Psychon Bull Rev
DOI 10.3758/s13423-011-0096-7

BRIEF REPORT

The diagnosticity of individual data for model selection: Comparing signal-detection models of recognition memory

Yoonhee Jang · John T. Wixted · David E. Huber

Abstract We tested whether the unequal-variance signal-detection (UVSD) and dual-process signal-detection (DPSD) models of recognition memory mimic the behavior of each other when applied to individual data. Replicating previous results, there was no mimicry for an analysis that fit each individual, summed the goodness-of-fit values over individuals, and compared the two sums (i.e., a single model selection). However, when the models were compared separately for each individual (i.e., multiple model selections), mimicry was substantial. To quantify the diagnosticity of the individual data, we used mimicry to calculate the probability of making a model selection error for each individual. For nondiagnostic data (high model selection error), the results were compatible with equal-variance signal-detection theory. Although neither model was justified in this situation, a forced-choice between the UVSD and DPSD models favored the DPSD model for being less flexible. For diagnostic data (low model selection error), the UVSD model was selected more often.

Keywords Model mimicry · Model flexibility · Recognition memory · Unequal-variance signal-detection model · Dual-process signal-detection model

When comparing models based on goodness of fit (GOF), *model flexibility* (or *complexity*) is an important issue to address. Model flexibility refers to the ability of a model to flexibly capture any data pattern (Myung, 2000). A useful concept for understanding model flexibility is the response surface methodology (Bates & Watts, 1988), which is a plot of all possible results that a model can explain (an area in the data space). A more flexible model will cover a larger proportion of the data space, which indicates that it is difficult to find data to reject that model, as compared to all possible alternative models. However, sometimes researchers want to directly compare two leading candidate models. The relevant consideration is the area of overlap between the two models (the region in which they mimic each other), as compared to the area in the data space that is unique to each model, and in this case, *model mimicry* is an important consideration (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). In contrast to absolute flexibility, model mimicry can be thought of as a measure of relative flexibility when comparing two particular models.

Another important issue to consider when comparing two models is whether to analyze data at the individual or the group level. Cohen, Sanborn, and Shiffrin (2008) conducted model mimicry simulations fitting both individual and group data in order to determine which method was more effective in recovering the true underlying model (see also Cohen, Rotello, & Macmillan, 2008). They found that model selection based on the sum of GOF values was more accurate when the data of each individual were fit separately, provided that there were a sufficient number of data for each individual; otherwise, model selection was superior when based on group data. The experiments we analyzed collected at least 140 observations per individual—enough observations to favor individual analysis over group analysis. Thus, we only consider the results of applying models separately to each individual, which allowed us to investigate the possibility that model mimicry is not the same for all individuals.

An issue that arises when fitting two competing models at the individual level concerns the degree to which each data set yields diagnostic information (i.e., information that

Electronic supplementary material The online version of this article (doi:10.3758/s13423-011-0096-7) contains supplementary material, which is available to authorized users.

Y. Jang (✉) · J. T. Wixted · D. E. Huber
Department of Psychology, University of California, San Diego,
9500 Gilman Drive,
La Jolla, CA 92093-0109, USA
e-mail: yhjang@ucsd.edu

Published online: 03 May 2011

is capable of distinguishing between the two models). Not all participants yield diagnostic data, and it is important to consider how model mimicry varies as a function of diagnosticity. In the studies by Cohen and colleagues, the models were compared only once, based on the sum of the separate individual GOF values. Therefore, these studies did not address diagnosticity at the level of individual data (some individuals could yield more informative data than others). In the present study, we used a new individual analysis that compared models separately for each individual without summing the GOF values across individuals. Hereafter, we refer to the former as *single-comparison analysis* and the latter as *multiple-comparison analysis*. Both are based on fits of individual data but ask fundamentally different questions. The single-comparison analysis asks which model is better in a simultaneous fit of all of the separate individuals' data whereas the multiple-comparison analysis asks which model is better at fitting each individual separately. As applied to model mimicry simulations, the latter analysis allows for the possibility that model mimicry may differ for different regions of the data space.

Model comparison using model mimicry simulations

Wagenmakers et al. (2004) presented a general method to quantify model mimicry, termed the *parametric bootstrap cross-fitting method*, PBCM (see Navarro, Pitt, & Myung, 2004, for a similar technique). Consider a comparison between Model A and Model B. The PBCM begins with a nonparametric bootstrap sample: sampling with replacement from the observed data to produce a single data set that contains as many observations as exist in the original data. Next, both models are fit to this sample to estimate model-specific parameters, which are used to produce a parametric bootstrap sample according to each model. That is, using the estimated model-specific parameters for each model, a simulated data set is generated from Model A, and another simulated data set is generated from Model B. Then, both models are fit to the simulated data generated by Model A and to the simulated data generated by Model B. Finally, a GOF difference (ΔGOF) is calculated for the situation when Model A generated the simulated data, and another when Model B generated the simulated data. The entire sequence of steps is repeated *M* times, where *M* is usually a large number, to produce two ΔGOF distributions (one when Model A is true, and the other when Model B is true). The PBCM, based on simple resampling techniques, is a useful tool to measure relative flexibility given a limited data set.

The extent to which competing models are able to mimic each other is determined by the overlap between these ΔGOF distributions. If the models do not mimic each other, the distributions will not overlap, and a simple comparison is equivalent to using a criterion of zero. Another important question is whether the two distributions are symmetric about the zero point, both in terms of their average positions and their variances (generally, the shapes of the two distributions). If not, this asymmetry indicates that the two models are not equally flexible. If Model A is found to be more flexible than Model B, Model A needs to exceed Model B's fit by a certain amount (nonzero criterion), which is a penalty term to offset Model A's extra flexibility. Cohen and colleagues set this penalty to the *optimal* criterion, which is the value at which the fit of each model is equally likely (the crossover point between the distributions: for details, see Wagenmakers et al., 2004). However, as we will see below, this method does not necessarily equate the probability of making a selection error in favor of each model.

Model mimicry between two signal-detection models of recognition memory

Using model mimicry, the present study compared two signal-detection models of recognition memory: the unequal-variance signal-detection (UVSD) model (Egan, 1975) and the dual-process signal-detection (DPSD) model (Yonelinas, 1994). The two models were advanced to explain the shape of a receiver operating characteristic (ROC), which is a plot of the hit rate (HR) versus the false alarm rate (FAR), or the slope of the *z*-ROC where HR and FAR are converted to *z* scores (Green & Swets, 1966). For recognition memory, target variance is typically greater than lure variance (the *z*-ROC slope is less than 1.0), and each of these models provides a different explanation of this departure from the equal-variance signal-detection (EVSD) model. The UVSD model assumes that both distributions are Gaussian and allows a different variance for the distribution of memory strength in response to targets as compared to the distribution in response to lures. In contrast, the DPSD model assumes that responses to targets are either based on categorical recollection, which produces high confidence, or on a sense of familiarity that is drawn from a Gaussian distribution with variance equal to that of the lure distribution. Thus, both of these models collapse to the EVSD model as a special case.

Although numerous studies have compared the two models (for a review, see Wixted, 2007; see also Yonelinas & Parks, 2007), one concern is that these comparisons were usually made by raw GOF only from a single experiment, without any consideration of relative flexibility differences, or sometimes by adjustment for flexibility based only on different numbers of parameters rather than functional complexity (for details, see, e.g., Myung, 2000). Providing one exception, Cohen et al. (2008) performed model mimicry simulations with the UVSD and DPSD models (which were referred to as the one-dimensional and standard dual-process models, respectively) for experiments that collected remember–know recognition judgments. Their study was an apt application of the technique because remember–know judgments are often used to measure the separate contributions of familiarity and recollection. They found that the summed ΔGOF distributions of the two models did not overlap when fitting individual data, and concluded that the two models "are approximately equal in complexity" (p. 915).

However, there are at least two limitations to their conclusion. First, the absence of mimicry is not the same as equal flexibility. For instance, Myung, Pitt, and Navarro (2007) found situations in which models differed in flexibility, as determined by other measures, even though these models did not mimic each other. In fact, Cohen et al. (2008) found that the UVSD model distribution fell farther from zero than the DPSD model distribution, which implies that the UVSD model is more flexible. Second, and more importantly, their conclusion was based on a single-comparison analysis, and it is unclear whether it would generalize to comparisons for each individual. To examine these limitations, we compared the UVSD and DPSD models separately for each individual. To test a range of situations that have been addressed in the literature, we applied these models to simultaneous fits of yes/no and two-alternative forced-choice (2AFC) data from the same individual, and we also fit yes/no recognition data in isolation, which is the more common situation.

Method

Two previous studies: Yes/no and 2AFC recognition memory experiments

For the model mimicry applications to recognition memory, we adopted data from two previous studies that collected both yes/no and 2AFC recognition judgments (33 participants in Jang, Wixted, & Huber, 2009; 29 in Smith & Duncan, 2004), and conducted another experiment (which is described below). In both previous experiments, participants studied a list of 280 words that were presented one at a time for 5 s each. During the test phase, participants were given a randomly ordered mixture of yes/no test trials (70 targets and 70 lures) and 2AFC test trials (70 left responses correct and 70 right responses correct). Responses were collected on a 6-point rating scale for each test trial (for the yes/no test trials, *Certain YES/NO*, *Probably YES/NO*, and *Guess YES/NO*, and for the 2AFC test trials, *Certain LEFT/RIGHT*, *Probably LEFT/RIGHT*, and *Guess LEFT/RIGHT*).

Smith and Duncan (2004) fit the yes/no and 2AFC data separately. However, Jang et al. (2009) demonstrated that separate fits are inadequate tests of whether a model can generalize between test formats. Instead, they found that fitting both test formats simultaneously allowed recovery of the true model. To reach this conclusion, they used a single-comparison model mimicry analysis that only produced a probability of recovery rather than the full recovery distributions that would be necessary to determine relative flexibility.

A new yes/no recognition memory experiment: Weak versus strong memory

Because the aforementioned studies tested yes/no and 2AFC recognition in the same test block, which might influence retrieval strategies, we ran a new experiment with only yes/no recognition memory. In addition, this study manipulated memory strength (weak vs. strong). According to dual-process theories, stronger memories more often produce recollection, which tends to increase the number of highest-confidence responses for targets. Therefore, a manipulation of memory strength may be useful for distinguishing between these models.

Participants A total of 35 undergraduate students at the University of California, San Diego, participated in this experiment for course credit.

Materials The stimuli were 720 moderately high-frequency words (an average frequency of 60, according to Kučera & Francis, 1967) from four to eight letters in length.

Procedure The experiment consisted of four blocks that used the same procedure. In each block, participants studied 90 words that were presented one at a time for 1 s each. Half of the items were presented once (weak), and the other half three times (strong). The words were randomly assigned to weak and strong conditions and presented in random order, except that the items of the strong condition were not presented consecutively (to avoid massed presentations). Memory was tested using the same 6-point rating scale as in the previous experiments. Each test list contained, in random order, 45 weak and 45 strong study words for targets and 90 new words for lures.

Model mimicry simulations

According to the PBCM, 1,000 simulated data were generated for each model. The UVSD and DPSD models were fit to the data by using maximum likelihood estimation (GOF measure). The ΔGOF value was calculated by subtracting the GOF value of the UVSD model from that of the DPSD model. The model mimicry method of Cohen et al. (2008) was also used, which is identical to the PBCM, except that it does not include the initial nonparametric sample (we refer to it as the short-version method), and the results are reported in the supplemental material. By including nonparametric sampling, the PBCM entails extra variability that is related to sampling from the population of possible observations, although both methods produced qualitatively similar results.

In total, we analyzed seven data sets from 97 individuals. For both of the previously published data sets, we performed model mimicry simulations on the yes/no and 2AFC data fits simultaneously (as in Jang et al., 2009) and also on the yes/no data fits in isolation (as in Smith & Duncan, 2004). These combinations produced the first four data sets. The remaining three came from our new experiment, which only included yes/no data. We considered the weak memory data in isolation, the strong memory data in isolation, and a simultaneous analysis across the weak and strong memory data.

Results

The results below focus mainly on the multiple-comparison analysis, and Table 1 reports them in summary form.

Model comparison without adjusting for flexibility: Model fit results

Without adjusting for model flexibility (using the zero criterion), as seen in the table, 65% (ranging from 57% to 76%) of the data were fit better by the UVSD model. The ΔGOF values were negative (not significantly below zero for four of the seven data sets). On the surface, these findings suggest that the UVSD model is somewhat better able to account for yes/no and 2AFC recognition memory.

Model comparison with adjustment for flexibility: Model mimicry results

First, we briefly consider the single-comparison analysis (as in Cohen et al. 2008). The top panel of Fig. 1 shows a typical example of the ΔGOF histograms from the single-comparison analysis. This particular example shows the simultaneous fit to yes/no and 2AFC testing for the experiment of Smith and Duncan (2004). The *x*-axis indicates the ΔGOF values from the 1,000 simulated experiments when the data were generated by the UVSD model (left) versus the DPSD model (right). The *y*-axis indicates how many simulated experiments fell into each bin of the ΔGOF scale. All values of the UVSD model fall to the left side of zero, and those of the DPSD model fall to the right side of zero: The two models do not mimic each other. The two distributions are slightly asymmetrical around zero, which suggests that the UVSD model is somewhat more flexible than the DPSD model. These findings are in agreement with those of Cohen et al. (2008).

Next, we consider the multiple-comparison analysis, which produced a different set of results by considering the possibility that different models apply to different individuals. The bottom panel of Fig. 1 illustrates the ΔGOF histograms of a representative individual (who participated in the experiment of Smith & Duncan, 2004). Unlike the results of the single-comparison analysis, 10% of the white histogram (the UVSD model) falls above zero, and 14% of the gray histogram (the DPSD model) falls below zero. These areas represent how often the nongenerating (i.e.,

Table 1 Model comparison results using a multiple-comparison analysis

	Number of subjects	No Adjustment for Flexibility				PBCM Adjustment			
		UVSD (%)	ΔGOF	*t*	*p*	UVSD (%)	Optimal criterion	*t*	*p*
Smith & Duncan (2004): Y/N & 2AFC	29	76	–0.83 (1.67)	2.69	<.01	72	–0.22 (0.31)	3.80	<.001
Jang et al. (2009): Y/N & 2AFC	33	58	–0.44 (1.92)	1.32	.10	54	–0.51 (0.78)	3.76	<.001
Smith & Duncan (2004): Y/N	29	62	–0.10 (0.97)	0.55	.29	34	–0.22 (0.50)	2.40	<.05
Jang et al. (2009): Y/N	33	58	–0.33 (1.27)	1.50	.07	48	–0.46 (0.66)	4.08	<.001
Weak memory: Y/N	35	57	–0.12 (2.01)	0.36	.36	54	–0.10 (0.76)	0.76	.23
Strong memory: Y/N	35	74	–0.78 (2.05)	2.26	<.05	68	–0.22 (0.31)	4.17	<.001
Both weak & strong memory: Y/N	35	69	–0.86 (2.99)	1.69	<.05	60	–0.46 (0.62)	4.36	<.001

Standard deviations are in parentheses. UVSD = unequal-variance signal-detection model; Y/N = yes/no; 2AFC = two-alternative forced-choice; PBCM = parametric bootstrap cross-fitting method. UVSD (%) indicates how many data sets were in line with the UVSD model. ΔGOF, GOF difference, represents Log*L*(DPSD) – Log*L*(UVSD), where *L* = maximum likelihood and DPSD = dual-process signal-detection model.

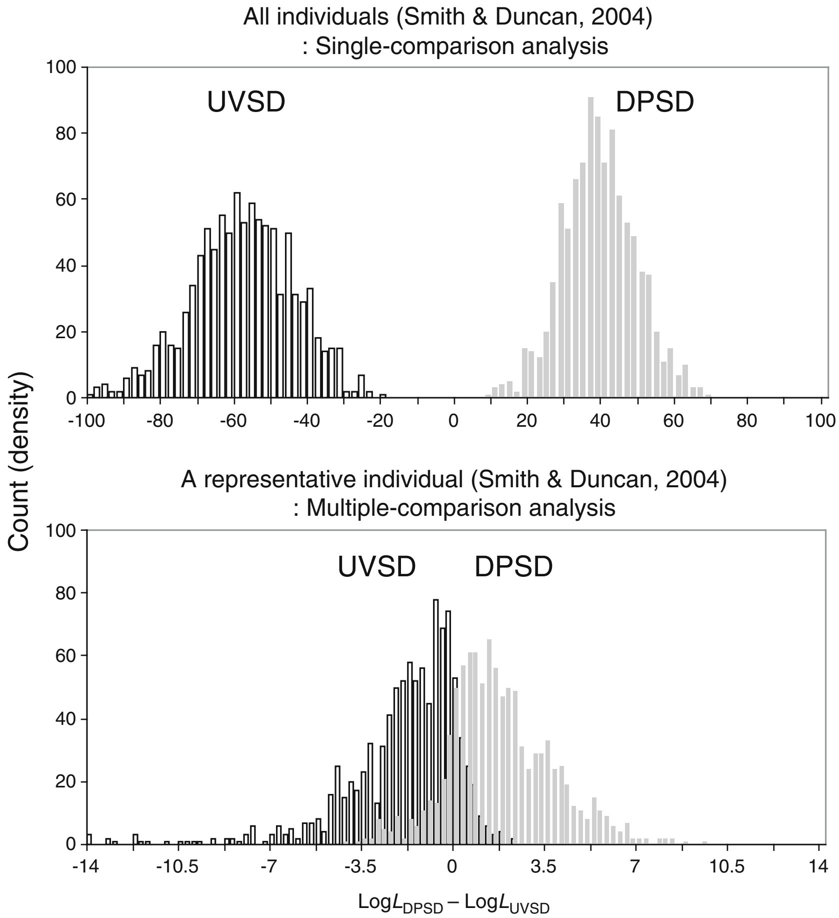

Fig. 1 Histograms of the goodness-of-fit (GOF) difference for the unequal-variance signal-detection (UVSD) and dual-process signal-detection (DPSD) models. The top panel shows results for the individual analysis on which the GOF differences were summed across all individual fits (single-comparison analysis), and the bottom panel shows results for the individual analysis of a single participant's data (an example of the multiple-comparison analysis). Note that the axes differ in scale across the panels

incorrect) model provided a better fit (in this example, the UVSD model is more flexible than the DPSD model). Such overlapping distributions were found for every individual. Thus, the default criterion of zero might not properly adjust for differences in flexibility. We therefore assessed relative flexibility for each individual separately based on the distributions from each individual data set.

As seen in Table 1, the optimal criteria were significantly below zero for six of the seven data sets. When using the optimal criteria, only 56% (ranging from 34% to 72%) of the model selections favored the UVSD model (i.e., a 9% decrease after adjusting for flexibility). These findings suggest that in general, the UVSD model is more flexible than the DPSD model.

An important caveat to the above conclusion is the quality of the data produced by each individual. In what follows, we address this limitation by calculating the probability of making a selection error. The selection error probability is the proportion of the rejected model ΔGOF distribution that is above the optimal criterion (if the rejected model ΔGOF distribution lies on the left side; otherwise, below the optimal criterion), as compared to the proportion of the selected model ΔGOF distribution. This measure assumes that the two particular models are the only possible models, and that they have equal priors. This reverses the conditional probability using Bayes' rule, resulting in a simple calculation from the cumulative frequency distributions. The advantage of this measure is that it presents the diagnosticity of the observed data, as indicated by the degree of overlap between the ΔGOF distributions.

Figure 2 shows each instance of model selection as a function of selection error rate, as indicated by the different symbols, which are placed at the arbitrary values of −0.5 (UVSD) and +0.5 (DPSD) to indicate which model won. Based on these separate model selections, the solid line plots the log ratio of the DPSD model wins to the UVSD model wins, as calculated using Gaussian smoothing over the probability of selection error. The graph clearly shows that when the data were diagnostic (i.e., low selection error), selections favored the UVSD model whereas when the data were nondiagnostic (i.e., high selection error), selections favored the DPSD model. For instance, among individuals with error rates less than 15%, the ratio of selections favoring the UVSD model versus the DPSD model was 32% to 16%. In contrast, among individuals with error rates greater than 35%, the ratio of selections favoring the UVSD model versus the DPSD model was 11% to 44%.

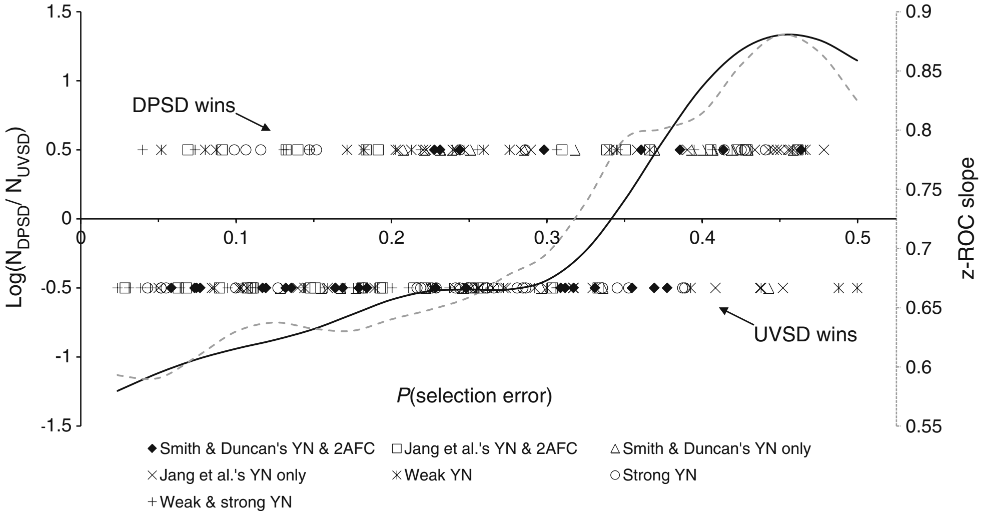

Fig. 2 Log ratio of model wins and *z*-ROC slope as a function of selection error rate. Each symbol indicates the selected model for an individual data set. The symbols are arbitrarily placed at −0.5 (UVSD) and +0.5 (DPSD) to indicate which model won. Based on these model selections, the solid line indicates the log ratio of the DPSD model wins to the UVSD model wins using Gaussian smoothing over the probability of selection error. The dashed line indicates the *z*-ROC slope using Gaussian smoothing over the probability of selection error. N = the number of wins; UVSD = unequal-variance signal-detection model; DPSD = dual-process signal-detection model; YN = yes/no; 2AFC = two-alternative forced-choice

The figure also shows the *z*-ROC slope, using the same technique of Gaussian smoothing over the probability of selection error, as indicated by the dashed line (which can be done equivalently using the recollection parameter of the DPSD model). The *z*-ROC slope indicates whether the data of an individual are compatible with the EVSD model: If the data yield a *z*-ROC slope of 1.0, neither the UVSD model nor the DPSD model is needed, and an equally good fit can be found with the EVSD model even though it contains one less parameter. As seen in the figure, as the *z*-ROC slope approaches 1.0, the selection error rate increases, and it becomes increasingly likely that the DPSD is selected. In other words, as the data become more compatible with the EVSD model, (1) the selection error rate becomes higher (it is hard to differentiate between the UVSD and DPSD models) because both models are underconstrained by the data and (2) the log ratio of wins favors the DPSD model because it is the less flexible of these two models that are both too complex in light of the data: The data that should have been assigned to the EVSD model produce a selection of the DPSD model because the selection was only between the UVSD and DPSD models.

Discussion

The present study conducted model mimicry simulations to assess the relative flexibility between the UVSD and DPSD models of recognition memory at the individual-data level. Replicating prior work, the two models did not mimic each other when the individual ΔGOF values were summed (single-comparison analysis). However, this analysis makes a single model selection under the implicit assumption that model mimicry is the same for all individuals, and it does not take into account the fact that some participants yield more diagnostic data than others. Using a single-comparison analysis, it cannot be determined whether the winning model is mainly favored when the data are nondiagnostic (which would be a less compelling outcome) or more diagnostic (which would be a more compelling outcome).

With the multiple-comparison analysis, we found that the UVSD model was generally more flexible. However, additional analyses revealed that a large number of the selections in favor of the DPSD model were based on situations of extreme mimicry in which case the probability of making a selection error was very high, and the *z*-ROC slope was close to 1.0 (in which case, neither model would be needed). When only diagnostic data were considered, corresponding to a low selection error rate, more individuals produced data in line with the UVSD model, which provided a better explanation of how the data departed from the EVSD model. These analyses demonstrate an important but often overlooked aspect of model selection: The binary decision in favor of one model over the other needs to be tempered by the diagnosticity of the data and the probability of making a selection error.

Recently, Kapucu, Macmillan, and Rotello (2010) concluded that different individuals appear to have different underlying models. We replicated this finding but also demonstrated that the assignment of different models to different individuals might be confounded with the diagnosticity of the data. From our results, we reach two conclusions: (1) When limiting model selection to the data that are capable of differentiating between these two models (i.e., data that are incompatible with the EVSD model), model selection favors the UVSD model; and (2) When the data cannot differentiate between these two models (i.e., they are compatible with the EVSD model), the DPSD model is selected by default because the UVSD model was found to be relatively more flexible. Occam's razor dictates that in the absence of any diagnostic information, the less flexible model is selected for its characteristic of being simple. However, in such situations, it is important to consider whether either model is needed, or if an even simpler alternative may suffice. The fact that this was the case for many of the individual data sets suggests that a single model comparison based on summing GOF values across individuals may be misleading by mixing together both diagnostic and nondiagnostic data.

Author Note This research was supported by National Institute of Mental Health Grant RMH081084A to D.E.H. We thank David Smith and Matthew Duncan for making their data available and Daniel Navarro for his valuable comments.

References

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember–know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review, 15*, 906–926.

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

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.

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

Kapucu, A., Macmillan, N. A., & Rotello, C. M. (2010). Positive and negative remember judgments and ROCs in the plurals paradigm: Evidence for alternative decision strategies. *Memory & Cognition, 38*, 541–554.

Kučera, H., & Francis, W. H. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Myung, J. I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*, 190–204.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review, 14*, 1043–1050.

Navarro, D. J., Pitt, M. A., & Myung, J. I. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49*, 47–84.

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.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28–50.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.

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.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics in recognition memory: A review. *Psychological Bulletin, 133*, 800–832.