most prominent empirical regularities in recognition memory: the slope of the *z*-transformed receiver operating characteristic (*z*ROC) function. *z*ROCs are formed by plotting the *z*-transformed hit rate on the *z*-transformed false-alarm rate across multiple levels of bias. Item recognition *z*ROC functions are linear (or very close to linear) with a slope less than 1, typically around .8 (Glanzer, Kim, Hilford, & Adams, 1999; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Wixted, 2007; Yonelinas & Parks, 2007). Two popular signal detection models offer competing explanations for slopes below 1. The dual process signal detection (DPSD) model accommodates this phenomenon by mixing trials that are based on recollection versus familiarity, where the former describes the threshold retrieval of qualitative information and the latter describes a vague sense of prior occurrence that varies continuously in strength (Yonelinas, 1994). The unequal variance signal detection (UVSD) model accommodates slopes in terms of purely continuous evidence that is more variable for target items than for lure items (Egan, 1958).

Allowing higher variability in target versus lure evidence has been maligned as psychologically unjustified and purely data driven (e.g., DeCarlo, 2010). It is true that the relative variability is simply estimated as a free parameter, but the unequal-variance model does have theoretical backing. A number of recognition matching models produce more variable memory match values for targets than for lures (see Ratcliff et al., 1992). More generally, the additional variability for target items is sometimes attributed to variability in the encoding process itself (Wixted, 2007). This encoding variability hypothesis proposes that the evidence for lures follows a Gaussian distribution reflecting factors such as pre-experimental familiarity and overlap with the studied items. The evidence for targets reflects these same factors *plus* a Gaussian distribution of the encoding success for that item in the study list (Wixted, 2007). Convolving these two Gaussian distributions produces another Gaussian with higher variability, consistent with the UVSD model. Many possible sources could theoretically contribute to the increased variability due to encoding, including experimental properties such as differences in memorability across

Jeffrey J. Starns and Caren M. Rotello, Department of Psychology, University of Massachusetts Amherst; Roger Ratcliff, Department of Psychology, Ohio State University.

Preparation of this article was supported by National Institute of Mental Health Grant R37-MH44640 and National Institute on Aging Grant R01-AG17083 to Roger Ratcliff, as well as by National Institute of Mental Health Grant R01 MH60274 to Caren M. Rotello.

Correspondence concerning this article should be addressed to Jeffrey J. Starns, Department of Psychology, 441 Tobin Hall, University of Massachusetts Amherst, Amherst, MA 01003. E-mail: jstarns@psych.umass.edu

items and intrinsic factors such as variation in attention (Pratte, Rouder, & Morey, 2010).

Before continuing, we should note that the encoding variability account and the UVSD model have only partial overlap. The general UVSD model is not tied to any specific account for why variance differs between targets and lures (e.g., DeCarlo, 2010). Moreover, although we presented the encoding variability account using Gaussian distributions, the general hypothesis could be implemented under a variety of distributional assumptions. Throughout this comment, we limit discussion to a Gaussian version of the encoding variability account so we can define its predictions using the UVSD model (K&Y also used the UVSD model to get their encoding variability predictions).

To explore the role of encoding variability, K&Y evaluated *z*ROC data pooled across targets at different levels of memory strength. In the mixed condition, participants studied a list with half of the words presented for 1 s and the remaining words presented for 4 s. In the pure condition, participants studied each word for 2.5 s. Participants completed a recognition test with 6-point confidence ratings in each condition, and *z*ROCs were formed from the rating data. For the mixed condition, data from the strong (4 s) and weak (1 s) targets were combined to create a single class of target items. K&Y reasoned that if *z*ROC slope reflects variability in encoding, then mixing targets with different degrees of learning should result in a lower slope. Their results did not support this prediction: Slopes did not differ between the mixed and pure conditions, and no difference was observed even after K&Y performed a median split to evaluate just the participants with the largest strength differences in the mixed condition. Moreover, K&Y used the separate strong and weak *z*ROCs to derive direct predictions for the mixed *z*ROC slope. The predicted slopes were significantly lower than the actual mixed slopes.

K&Y also supported the dual process approach by noting its consistency with remember–know (RK) data, and they challenged the UVSD model by noting that the RK *z*ROC slopes were not consistent with the confidence rating slopes. The debate about whether RK data support a unidimensional or a dual process approach has been going on for some time, and it is not our intention to add to it here (Donaldson, 1996; Dougal & Rotello, 2007; Dunn, 2004, 2008; Hirshman & Master, 1997; Koen & Yonelinas, 2010; Rotello & Macmillan, 2006; Rotello, Macmillan, Hicks, & Hautus, 2006; Starns & Ratcliff, 2008; Wixted & Stretch, 2004; Yonelinas, 2001, 2002; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998; Yonelinas & Parks, 2007). We simply note that the link between RK and ROC estimates of recollection is not always replicated (Rotello, Macmillan, Reeder, & Wong, 2005) and that the difference between *z*ROC slopes formed from confidence and RK data has been shown to reflect a greater degree of criterion variability for RK judgments (Starns & Ratcliff, 2008; Wixted & Stretch, 2004). Readers who wish to weigh the viability of the competing accounts will find relevant discussion in the references we have provided.

For this comment, we focus on the new result reported by K&Y: the fact that mixing across strength does not decrease the *z*ROC slope. Our comment has several goals. First, we wish to show that mixing across strength is not a way to test the encoding variability account. K&Y's design creates a mixture of discrete strength classes, which is not the same as adding continuously varying amounts of target strength across items. In their analyses, K&Y actually used a mixture UVSD model with discrete strength classes to generate the predictions that they ascribed to the encoding variability account, thereby confusing two different models. To reinforce our contention that mixing and encoding variability are separate issues, we show that a decrease in *z*ROC slope based on mixing is not a unique prediction of the encoding variability account or the UVSD model: The DPSD model predicts a mixing effect on slope for the same reason as the UVSD model.

Second, we wish to show that K&Y's results are fully consistent with the mixture UVSD model, contrary to the claims of K&Y (keeping in mind that they referred to the predictions of the mixture UVSD model as the predictions of the encoding-variability hypothesis). For the comparison between the pure and mixed conditions, we show that K&Y had extremely low levels of power to find a mixing effect. K&Y also reported that their predicted mixed slopes did not match the data for individual participants. We show that the inaccurate predictions were created by an error in their modeling methods. We re-analyzed K&Y's data as well as a similar experiment from Ratcliff et al. (1994, Experiment 5) using a more appropriate procedure to define the mixed slope predictions. These analyses revealed that the predictions of the mixture UVSD model were accurate. Moreover, we note that it is not possible for a model that accounts for the unmixed data to fail to match the mixed data, and K&Y acknowledge that the UVSD model matches the separate strong and weak *z*ROC functions.

Third, we briefly discuss modeling developments that have been relatively ignored in the ROC debate up to this point. We review recent work that fits both *z*ROC functions and response time (RT) distributions with sequential sampling decision models. We note that models have successfully accommodated *z*ROC and RT data based on the unequal-variance assumption, whereas the dual process assumption has not been extended to RT data. We also discuss recent developments in hierarchical modeling of item and participant effects and note how this approach can provide stronger tests of the UVSD and DPSD models (Pratte & Rouder, 2011; Pratte et al., 2010).

The Mixing Effect on *z*ROC Slope

The first aspect of K&Y's data that they claimed was inconsistent with the encoding variability account is that *z*ROC slopes did not differ between the mixed and pure-strength conditions. In this section, we review why mixing decreases slopes and offer an explanation for the lack of a mixing effect in K&Y. The first issue to consider is how to get predictions for mixed data, and it is clear that this can be achieved only by applying a mixture model. DeCarlo (2002) proposed a mixture approach to accommodate latent classes that may be created by psychological processes, such as shifts in attention. In this way, DeCarlo's model is similar to the encoding variability account, with the critical difference that encoding processes are assumed to create distinct classes of items instead of adding continuously distributed offsets in strength. We think that encoding factors are best characterized as continuous; thus, we are not advocating the *latent* mixture approach (i.e., we do not apply the mixture model to the pure condition). However, K&Y's mixed condition introduces a case in which an *explicit* mixture is imposed by the experimenter: Targets from two strength classes are analyzed as if they were a unified set of items. For example, if one begins by assuming that the UVSD model underlies memory performance, then one assumes separate Gaussian distributions for strong and weak targets in the mixed condition. If strong and weak targets are treated as a single class, then the underlying distribution is a mixture of the two original distributions with a mixing parameter of .5 (there were an equal number of strong and weak targets on the test in K&Y, so each target trial had an equal chance of coming from the strong or weak distribution).

Critically, the mixture of two Gaussian distributions with different means and/or standard deviations is not itself a Gaussian distribution. In contrast, the encoding variability account envisions target evidence in pure-strength conditions as the *sum* of Gaussian distributions for baseline strength and encoding effectiveness, which *is* Gaussian in form (Wixted, 2007). The difference in distributional shape translates to different *z*ROC functions, with the mixture model producing nonlinear *z*ROCs in contrast to linear functions in the standard Gaussian model.

The top row of Figure 1 shows the effect of mixing across target strength under the assumption that the UVSD model generated the unmixed data. The left panel shows ROC data for strong targets, weak targets, and the mixture that results when they are combined. The mixed hit rate is simply the average of the strong and weak hit rates, so each mixed point lies midway between the strong and weak points. Mixing produces a different result in *z*-space (the right panel). Specifically, the mixed points start out about halfway between the strong and weak points but get progressively closer to the weak points as the function moves to the right (i.e., as responding becomes more liberal). This produces a lower slope for the mixed function than for either the strong or weak function; indeed, the slope of the mixed function may go below 1 even if an equal-variance model applies to the separate strong and weak data (Pratte et al., 2010; Ratcliff et al., 1992).

Figure 2 shows why mixing produces a lower *z*ROC slope by showing how raw hit rates relate to *z*-transformed hit rates. The solid line shows the cumulative distribution function for a unit normal distribution (the function used to translate hit rates to *z*-scores). The horizontal lines in each panel show a weak hit rate, a strong hit rate, and the corresponding mixed hit rate at the midpoint between them. The vertical lines show the corresponding *z*-transformed hit rates. On the *z*-dimension, the mixed hit rate is closer to the strong than to the weak hit rate when both are close to 0 (Panel A), near the midpoint of strong and weak when both are near .5 (Panel B), and closer to the weak than to the strong hit rate when both are close to 1 (Panel C). Thus, the mixed points get increasingly closer to the weak points from left to right across the *z*ROC function, creating a mixed slope lower than the original slopes.

K&Y present the lower mixed slope as a prediction of the encoding variability account based on the logic that combining distributions with two different means (different degrees of learning) should create more variability than in either of the original distributions. We do not dispute that mixing creates a more variable distribution in many situations, but it also affects distribution shape (i.e., it creates a mixture of two Gaussians as opposed to a standard Gaussian). Both of these effects contribute to the mixed

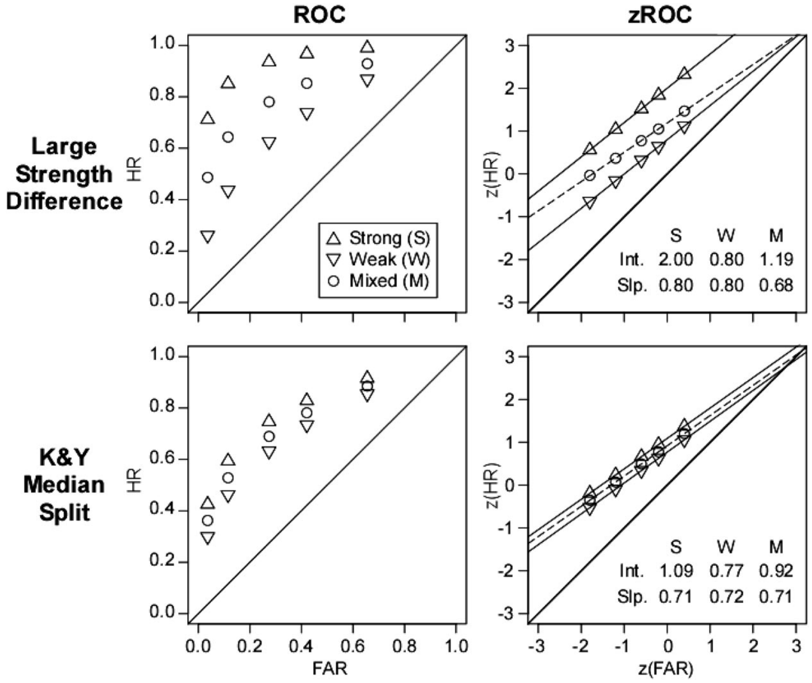


Figure 1. Effects of mixing across strong and weak targets for ROC and *z*ROC data generated from the UVSD model. The top row shows strong (S), weak (W), and mixed (M) ROC and *z*ROC functions with a large strength difference, and the bottom row shows the same with the strength difference observed by K&Y after they performed a median split to isolate the participants with the largest strength effect. HR = hit rate; FAR = false-alarm rate; Int. = *z*ROC intercept; Slp. = *z*ROC slope.

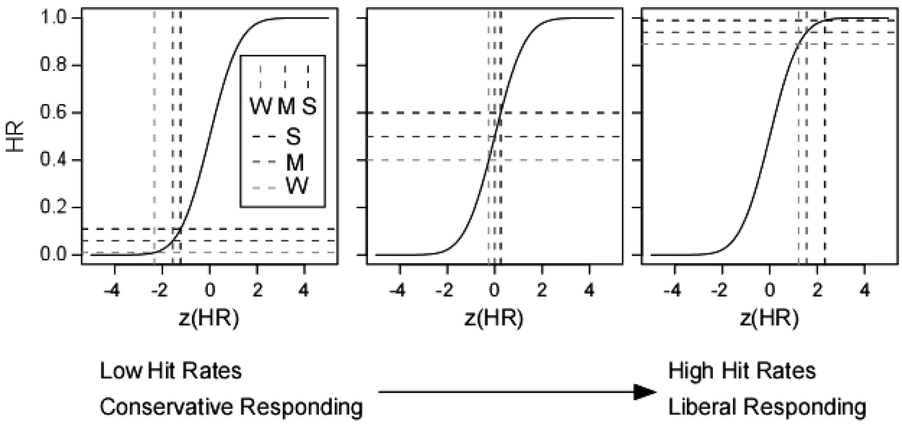


Figure 2. Demonstration of the different effects of mixing across strength in probability space and in *z*-space. Each plot shows the cumulative distribution function for a standard normal, the function used to translate hit and false-alarm rates to *z*-scores. The horizontal lines show a strong hit rate (S, top line), a weak hit rate (W, bottom line), and the hit rate that results from mixing the two (M, middle line). The vertical lines show the corresponding *z*-scores. When hit rates are low overall, the mixed *z*-score is closer to strong than to weak. When hit rates are high overall, the mixed *z*-score is closer to weak than to strong. HR = hit rate; *z*(HR) = *z*-transformed hit rate.

hit rate, and when they are combined the mixed hit rate turns out to be the average of the original strong and weak hit rates. Our discussion in the last paragraph simply explores a more general way to think about why mixing should produce lower slopes. The mixed hit rate will be the average of the strong and weak hit rates *regardless of the model that is assumed to underlie the strong and weak data*, and averaging in probability space corresponds to a lower slope in *z*-space. To emphasize this point, Figure 3 shows functions generated from the DPSD model such that the strong and weak *z*ROCs have similar slopes. Mixing the predictions across strength produces a function with a lower slope than either of the component functions. So, K&Y's claim that mixing across strength will test "novel predictions of the encoding variability and dual-process accounts" (p. 1537) is not accurate. Both accounts make the same prediction for the same reason.

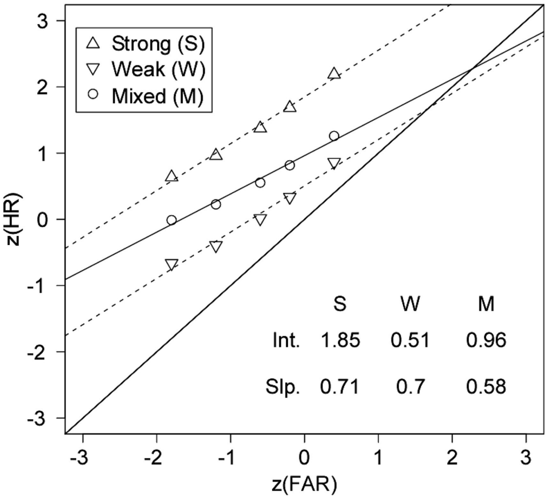


Figure 3. Predicted effects of mixing across target strength for the DPSD model with a large strength difference. As with the UVSD predictions (see Figure 1), the mixed (M) slope is lower than both the original strong (S) and weak (W) slopes. HR = hit rate; FAR = false-alarm rate; Int. = *z*ROC intercept; Slp. = *z*ROC slope.

So why did the results fail to confirm this prediction of both models? To understand the lack of an effect, one should note that a fairly large difference in strength is needed to produce a noticeable change in slope based on mixing (see Ratcliff et al., 1992, p. 527, and Pratte et al., 2010). The bottom row of Figure 1 shows the predictions for mixing *z*ROC functions with the same size strength effect as observed by K&Y. The data for both panels were generated from the UVSD model with the parameters set to create functions with the same intercepts and slopes from K&Y's median split analysis (these values are reported on page 1539 of K&Y). With a small effect on strength, mixing produces no noticeable change in *z*ROC slope. Again, the panel represents K&Y's strength difference *after* they performed a median split to isolate the participants with the largest strength effects, so the full dataset had an even smaller strength difference.

The tiny mixing effect shown in Figure 1 of course raises questions of power, so we estimated K&Y's chance of finding a mixing effect in their median-split analysis. In the mixed condition, the mixed slope changed by 0.004 relative to the strong slope and by 0.014 relative to the weak slope. Assuming that the slope for the words studied for 2.5 s in the pure condition would fall between the slopes for the weak words (1 s) and the strong words (4 s), one would expect a difference between the pure and mixed conditions of 0.004 to 0.014. Even the high end of that range is a very small difference given the variability in slope estimates (Macmillan, Rotello, & Miller, 2004). K&Y used a dependent-samples *t* test, so estimating the standardized effect size requires an estimate of the standard deviation of the difference scores. We used K&Y's reported *t* value and the slope difference between pure and mixed to solve for the standard error in the difference scores. We then multiplied this standard error by the square root of the sample size

(16) to get the standard deviation, which was .15. With a raw effect size of 0.014, the standardized effect size would be 0.093 (0.014/0.15). Achieving an 80% chance of detecting a difference with this effect size would require 910 participants with $\alpha = .05$. K&Y's actual sample size yields a power of .064, which is near the rejection rate expected with no effect at all (.05).[1]

Ironically, Figure 3 shows that K&Y should be relieved that their power was so low. If they had had appropriate levels of power and still failed to find a mixing effect, this result would have violated the predictions of the mixture DPSD model. Since a mixing effect was all but undetectable, neither UVSD nor DPSD proponents should be concerned about the result.

Participant-Level Analyses

K&Y did not base their dismissal of the encoding variability account entirely on the null effect of mixing, however. They also used the mixture UVSD model to get predicted mixed slopes for each participant in their study. The predicted slopes were significantly lower than the observed slopes, suggesting that the mixture UVSD model actually does predict a mixing effect larger than the one observed in the data. We explored this result with our own analyses. We first fit the UVSD model to the unmixed data from their mixed condition. The model had four free memory evidence parameters (the mean and standard deviation of the distributions for weak and strong targets, with the lure distribution fixed at $M = 0$ and $SD = 1$) and five criteria parameters to divide responses into the six levels of the confidence scale. We fit the model to the response frequencies for strong targets, weak targets, and lures by minimizing G^2. Next, we applied the mixture model with no free parameters: The means and standard deviations of the target distributions and the positions of the five response criteria were fixed to the values estimated in the original fits to the unmixed data, and the mixing parameter was set to .5.

Before discussing the results, we note an important property of the mixture model. When predictions are expressed as the frequency of responses in each confidence category, deriving the mixture predictions from the unmixed predictions is simple: The predicted response frequencies for lures do not change, and the predicted frequencies for targets are just the sum of the separate strong and weak frequencies at each confidence level. In other words, mixing the model predictions involves exactly the same procedure that is used to mix the data. If a model fits the unmixed data, it will necessarily fit the mixed data. No new information about the quality of the model is revealed by mixing; indeed, mixing results in a loss of information and can produce distortions in nonlinear models such as UVSD (Rouder & Lu, 2005). As noted, when predictions are expressed as hit rates and false-alarm rates, the predictions of the mixture model can be defined simply by averaging the weak and strong hit rates (the false-alarm rate predictions do not change).

Given that misfits in the mixed data can come only from misfits to the unmixed data, K&Y's report of mispredictions for the UVSD model is surprising. The UVSD model almost always provides an impressive fit to item recognition *z*ROC data (Wixted, 2007; Yonelinas & Parks, 2007), and K&Y's dataset is not an exception to this rule. Only .125 (4/32) of the individual-participant G^2 values were larger than the critical value with $\alpha = .05$ (6 degrees of freedom). Mirroring previous comparisons, the DPSD and UVSD models provided roughly equivalent fits. Matching the UVSD model, .125 of the DPSD G^2 values passed the critical value, with 17 participants better fit by the DPSD model and 15 participants better fit by the UVSD model. So one would expect that neither model would substantially miss the mixed data, yet these data were reported as uniquely challenging for the "encoding variability account" that K&Y implemented using the mixture UVSD model. We explored this mystery with our own participant-level analyses.

Figure 4 plots the predicted against the observed mixed slope for each participant. All of the points are practically on the diagonal line, indicating an extremely close match between the observed and predicted slopes. The extremely accurate predictions displayed in Figure 4 should not be interpreted as compelling new support for the UVSD model. As we mentioned, any model that fits the original strong and weak *z*ROC functions will produce an accurate prediction for the mixed slope, so the mixing procedure adds nothing to model evaluation. However, Figure 4 clearly shows that the mixed strength results cannot be interpreted as evidence *against* the mixture UVSD model or the encoding variability hypothesis that K&Y claimed to test by applying the mixture model.

An important remaining question to address is why we have reported results that contradict those of K&Y. They derived their mixture predictions by fitting the UVSD model to the separate strong and weak *z*ROCs, just as we did. Moreover, K&Y computed predicted mixed hit rates by averaging the strong and weak hit rates, which is equivalent to using the mixture model that we applied. The critical difference was the way that the predicted mixed slopes were computed. K&Y based their slopes on a large number of hypothetical criteria that corresponded to *z*-transformed false-alarm rates of -5 (essentially all "old" responses) to 5 (essentially all "new" responses).[2] We based our slopes on just the five criteria that were actually used to fit the unmixed data.

Figure 5 demonstrates how the different procedures lead to differences in the slope estimate. Panel A shows predicted mixed *z*ROC points for one of K&Y's participants. The circles show predictions across criteria ranging from *z*-transformed false-alarm rates of -5 to 5, and the solid line is the best-fitting line across this entire range. Note that the predictions of the mixture model do not exactly follow the line. Because the actual predictions follow a slightly curving path, different linear slopes can be observed by looking at different ranges of the data. For example, the triangles show the five predicted points corresponding to the five criteria used to fit the participant's data. The slope through these five

[1] A reviewer suggested that the variance of the lure distribution might be different between the pure and mixed strength conditions if participants were using a likelihood decision rule, with greater lure variance in the mixed condition. This would drive the mixed *z*ROC slope closer to 1.0 and make it even less likely that K&Y would observe a decreased slope in the mixed condition.

[2] In the original article, K&Y did not go into detail regarding their method for calculating the predicted mixed slope. We thank them for answering questions about their method via e-mail and for sending the computer file that they used to get the predictions.

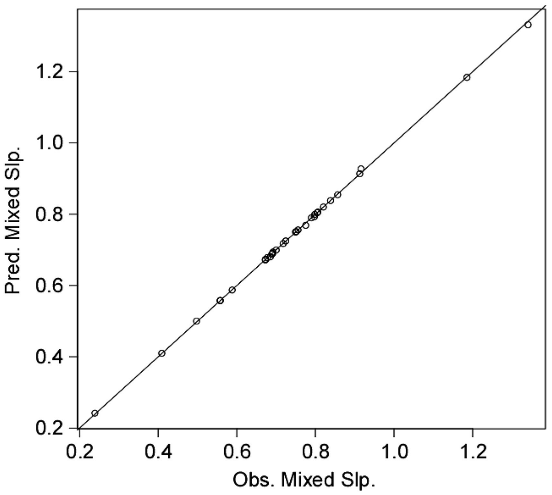


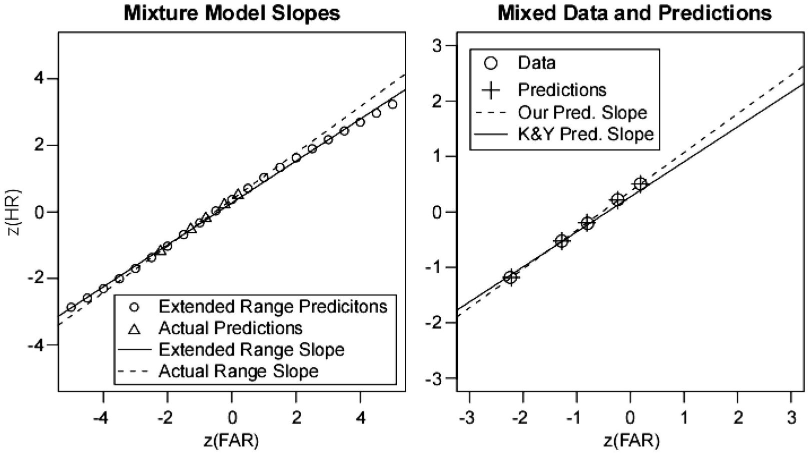
Figure 4. Observed slopes for the mixed *z*ROC functions versus the predicted slopes from the mixture UVSD model. The results are based on our own fits to K&Y's data. Each point represents a participant. Pred. = Predicted; Obs. = Observed; Slp. = *z*ROC slope.

points is shown with the dashed line, and it is noticeably higher than the slope for the full range.

The second panel of Figure 5 shows the mixed *z*ROC data and predictions. The predictions of the mixture model closely matched the data for this participant, and we again stress that the methods used by both sets of researchers agree on the positions of those predicted points. The disagreement comes in terms of the predicted *slopes*, with our prediction shown as the dashed line and K&Y's prediction shown as the solid line. The slope for the data (0.70) almost exactly matched our predicted slope (0.70), which one would expect given that the observed and predicted points are nearly in the same locations. In contrast, K&Y's slope (0.63) appears to characterize neither the predictions nor the data, falling below the predicted and observed points on the right side of the graph. The same distortion arose consistently across participants: 26 of 32 participants had lower predicted slopes over the extended range used by K&Y than over the actual range of the data points. The mean difference between the data-range and extended-range predictions was 0.030, quite similar to the mean difference between the observed and predicted results in K&Y's analyses (0.0321; the difference looks to be 0.04 in the original article because 0.7365 was rounded to 0.74 and 0.7044 was rounded to .70). Therefore, our analyses show that the misfits reported by K&Y arose because they used a different range of *z*ROC points to get their predicted and observed slopes; thus, they cannot be interpreted as inaccurate predictions of the mixture UVSD model or the encoding variability account.

Analyses of Ratcliff et al. (1994, Experiment 5)

Ratcliff et al. (1994) reported an experiment in which low and high frequency words were studied for either 1.5 s or 5 s. We used this dataset to further explore the effects of mixing targets across strength. We fit each participant's data with the UVSD model, mixed both the observations and the predictions across target strength, and evaluated the observed and predicted mixed *z*ROC slope. Replicating our results with K&Y's data, the predicted mixed slopes matched the data very closely. For high frequency words, the observed mixed slopes matched the predicted slopes to the second decimal place (M = 0.71 in both cases). Moreover, the correlation between the data and predictions across participants was 0.99. For low frequency words, the average slopes were 0.65 for the data and 0.64 for the predictions, with a correlation of 0.94. Therefore, this dataset provides additional evidence that the UVSD model correctly predicts the results of mixing across different levels of strength.


Mixture Model Slopes

Mixed Data and Predictions

Figure 5. Comparison of our method and K&Y's method for defining predicted mixed slopes. The left panel shows the predictions of the mixture model for one of the participants. The circles show the predictions spanning from *z*(FAR) = −5 to 5, and the triangles show just the five actual predictions for the participant. Computing the slope across the entire range (K&Y's method) produces a lower predicted slope than computing the slope for the five actual predictions (our method). The right panel shows the mixed *z*ROC data, the predictions of the mixture model, and the predicted slope based on K&Y's method versus our method. The mixture model very closely fits the data, but the predicted slope by K&Y's method characterizes neither the data nor the predictions.

Future Directions for ROC Research

Modeling RT and ROC Data

A critical shortcoming of both the UVSD and DPSD models is that they fail to explain RT data; thus, they cannot provide a complete account of decision processes. Attempts to model both RT and *z*ROC data have been rare. Van Zandt (2000) explored a Poisson-counter model for confidence judgments made after a two-choice ("old"/"new") decision. She showed that the model accommodated *z*ROC data, but other work indicates that Poisson-counter models do not predict the appropriate shape for RT distributions (Ratcliff & Smith, 2004). More recently, Ratcliff and Starns (2009) developed the RT confidence (RTCON) model for confidence-rating data. The model was able to account for *z*ROC slopes less than 1 while simultaneously fitting full RT distributions for each level of the confidence scale. Like the UVSD model, the RTCON results suggested that memory evidence was more variable for targets than for lures. Starns, Ratcliff, and McKoon (2012) also successfully implemented the unequal-variance assumption in the diffusion model to accommodate *z*ROCs formed from a two-choice task with a target proportion manipulation.

These studies show that the unequal-variance account of *z*ROC slopes can be extended to RT data. In contrast, the notion that slopes reflect an independent recollection process has not been implemented in a model that can fit RT distributions. Indeed, Starns et al. (2011) reported evidence that this may be a particularly challenging extension for the dual process approach. Specifically, *z*ROC slopes remained substantially below 1.0 when participants were pushed to respond very quickly (mean RT = 526 ms) even though this would presumably force participants to rely predominantly on the fast familiarity process and not on the slower recollection process (Gronlund & Ratcliff, 1989; McElree, Dolan, & Jacoby, 1999; Yonelinas, 2002). Applying the DPSD model suggested that familiarity was impaired when participants were placed under speed pressure, but recollection was not affected. Thus, the model results were exactly the opposite of what one would expect given the psychological processes purportedly measured by the familiarity and recollection parameters. Ratcliff and Starns (2009) also found that slopes were as far below 1.0 with speed-emphasis instructions as with accuracy-emphasis instructions. Perhaps difficulties such as these can be overcome, but for now only the unequal variability account successfully explains both *z*ROC and RT data.

Critically, fits of RTCON and the diffusion model show that explaining RT has profound implications for the interpretation of *z*ROC slopes. First, the sequential sampling decision process introduces variability independent of the variability in memory evidence, and this decision variability affects slope predictions. Second, changing decision criteria in the sequential sampling approach can produce changes in *z*ROC slope and even *z*ROC shape (Mueller & Weidemann, 2008; Ratcliff & Starns, 2009; Starns et al., 2011; Van Zandt, 2000). Therefore, neither slope nor shape uniquely reflects memory processes, and evaluating *z*ROC data alone cannot resolve debates about the nature of memory evidence.

Hierarchical ROC Modeling

Pratte et al. (2010) recently explored ROC data with a hierarchical Bayesian model. The model allowed them to estimate variation resulting from differences among participants or differences among specific words used in the experiment. The variation left over once these surface variables were accounted for was assumed to reflect processes intrinsic to the memory system. Their results showed that ROC functions remained asymmetrical even after participant and item effects were removed. Although they did not plot *z*ROCs directly, the asymmetry they observed in probability space corresponds to a slope less than 1.0 in *z* space. As such, Pratte et al. provided strong evidence against an encoding-variability account in which participants and items are the only influences on encoding effectiveness. However, there is no reason to limit the encoding variability account in this way. Indeed, Pratte et al. describe intrinsic memory processes that would produce variation in encoding, such as shifts in attention across learning trials (p. 226). Also, the hierarchical model cannot estimate effects resulting from the interaction of participants and items, and this interaction could be a source of encoding variability (e.g., Participant 1 vividly remembers studying "bulldog" because it was his nickname in high school, whereas Participant 2 vividly remembers studying "bridle" because she owns horses). Pratte et al. suggested that the influence of such interactions would be negligible. They simulated an equal-variance model but added an item-by-participant interaction for targets that was the same size as the item effect in their experiment. This resulted in an aggregated *z*ROC slope below 1, and a model fit to the simulated data concluded that target evidence was about 20% more variable than lure evidence (even though the data came from an equal-variance model). The authors suggested that this would be a worst case scenario, but there is no way to compare the size of the interaction effect in the simulation to the interaction effect in actual experiments (as the latter cannot be estimated). Therefore, we believe that participant-by-item interactions cannot be ruled out as at least a partial explanation for *z*ROC slopes less than 1.0.

Beyond the specific encoding-variability account, hierarchical modeling could offer a more powerful way to discriminate the general UVSD and DPSD models. For example, Pratte and Rouder (2011) developed hierarchical versions of each model and showed that the DPSD model was preferred by a fit index called the deviance-information criterion (DIC). However, the authors warn that their results should be seen as tentative. Adding RT modeling to the hierarchical approach would improve model discriminability and perhaps change key interpretations as it has for standard *z*ROC modeling (for examples of hierarchical RT modeling, see Rouder, Lu, Speckman, Sun, & Jiang, 2005; Vandekerckhove, Tuerlinckx, & Lee, 2011). Pratte and Rouder also suggested that manipulations targeted to each model's assumptions would provide a stronger basis for model discrimination. By separately modeling the effects of these manipulations on surface variables versus intrinsic memory processes, future applications of the hierarchical approach may yield important discoveries. Recognition theorists should seek ways to average over fewer variables, not mix even more together by combining strength conditions.

Conclusion

K&Y reported that mixing targets across different levels of strength violated the predictions of the encoding-variability account of *z*ROC slope. We challenged this claim on both theoretical and empirical grounds. We showed that mixing two strength classes does not produce the same outcome as adding encoding variability. Moreover, a lower slope for mixed *z*ROC functions is not a unique prediction of the encoding variability account; instead, it is a general product of the *z*-score transformation. We also showed that mixing produces noticeable effects only when strength differences are large, and K&Y's strength difference was too small to produce a slope change even after a median split of the participants. At the participant level, the mixture UVSD model accurately predicted mixed slopes for K&Y's data as well as Ratcliff et al.'s (1994) Experiment 5. The inaccurate predictions of the mixture UVSD model reported by K&Y resulted because they used an artificially extended range of *z*ROC points to calculate the predicted mixed slopes.

In reviews of this comment, K&Y repeatedly stressed that their mixing results do not provide any evidence in favor of the encoding variability hypothesis. To be clear, we completely agree. However, the mixing results also provide no evidence *against* the encoding variability hypothesis. Mixing is simply irrelevant. Directly testing the encoding variability account is an important goal, and the encoding variability account can only be regarded as provisional until it is supported or refuted by strong tests. Despite their claims, K&Y did not achieve such a test. We suggest that RT modeling and hierarchical modeling can provide new directions forward for ROC research. Strength mixing, in contrast, will play no role in advancing theories of ROC data.

References

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109,* 710–721. doi:10.1037/0033-295X.109.4.710

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

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523–533. doi:10.3758/BF03200940

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14,* 423–429. doi:10.3758/BF03194083

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review, 111,* 524–542. doi:10.1037/0033-295X.111.2.524

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review, 115,* 426–446. doi:10.1037/0033-295X.115.2.426

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

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500–513. doi:10.1037/0278-7393.25.2.500

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 846–858. doi:10.1037/0278-7393.15.5.846

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory & Cognition, 25,* 345–351. doi:10.3758/BF03211290

Koen, J. D., & Yonelinas, A. P. (2010). Memory variability is due to the contribution of recollection and familiarity, not encoding variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1536–1542. doi:10.1037/a0020448

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

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: A time course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 563–582. doi:10.1037/0278-7393.25.3.563

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

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology, 55,* 36–46. doi:10.1016/j.jmp.2010.08.007

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

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

Ratcliff, R., Sheu, C.-F., & Gronlund, S. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535. doi:10.1037/0033-295X.99.3.518

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367. doi:10.1037/0033-295X.111.2.333

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

Rotello, C. M., & Macmillan, N. A. (2006). Remember-know models as decision strategies in two experimental paradigms. *Journal of Memory and Language, 55,* 479–494. doi:10.1016/j.jml.2006.08.002

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition, 34,* 1598–1614. doi:10.3758/BF03195923

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review, 12,* 865–873.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12,* 573–604. doi:10.3758/BF03196750

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12,* 195–223.

Starns, J. J., & Ratcliff, R. (2008). Two dimensions are not better than one: STREAK and the univariate signal detection model of RK performance. *Journal of Memory and Language, 59,* 169–182. doi:10.1016/j.jml.2008.04.003

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of *z*ROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64,* 1–34.

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

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16,* 44–62. doi:10.1037/a0021765

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

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354. doi:10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General, 130,* 361–379. doi:10.1037/0096-3445.130.3.361

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. doi:10.1006/jmla.2002.2864

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember-know, process dissociation, and receiver operating characteristic data. *Neuropsychology, 12,* 323–339. doi:10.1037/0894-4105.12.3.323

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

Received December 1, 2010
Revision received August 30, 2011
Accepted September 7, 2011 ■