Psychological Review
2011, Vol. 118, No. 1, 164–173

0033-295X/11/$12.00 DOI: 10.1037/a0020698

COMMENT

Assessing the Belief Bias Effect With ROCs: Reply to Dube, Rotello, and Heit (2010)

Karl Christoph Klauer and David Kellen
Albert-Ludwigs-Universität Freiburg

Dube, Rotello, and Heit (2010) argued (a) that the so-called receiver operating characteristic is nonlinear for data on belief bias in syllogistic reasoning; (b) that their data are inconsistent with Klauer, Musch, and Naumer's (2000) model of belief bias; (c) that their data are inconsistent with any of the existing accounts of belief bias and only consistent with a theory provided by signal detection theory; and (d) that in fact, belief bias is a response bias effect. In this reply, we present reanalyses of Dube et al.'s data and of old data suggesting (a) that the receiver operating characteristic is linear for binary "valid" versus "invalid" responses, as employed by the bulk of research in this field; (b) that Klauer et al.'s model describes the old data significantly better than does Dube et al.'s model and that it describes Dube et al.'s data somewhat better than does Dube et al.'s model; (c) that Dube et al.'s data are consistent with the account of belief bias by misinterpreted necessity, whereas Dube et al.'s signal detection model does not fit their data; and (d) that belief bias is more than a response bias effect.

Keywords: reasoning, belief bias, multinomial models, signal detection models

Supplemental material: http://dx.doi.org/10.1037/a0020698.supp

Dube, Rotello, and Heit (2010), henceforth referred to as DRH, presented a signal detection theory (SDT) analysis and a series of model comparisons for data from three experiments on syllogistic reasoning, manipulating the perceived (Experiment 1) and actual (Experiment 3) base rate of valid versus invalid syllogisms and conclusion believability (Experiments 2 and 3). A number of strong claims were based on these analyses, and if true, these claims would have important implications for modeling data on syllogistic reasoning and for accounts of belief bias. The purpose of this reply is to examine these claims on the basis of both DRH's data and old data. Before we do so, it is necessary to describe central features of DRH's procedure.

Responses were collected via confidence ratings. More precisely, for each syllogism, participants were asked to first decide whether it was valid or invalid and then give a rating of confidence in their response on a 3-point rating scale. Responses were subsequently recoded using the numbers 1 to 6, where 1 reflects a high-confidence "valid" judgment, 3 a low-confidence "valid" judgment, 4 a low-confidence "invalid" judgment, and 6 a high-confidence "invalid" judgment.

DRH based their argument on the so-called receiver operating characteristic (ROC). An ROC is a two-dimensional plot plotting two aspects of performance across different levels of response bias. In the present context, it plots the proportion of correct "valid" responses for valid syllogisms (hit rate) against the proportion of false "valid" responses for invalid syllogisms (false alarm rate). Different levels of response bias can be obtained by varying the perceived base rate of valid relative to invalid syllogisms or by varying payoff schedules (Macmillan & Creelman, 2005; McNicol, 1972; Wickens, 2002). Confidence ratings have been used to emulate ROCs in a less expensive manner. For this purpose, the differences between response levels from 1 to 6 are construed as differences in response bias. For example, to obtain the point of the ROC corresponding to the most liberal response-bias condition, only Response 6 (high-confidence "invalid" judgment) is considered an "invalid" response, whereas all other responses are treated as though the participant had responded "valid," including "invalid" Responses 4 and 5, and the hit rate and false alarm rate are computed by aggregating over Responses 1–5. For the point of the ROC corresponding to the strictest response bias, only Response 1 (high-confidence "valid" judgment) is considered a "valid" response; all other responses, including the "valid" Responses 2 and 3, are treated as though the participant had responded "invalid." Moving across the response scale in this fashion, an ROC with 5 points is emulated. In what follows, we will refer to the emulated ROC as a confidence-based ROC, whereas ROCs based on a binary response format will be referred to as binary ROCs.

DRH fitted an SDT model, implying nonlinear ROC, to data from three experiments and judged that it provided reasonable fits to their data. In contrast, a number of multinomial models (Batchelder & Riefer, 1999) with linear ROCs did not fit. They concluded that ROCs are nonlinear for data on belief bias in syllogistic reasoning. As DRH pointed out, nonlinear ROCs would invalidate

Karl Christoph Klauer and David Kellen, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

The research reported in this article was supported by Grant Kl 614/31-1 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer.

Correspondence concerning this article should be addressed to Karl Christoph Klauer, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany. E-mail: christoph.klauer@psychologie.uni-freiburg.de

traditional analyses of the data in terms of linear models such as analyses of variance.[1]

DRH argued that the particular multinomial models they fitted represent appropriate extensions of Klauer, Musch, and Naumer's (2000) multinomial model of belief bias. That model was developed to account for data collected in a binary (valid vs. invalid) format and cannot be applied to confidence-rating data without modification. DRH concluded that their data was not consistent with Klauer et al.'s model.

On the basis of two null findings for effects of conclusion believability on parameters quantifying reasoning performance (Experiments 2 and 3), DRH furthermore concluded that belief bias in syllogistic reasoning is just a response-bias effect and that there are no effects of conclusion believability on reasoning. They also concluded that their data were inconsistent with accounts of belief bias in terms of selective scrutiny, misinterpreted necessity, mental models, metacognitive uncertainty, verbal reasoning theory, modified verbal reasoning theory, and the selective processing theory,[2] whereas the only theory of belief bias consistent with their data was claimed to be the one provided by SDT.

In what follows, we consider each of these conclusions in turn, beginning with the issue of nonlinearity of ROCs.

The Shape of the ROC in Reasoning Data

The bulk of data collected on belief bias in syllogistic reasoning has employed a binary (valid vs. invalid) response format (for a review, see Klauer et al., 2000). Unfortunately, it is an open question how confidence-based ROCs relate to binary ROCs. DRH drew heavily on the literature on recognition memory in which the shape of ROCs has been examined for a couple of decades and in which nonlinear confidence-based ROCs have frequently been observed. There is, however, a recent article by Bröder and Schütz (2009) in that literature that has received surprisingly little attention.[3] Bröder and Schütz argued as others have (Erdfelder & Buchner, 1998; Klauer & Kellen, in press; Malmberg, 2002) that confidence ratings may create rather than reveal nonlinear ROCs due to variations, across and within participants, in response style. A meta-analysis of data from 59 studies and three new experiments conducted by Bröder and Schütz suggests that binary ROCs may be linear in the field of recognition memory, underlining empirically the important theoretical argument that nonlinear confidence-based ROCs do not imply nonlinear binary ROCs. We elaborate on this point later.

Considering binary ROCs, there is evidence for nonlinear shapes in some domains (e.g., in perception; Egan, Schulman, & Greenberg, 1959) and evidence for a linear shape in others (e.g., in working memory; Rouder et al., 2008). The issue is moot in recognition memory (Bröder & Schütz, 2009).

Most studies on belief bias in syllogistic reasoning collected binary (valid vs. invalid) responses. Taking the previously mentioned arguments into account, there may therefore not be a nonlinearity problem to begin with in this literature. The nonlinearity problem postulated by DRH may instead be created by their use of confidence ratings. It would therefore be good to have more positive reassurance for the possibility that nonlinearity of ROCs generalizes beyond the use of confidence ratings in the field of syllogistic reasoning.

Fortunately, there are published data sets, some of them of considerable size, that can be used to address the issue. We fitted Klauer et al.'s (2000) multinomial model and the SDT model to the 10 data sets reported by Klauer et al. that employed a manipulation of response bias via base rate and a binary response format. For the binary response format, the multinomial model implies linear ROCs, whereas the SDT model implies nonlinear ROCs.[4]

Table 1 shows the results for these 10 data sets (Study 8 in Klauer et al., 2000, did not employ a base rate manipulation, so the models cannot be fitted to the results of that study). The table reports the goodness-of-fit index G^2, the associated p values (small values indicating misfit), and model-selection indices Akaike's information criterion (AIC) and Bayesian information criterion (BIC) as used by DRH for comparing models (models with smaller values are preferred).[5] Because DRH did not employ syllogisms with neutral conclusions, syllogisms with neutral conclusions as used in Klauer et al.'s Studies 1, 3, 5, and 7 were excluded for the fits reported in Table 1. Including them leads to the same conclusions; in fact, the results become even stronger in the direction summarized next.

As can be seen in Table 1, the multinomial model outperforms the SDT model in 8 of 10 cases in terms of AIC and in 10 of 10 cases in terms of BIC. The differences in AIC values, but not those in BIC values, are numerically small for each individual study, but it is possible to enhance the information-to-noise ratio via the aggregation principle.

One way to do this is to test whether the differences in AIC values and BIC values are significant across the 10 data sets. The difference in AIC values between the two models is significant in a Wilcoxon test ($Z = -2.29$, $p = .02$), as is that in BIC values ($Z = -2.80$, $p = .01$). Another way to do this is to consider the 10 data sets as one big data set and to compute AIC and BIC for the joint

[1] As pointed out by Klauer et al. (2000), even linear ROCs would question such analyses unless the slope of the linear ROC is 1. Klauer et al. proposed a model-based approach to remedy this problem.

[2] DRH argued, however, that their data and model were broadly consistent with broader theories of reasoning such as Chater and Oaksford's (1999) probability heuristics model of syllogistic reasoning.

[3] DRH also did not deal with current criticisms suggesting that the interpretation of confidence-based ROCs entertained in their article is inadequate (Benjamin, Diaz, & Wee, 2009; Ratcliff & Starns, 2009; Rosner & Kochanski, 2009).

[4] Details on how these analyses were done including R scripts (R Development Core Team, 2009), HMMTree equation files (Stahl & Klauer, 2007), and data files can be found through the supplemental materials link at the beginning of the article or at http://www.psychologie.uni-freiburg.de/Members/klauer/r-scripts.zip/

[5] In fitting the SDT model to the data from Study 7, we had to put an upper bound on the model parameters, because maximum likelihood estimation led to unrealistically large values for these parameters. The upper bound was 3, an unrealistically large value for any of the model parameters. The problem arises because in this study invalid believable syllogisms were accepted as frequently as valid believable syllogisms, as predicted by Klauer et al. (2000). This occasionally occurs (see, e.g., the data set presented as an introductory example by DRH in their Table 1) but should not happen, according to the SDT model. This pattern of belief bias causes problems in estimating the SDT variance parameter. Excluding Study 7 from the analyses altogether does not change the results, including the outcome of the significance tests reported next.

Table 1
Fit Indices and Model-Selection Indices for Data Sets by Klauer et al. (2000)

Study and data set[a]	Multinomial model				Signal detection model			
	G^{2b}	*p*	AIC	BIC	G^{2c}	*p*	AIC	BIC
Study 1	3.43	.49	63.61	98.46	3.19	.20	67.36	110.92
Study 2								
Naive	4.53	.34	94.19	148.75	3.75	.15	97.40	165.60
Expert	4.08	.40	95.69	153.62	3.05	.22	98.66	171.08
Study 3	1.42	.84	62.53	95.92	0.59	.75	65.70	107.44
Study 4								
Naive	9.45	.05	104.12	160.32	5.47	.07	104.13	174.39
Expert	7.94	.09	104.26	162.56	2.75	.25	103.07	175.94
Study 5	3.40	.49	55.36	88.75	1.65	.44	57.60	99.34
Study 6								
Naive	1.74	.78	84.95	139.31	1.38	.50	88.60	156.54
Expert	21.21	<.01	106.70	165.36	15.52	<.01	105.01	178.33
Study 7	10.15	.04	57.45	89.05	10.27	.01	61.56	101.07

Note. G^2 = goodness-of-fit index; AIC = Akaike's information criterion; BIC = Bayesian information criterion.
[a] Naive and expert refer to data from participants who reported no prior experience with formal logic and who did report such experience, respectively. [b] $df = 4$. [c] $df = 2$.

data (with different parameters estimated for each individual study). This yields a difference in AIC of 20.24 in favor of the multinomial model and a difference in BIC of 198.91. According to the rules of thumb stated by Burnham and Anderson (2005, Chapter 2), a difference in AIC values larger than 10 observed in a large data set means that the model with the larger AIC (i.e., the SDT model) has essentially no empirical support (Burnham & Anderson, 2005, p. 70) relative to the model with the smaller AIC (i.e., the multinomial model).

In sum, there is surprisingly strong evidence for the multinomial model.[6] These findings thereby parallel those obtained by Bröder and Schütz (2009) in the field of recognition memory. The conclusion, by DRH's standards, is that as far as we can tell on the basis of the available data, ROCs are linear for binary response formats. This suggests that the use of confidence ratings creates rather than reveals problems of nonlinearity (for reasons elaborated on later). Because most studies on belief bias are based on binary responses, the nonlinearity problem postulated by DRH may be largely nonexistent. Another conclusion is that the multinomial model describes the old data significantly better than does the SDT model.

DRH presented one analysis involving Klauer et al.'s (2000) original model for DRH's Experiment 3 with data dichotomized (see DRH's Table 9). Unfortunately, they used 3 as degrees of freedom for that model, but the degrees of freedom of the model equals 4.[7] This changes the values of *p*, AIC, and BIC for the multinomial model. Less importantly, the likelihood terms in AIC and BIC are computed wrongly. According to our reanalysis of the multinomial and the SDT models, respectively, *p* values are .23 and .85, AIC values are 93.01 and 91.72, and BIC values are 143.14 and 154.37. Considering model fit, there is no indication in the *p* values of significant model violations for either model. Considering the model-comparison indices, the two models are more or less tied on AIC, whereas the multinomial model performs considerably better in terms of BIC. This suggests that Klauer et al.'s model provides, if anything, a better description of DRH's data than does the SDT model.

Note that Klauer et al.'s (2000) model also shows the null effect of believability on reasoning parameters that the SDT model analyses exhibit. The reasoning parameters of the multinomial model—r_{vb}, r_{vu}, r_{ib}, r_{iu}—measure the participants' ability to determine the validity or invalidity of syllogisms, separately, for valid (*v*) and invalid (*i*) syllogisms with believable (*b*) and unbelievable (*u*) conclusions. For the dichotomized data of DRH's Experiment 3, the H_0 of no effect of belief on reasoning, $r_{vb} = r_{vu}$ and $r_{ib} = r_{iu}$, can be maintained ($\Delta G^2 = 0.96$, $df = 2$, $p = .62$). DRH chose not to report this, although they presented the analogous information for the SDT model applied to the dichotomized data (i.e., the H_0: $\mu_{vb} = \mu_{vu}$ and $\sigma_{vb} = \sigma_{vu}$ can be maintained). Instead, they reported that $r_{vb} = r_{iu} = r_{vu}$ can be maintained, whereas it is not possible to set all four *r* parameters equal to each other, suggesting that r_{ib} differs from the other three once

[6] One statement found in the SDT literature is that empirical ROCs with more than 3 points are more diagnostic for discriminating between models with differently shaped predicted ROCs than empirical ROCs based on 3 points, as in the Klauer et al. (2000) data (e.g., Bröder & Schütz, 2009). If so, this would make it difficult to obtain a clear decision in favor of one of the two models on the basis of the Klauer et al. data, rendering the current outcome the more impressive.

[7] Only the ratio of parameters β_u and β_b, but not their absolute values, is identified in Klauer et al.'s (2000) model. To fix the scale, β_b is set equal to 1 a priori and is therefore not a free parameter. This does not imply that the "true" value of β_b, if it could be identified, is 1.

[8] Consider an analogously focused test strategy for the SDT model. The parameters governing reasoning performance in the SDT model are the means μ_{xy} and standard deviations σ_{xy} of the distributions of valid ($x = v$) and invalid ($x = i$) syllogisms with believable ($y = b$) and unbelievable ($y = u$) conclusions with $\sigma_{iu} = \sigma_{ib} = 1$ and $\mu_{iu} = \mu_{ib} = 0$ imposed a priori. The four σ parameters and the four μ parameters cannot simultaneously be set equal; that is, the σ and/or μ parameters differ as a function of validity and/or belief ($\Delta G^2 = 345.66$, $p < .01$, with $df = 4$ due to the a priori constraints). In a second step, it is seen, however, that the four σ parameters can be set equal (to one): $\Delta G^2 = 2.86$, $df = 2$, $p = .24$. Once these have been set equal, effects of belief on the two μ parameters not constrained a priori are "revealed"; that is, the H_0: $\mu_{vb} = \mu_{vu}$ must be rejected ($\Delta G^2 = 5.28$, $df = 1$, $p = .02$).

these have been set equal. DRH asserted that the multinomial model concludes that there are effects of belief on the reasoning stage. There is, however, little justification for this focused test strategy, and none such is pursued for the SDT model.[8]

Klauer et al.'s (2000) Model for Confidence Ratings

Klauer et al.'s (2000) model was designed for binary data. Its purpose was to provide a measurement tool to measure reasoning accuracy for the four kinds of syllogisms typically investigated in studies of belief bias (i.e., valid and invalid syllogisms crossed with believable and unbelievable conclusions), correcting for response bias and possible effects of belief on it. One issue in Klauer et al. was that response bias should be controlled for in evaluating reasoning performance for each kind of syllogism.

To extend the model to confidence ratings, we found it convenient to present the model in two parts, a stimulus-state mapping and a state-response mapping (Klauer & Kellen, in press).

Stimulus-State Mapping

The stimulus-state mapping specifies how each kind of syllogism is mapped on a number of unobservable mental states. In Klauer et al.'s (2000) most basic model, there are two detection states: M_1, in which a valid syllogism is correctly detected as valid, and M_2, in which an invalid syllogism is correctly detected as invalid. There are also two states of uncertainty—M_{3b} and M_{3u}—in which participants are uncertain about the syllogism's logical status and in which responses are based on an informed guessing process that draws on extralogical information such as conclusion believability. For M_{3b}, the logical status (valid vs. invalid) of a given believable syllogism is not detected, and for M_{3u}, the logical status of a given unbelievable syllogism is not detected.

The stimulus-state mapping is depicted in Figure 1. The parameters r of the stimulus-state mapping provide the probabilities with which detection states M_1 and M_2 rather than the uncertainty states M_{3b} and M_{3u} are reached. For example, given a valid (v), believable (b) syllogism, the probability of reaching state M_1 is r_{vb} and that of reaching state M_{3b} is $1 - r_{vb}$. The stimulus-state mapping is independent of response format and should not be changed in adapting the model to deal with different response formats if the resulting model is to be consistent with Klauer et al.'s (2000) model for binary data.

State-Response Mapping

The state-response mapping specifies how states are mapped on responses. For binary responses, detection states M_1 and M_2 lead to "valid" and "invalid" responses, respectively, deterministically. In uncertainty states, response guessing occurs that may be biased by conclusion believability (and other extralogical cues such as base rate). Thus, in state M_{3b} (M_{3u}), the "valid" response is guessed with probability a_b (a_u), and the "invalid" response with probability $1 - a_b$ ($1 - a_u$).

In modeling confidence ratings, only the state-response mapping needs to be adapted; the stimulus-state mapping is independent of response format. Adapting the state-response mapping is straightforward. Table 2 shows a plausible state-response mapping following the one used by Klauer and Kellen (in press) for modeling confidence ratings. As for the case of binary responses, and in line with the definition of detection states, detection states M_1 and M_2 are mapped on "valid" and "invalid" responses in a deterministic fashion. There are, however, three "valid" and three "invalid" responses that can occur, and the probabilities of using these are modeled, following Klauer and Kellen, by three parameters: s_l, s_m, and s_h for the ratings expressing lowest, medium, and highest confidence, respectively. The three s parameters have to sum to 1, so there are only two free parameters to be estimated. Because people differ in their propensity to use extreme ratings (known as extreme response style; Hamilton, 1968), and because there is intraindividual variation in scale usage (e.g., Haubensak, 1992), it is not reasonable to assume that detection states such as M_1 and M_2 are invariably mapped on highest confidence responses (see Onyper, Zhang, & Howard, 2010, for a similar assumption in the SDT framework). The scale-usage parameters s_l, s_m, and s_h capture interindividual and intraindividual variation in scale usage. They are not a function of the syllogism's validity or believability. Note that the ROC implied by this model is nonlinear if s_h is smaller than 1. That is, interindividual and intraindividual variations in scale usage, leading to some less than highest confidence responses in detection states, cause nonlinear ROCs according to this model.

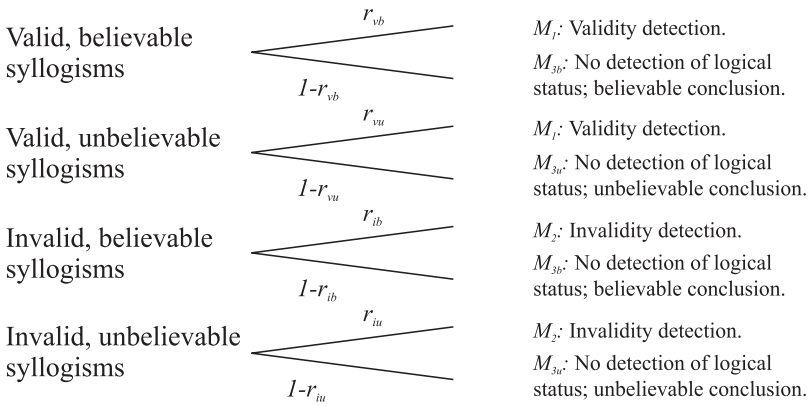

Figure 1. Stimulus-state mapping of Klauer et al.'s (2000) basic model.

Table 2
State-Response Mapping for Klauer et al.'s (2000) Model Extended to Confidence Ratings

Mental state	"Valid" response			"Invalid" response		
	1	2	3	4	5	6
M_1	s_h	s_m	s_l	0	0	0
M_2	0	0	0	s_l	s_m	s_h
M_{3b}	$a_b(1)$	$a_b(2)$	$a_b(3)$	$a_b(4)$	$a_b(5)$	$a_b(6)$
M_{3u}	$a_u(1)$	$a_u(2)$	$a_u(3)$	$a_u(4)$	$a_u(5)$	$a_u(6)$

Note. M_1 = detection state in which a valid syllogism is correctly detected as valid; M_2 = detection state in which an invalid syllogism is correctly detected as invalid; M_{3b} = state of uncertainty in which the logical status of a given believable syllogism is not detected; M_{3u} = state of uncertainty in which the logical status of a given unbelievable syllogism is not detected; s_h = highest confidence parameter; s_m = medium confidence parameter; s_l = lowest confidence parameter; $a_b(1)$ to $a_b(6)$ = guessing parameters for believable syllogisms; $a_u(1)$ to $a_u(6)$ = guessing parameters for unbelievable syllogisms.

The guessing parameters $a_b(1), \ldots, a_b(6)$ and $a_u(1), \ldots, a_u(6)$ correspond conceptually to the parameters a_b and a_u, respectively, for the binary response format. Because each of $a_b(1), \ldots, a_b(6)$ and $a_u(1), \ldots, a_u(6)$ have to sum to 1, there are only five parameters to be estimated per mental state M_{3b} and M_{3u}. The state-response mapping thus comprises 12 parameters (i.e., 10 nonredundant guessing parameters and two nonredundant scale-usage parameters). Taken together with the four r parameters governing the stimulus-state mapping, the model requires 16 parameters and thus two more than the SDT model for confidence ratings.[9]

Model Fits and Model Comparisons for DRH's Data

Table 3 presents the results of a reanalysis of DRH's data with Klauer et al.'s (2000) model for confidence ratings and DRH's SDT model. Parameter estimates for the multinomial model are shown in Table 4. In their own analyses, DRH presented the fit and model-selection indices per condition for each data set (e.g., separately for the data from believable and unbelievable syllogisms; see DRH's Tables 4, 6, and 12). However, when models such as the ones fitted by DRH and the present model have parameters (e.g., the present s parameters) that are shared by the different conditions, the condition-wise indices G^2, p, AIC, and BIC do not have the intended statistical interpretations. Table 3 therefore presents the results per experiment.[10]

In terms of goodness of fit, the SDT model is inconsistent with three of the four data sets; that is, the goodness-of-fit statistic G^2 indicates significant violations of the model assumptions with p smaller than .001 for three of four data sets. This is surprising given DRH's strong reliance on the SDT model (DRH did not report p values). Goodness of fit approaches more acceptable levels for the multinomial model, although there is certainly room for improvement (G^2 indicates significant model violations with $p < .01$ in one case and with $p < .05$ in a second case).

Note, however, that for large data sets, there is high power to detect even tiny model violations. To take this into account, several approaches have been considered in the literature. One possibility is to compute compromise power analyses with effect size $w = .10$ (a small effect according to Cohen, 1988) and a β/α ratio = 1 for each data set (Bröder & Schütz, 2009), adjusting the level of significance α according to sample size. Another possibility is to use relaxed criteria for p and G^2, as found in the literature on structural equation modeling (Schermelleh-Engel, Moosbrugger, & Müller, 2003), according to which $.05 < p \leq 1.0$ corresponds to a good fit and $.01 \leq p \leq .05$ to an acceptable fit, whereas $0 \leq G^2 \leq 2$ *df*s corresponds to a good fit and 2 *df*s $< G^2 \leq 3$ *df*s to an acceptable fit. Whichever of these criteria is used, the SDT model is rejected in three of four cases, whereas the multinomial model is rejected in one or two cases, depending upon the criterion used.

In terms of model-selection indices AIC and BIC, the multinomial model outperforms the SDT model in three of four cases for AIC and in two of four cases for BIC, although one of these latter cases is more or less a tie. Taken together with the goodness-of-fit results, the multinomial model would probably be preferred on the basis of the results.

Given the poor goodness of fit of the SDT model, it would probably be prudent not to interpret its parameter estimates further (see Footnote 10). Nevertheless, DRH reported that manipulations of base rate (Experiments 1 and 3) as well as manipulations of conclusion believability (Experiments 2 and 3) map on the response criteria, whereas there are mostly no significant effects on the parameters governing reasoning performance. The same pattern of results is obtained for the multinomial model: Manipulations of base rate (Experiments 1 and 3) as well as manipulations of conclusion believability (Experiments 2 and 3) map on the guessing parameters capturing response bias (smallest $\Delta G^2 = 21.48$, largest $p < .01$), whereas there are no effects on the r parameters governing reasoning performance (largest $\Delta G^2 = 4.39$, smallest $p = .13$).

Taken together, DRH's data are inconsistent with the SDT model and somewhat better described by Klauer et al.'s (2000) model for confidence ratings. This implies that what is wrong with

[9] In Experiment 1, DRH manipulated perceived base rate in two steps. Base rate here takes the role of conclusion believability in the multinomial model, as it does in the SDT model. In Experiment 3, they manipulated perceived base rate in three steps. In analyzing the data as a function of base rate, base rate again takes the role of conclusion believability, but there are now three levels for this factor. Here, the multinomial model has 23 parameters and the SDT model 21 parameters. The parameters for the multinomial models are six r parameters (2 [valid vs. invalid] $\times$ 3 [3 base rates]), 15 nonredundant guessing parameters (five per base rate condition), and two nonredundant s parameters for scale usage.

[10] Presenting the results per experiment also supports DRH's intention to compare parameters across conditions. For example, DRH wished to determine whether response criterion parameters differ significantly between syllogisms with believable and unbelievable conclusions. The log-likelihood ratio test for this comparison contrasts (a) a baseline model modeling the conditions with believable and unbelievable conclusions jointly with separate response criterion parameters for each condition and (b) a constrained model in which these parameters are set equal across conditions. The validity of this test hinges on the assumption that the baseline model is valid, that is, that its goodness-of-fit statistic G^2 does not already indicate significant model violations. Table 3 presents the appropriate G^2 values and associated p values.

Table 3
Fit Indices and Model-Selection Indices for Dube, Rotello, and Heit's (2010) Data Sets

Study and data set	Multinomial model					SDT model				
	G^2	*df*	*p*	AIC	BIC	G^2	*df*	*p*	AIC	BIC
Experiment 1	12.70	4	.013	44.70	136.35	7.25	6	.298	35.25	115.45
Experiment 2	14.04	4	.007	46.04	127.70	22.94	6	<.001	50.94	122.39
Experiment 3[a]										
Believability	6.66	4	.155	38.66	138.91	46.82	6	<.001	74.82	162.54
Base rate	13.12	7	.069	59.12	203.23	30.95	9	<.001	72.95	204.53

Note. AIC and BIC values are shown minus an additive constant (i.e., by adding the appropriate penalty to G^2). G^2 = goodness-of-fit index; AIC = Akaike's information criterion; BIC = Bayesian information criterion.
[a] Believability and base rate refer to analyses aggregating across base rate and believability, respectively.

the particular multinomial models fitted by DRH is not what they have in common with Klauer et al.'s original model; the culprit is the set of auxiliary assumptions that DRH used to extend that model to confidence ratings.[11] What we consider to be reasonable auxiliary assumptions (as already employed by Klauer & Kellen, in press) lead to reasonable results.

The present section addressed the question of whether Klauer et al.'s (2000) model is consistent with DRH's data and how it characterizes the nature of base rate and believability manipulations in that data. This is a different question from the one considered in the previous section, that is, whether ROCs in reasoning data are nonlinear. DRH suggested, however, that there is a strong link between the two questions. For example, DRH stated in the abstract of their article: "In all cases, the form of the empirical ROCs was curved and therefore inconsistent with the assumptions of Klauer, Musch, and Naumer's (2000) multinomial model of belief bias." DRH discussed one multinomial model, MPT-R, that they believe is capable of generating nonlinear confidence-based ROCs, but in fact it generates strictly linear ROCs.[12]

Additional Model Analyses and the Account by Misinterpreted Necessity

These model analyses also provide deeper insights for the DRH data. Reasoning parameters *r* for participants' performance in detecting the syllogisms' logical status are invariably very small for invalid syllogisms, whether they are believable or not (see Table 4). In fact, the hypothesis $r_{ib} = r_{iu} = 0$ can be maintained for each of DRH's data sets (largest $\Delta G^2 = 2.23$, smallest $p = .33$). Again, base rate takes the role of conclusion believability in the analyses reported in this section for the data sets in which base rate rather than believability was manipulated (see Footnote 9).

DRH's invalid syllogisms are indeterminately invalid; that is, the conclusion is possible, but not necessary, given the premises. $r_{ib} = r_{iu} = 0$ means that participants were unable to detect the invalidity of the invalid syllogisms and that instead they relied exclusively on extralogical information (such as base rate and believability) in responding to these syllogisms. This is what one would expect under the account of belief bias in terms of misinterpreted necessity (Evans, 1989).

As argued by Evans and Pollard (1990); Newstead, Pollard, Evans, and Allen (1992); and Klauer et al. (2000), a viable version of the account by misinterpreted necessity should permit response bias for valid syllogisms. Participants are not always able to determine the validity of valid syllogisms, but they are not allowed to withhold the response when they are not. It is likely that the response is then influenced by extralogical information in guessing under uncertainty even for valid syllogisms. Expressed in terms of Klauer et al.'s model, the account by misinterpreted necessity then reduces to the claims (a) that $r_{ib} = r_{iu} = 0$ and $r_{vb} = r_{vu}$, which can be upheld for each of DRH's data sets (largest $\Delta G^2 = 5.81$, smallest $p = .16$) and (b) that $r_v = r_{vb} = r_{vu}$ is larger than zero and, in fact, the hypothesis that $r_v = 0$ can be rejected for each of DRH's data sets (smallest $\Delta G^2 = 109.14$, largest $p < .01$).

To summarize, $r_{ib} = r_{iu} = 0$ means that responses to the invalid syllogisms are exclusively driven by response bias as predicted by misinterpreted necessity. In turn, $r_v = r_{vb} = r_{vu} > 0$ implies that response bias governs responses for valid syllogisms to a proportionally smaller extent, given by $1 - r_v < 1$, as predicted by misinterpreted necessity, whereas no effect of belief on reasoning must be assumed (i.e., $r_{vb} = r_{vu}$ and $r_{ib} = r_{iu}$) in line with the account by misinterpreted necessity. These hypotheses capture the essence of the version of the account in terms of misinterpreted necessity considered by Evans and Pollard (1990) and Newstead et al. (1992) and formalized by Klauer et al. (2000). If response bias favors syllogisms with believable conclusions, an interaction of

[11] Thus, DRH's model MPT1 predicts (for all practical purposes) zero frequencies of highest confidence "valid" and "invalid" responses for invalid and valid syllogisms, respectively. Because a substantial number of responses were observed in these cells, any model predicting zero frequencies for these cells produces pronounced misfit, whether it is consistent with Klauer et al.'s (2000) model or not and whether it predicts linear or nonlinear ROCs. Models MPT2 and MPT3 are not consistent with Klauer et al.'s model, because a proportion ε of the responses governed by the *r* parameters is in fact mapped on (highest confidence) wrong responses. For example, a proportion $r_{vu}(1 - \varepsilon)$ of responses is mapped on highest confidence "valid" responses consistent with Klauer et al.'s model, but a proportion $r_{vu}\varepsilon$ of responses flow into highest confidence "invalid" responses inconsistent with Klauer et al.'s model. Model MPT-R does not make this assumption, but it assumes that detection states are invariably mapped on highest confidence correct responses, which as explained in the body of the text is implausible.

[12] For MPT-R, it follows from the equations given in DRH's Appendix A (dropping the subscript *y*) that hit rate H and false alarm rate FA are related by $H_c = r_v + (1 - r_v)/(1 - r_i)FA_c$ for each point of the confidence-based ROC ($c = 1, \ldots, 5$) of a given believability or base rate condition.

Table 4
Multinomial Model Parameter Estimates for Dube, Rotello, and Heit's (2010) Data Sets

	r		*a*						*s*		
Study and condition	Valid	Invalid	1	2	3	4	5	6	s_h	s_l	s_m
Experiment 1									.72	.25	.03
Liberal	.41	.00	.31	.27	.08	.03	.16	.15			
Conservative	.27	.11	.11	.22	.03	.11	.28	.25			
Experiment 2									.69	.26	.05
Believable conclusion	.64	.00	.29	.27	.05	.06	.16	.17			
Unbelievable conclusion	.53	.00	.13	.15	.04	.06	.31	.31			
Experiment 3: Belief[a]									.70	.21	.10
Believable conclusion	.60	.01	.31	.19	.09	.11	.15	.16			
Unbelievable conclusion	.52	.04	.18	.12	.05	.11	.22	.31			
Experiment 3: Base rate[b]									.72	.20	.09
Liberal	.54	.00	.32	.19	.11	.06	.16	.17			
Neutral	.51	.11	.33	.14	.08	.12	.11	.22			
Conservative	.46	.02	.13	.16	.04	.13	.28	.26			

[a] Refers to the analysis aggregating across the base rate. [b] Refers to the analysis aggregating across believability.

validity and belief is thereby predicted for the raw (unmodeled) acceptance frequencies.

DRH claimed that their data are inconsistent with any of the existing accounts of belief bias and that the only theory of belief bias consistent with their data is the one provided by SDT. In fact, their data are consistent with the version of misinterpreted necessity considered by Evans and Pollard (1990) and Newstead et al. (1992) as formalized by Klauer et al. (2000). Although the strength of this conclusion is somewhat weakened by the less-than-optimal goodness of fit of Klauer et al.'s model, the goodness of fit of the SDT model to DRH's data is clearly unsatisfactory.

Are Multinomial Models More Flexible Than SDT Models?

DRH anticipated that multinomial models (other than Klauer et al.'s, 2000, model) may provide adequate fits of their data, but as they stated, their purpose was not to document the flexibility of the multinomial model framework. This raises the legitimate question whether the present relative success of Klauer et al.'s (2000) model reflects nothing more than the possibly greater flexibility of this model relative to SDT models in accounting for data in general. The only formal evidence comparing SDT models and multinomial models in terms of flexibility that we are aware of is a simulation study by Bröder and Schütz (2009). That study suggested that the SDT model with unequal variance is more flexible than a multinomial two-high-threshold model similar to Klauer et al.'s basic model.

From a broader point of view, a cursory review of the SDT literature reveals that the toolbox for SDT models to date includes moving from one to two dimensions (e.g., Hautus, Macmillan, & Rotello, 2008), adding an additional distribution for "unattended" targets (e.g., Hautus et al., 2008), adding a discrete detection state (e.g., Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996), adding guessing noise (e.g., Hautus et al., 2008), adding decision noise (e.g., Mueller & Weidemann, 2008), using other than normal distributions for noise and signal trials (e.g., DeCarlo, 1998), and so forth. This endows the SDT approach with a high level of structural flexibility and makes it likely that a modified SDT model might be found that provides adequate fits of DRH's data.

Is Belief Bias a Response Bias Effect?

On the basis of two null findings for effects of conclusion believability on reasoning parameters (Experiments 2 and 3), DRH concluded that belief bias does not affect reasoning performance and that instead, as the title of their article puts it, "it's a response bias effect." Null findings in the sense of an absence of an interaction of validity and believability are not unusual in the literature on belief bias, and the conditions under which the interaction is observed is an issue of theoretical debate (e.g., Newstead et al., 1992).

Belief bias often takes the form of lowered accuracy for conflict syllogisms in which logic and believability suggest opposite responses (i.e., for invalid believable and for valid unbelievable syllogisms). In the SDT model, this pattern of belief bias is mapped on response bias and does not show up in the parameters quantifying reasoning performance (i.e., in μ_i, μ_v, σ_i, and σ_v) unless the drop in accuracy is substantially asymmetric for the two kinds of conflict syllogisms. Alternatively, belief bias of this kind could be mapped as a belief effect on reasoning parameters μ_i and μ_v (Wickens & Hirshman, 2000), as acknowledged by DRH. This is because only the differences $\mu_v - c_k$ and $\mu_i - c_k$ between distribution means and response criteria c_k, but not their absolute values, are identified in the SDT model. Shifting the response criteria of, say, unbelievable syllogisms c_{ku} relative to those of believable syllogisms by an additive constant (an effect of belief on response bias) therefore has the same effect as shifting the distribution means of valid and invalid unbelievable syllogisms μ_{vu} and μ_{iu} in the opposite direction relative to those of believable syllogisms (an effect of belief on the reasoning stage). But the parameterization chosen by DRH rules out the second mapping (μ_{iu} is set equal to μ_{ib} a priori) and enforces the first. Thus, whatever the data, it cannot be ruled out that belief bias, where it exists, is more than response bias, using the SDT model.

Using the SDT model, it can, however, be shown that belief bias is more than response bias. We reanalyzed Klauer et al.'s (2000) data sets using the SDT model, performing tests for an effect of conclusion believability on reasoning performance for each data set in the same manner as DRH did for their data. Although Klauer et al.'s model describes Klauer et al.'s data sets better than the SDT model does, the overall goodness of fit of the SDT model was sufficient in this case to warrant further analyses of its parameters (see Table 1). Due to the just-mentioned limitation of the SDT model, only certain patterns of belief bias can be detected by the SDT analysis, so that one cannot expect to find an effect of belief bias in each and every data set. Nevertheless, as shown in Table 5, there were significant effects of conclusion believability on the SDT reasoning parameters in four of 10 cases. Meta-analytically, over the 10 data sets, the ΔG^2 values sum to 86.08, which is significant ($df = 20$, $p < .01$). In conclusion, the available data suggest, by DRH's standards, that belief bias is more than a response bias effect, even according to the less appropriate metric provided by the SDT model.

Implications for Theories of Belief Bias

DRH and Klauer et al. (2000) agree that traditional analyses of reasoning data in terms of raw acceptance rates are overly naive and susceptible to scale artifacts that can be mitigated through the use of appropriate models. What can be concluded from the present debate over and above this methodological point?

DRH's data are consistent with the account of belief bias in terms of misinterpreted necessity within the limits set by the less-than-optimal goodness of fit of Klauer et al.'s (2000) model to DRH's data. DRH's data thereby add to the considerable support that exists for this account in the literature (e.g., Klauer et al., 2000; Newstead et al., 1992). However, when more focused tests of it are implemented (e.g., through the use of simpler, one-model, indeterminately invalid syllogisms; Klauer et al., 2000; Newstead et al., 1992), its predictions are often not confirmed.

Table 5
Tests for Effects of Believability on Reasoning Accuracy in the Signal Detection Model

Study and data set[a]	G^{2b}	p
Study 1	1.26	.53
Study 2		
Naive	1.83	.40
Expert	0.47	.79
Study 3	0.77	.68
Study 4		
Naive	26.62	<.01
Expert	8.15	.02
Study 5	6.51	.04
Study 6		
Naive	1.75	.42
Expert	35.19	<.01
Study 7	3.53	.17

Note. G^2 = goodness-of-fit index.
[a] *Naive* and *expert* refer to data from participants who reported no prior experience with formal logic and who did report such experience, respectively. [b] $df = 2$.

There is little support for DRH's claim that belief bias is just response bias, which if true would have been an important addition to, and correction of, the literature. Empirically, the claim rests on two null findings and on a model that does not fit the data, whereas that same model applied to old data produces considerable counterevidence against it. Theoretically, the claim rests on a model that for mathematical reasons cannot rule out the possibility that belief bias, where it exists, has an effect on reasoning, whatever the data. Nevertheless, the idea that belief bias is primarily a response-bias phenomenon is an old one (e.g., Evans, Barston, & Pollard, 1983) and deserves further study, because it would offer an attractively simple account of belief bias.

From a broader perspective, it could be argued that SDT models are more in tune with probabilistic theories of reasoning (e.g., Oaksford & Chater, 2007). The probabilistic approach roughly claims that responses are based on the subjective probability of the conclusion in the presence of the premises or on considerations of measures of statistical information. Both the SDT model and probabilistic theories thereby invoke a latent strength-of-evidence scale on which responses are based, as noted by DRH.

In contrast, multinomial models incorporate discrete all-or-none detection states that are more in tune with theories invoking a kind of mental logic, be it in terms of rules or mental models. Such theories assume that reasoners are at least in some instances capable of arriving at a firm valid versus invalid decision that they then translate into responses (not necessarily using highest confidence response categories where rating responses are required). However, whether or not detection states are reached can by itself be modeled as deriving from a latent strength-of-evidence dimension with a criterion placed on it (e.g., Klauer, 2010), suggesting that these analogies do not carry much weight. For such reasons, we prefer to view the simple models used here as measurement tools.

Limitations of the SDT Model as a Measurement Tool

One, perhaps modest, way to think of SDT models and multinomial models is as measurement tools (rather than theories) quantifying task performance in terms of dependent variables that provide better controls for confounded processes than do analyses of raw data. Thus, Klauer et al. (2000) viewed their model as a measurement tool providing a means to assess reasoning performance with response bias controlled for, and they spelt out the predictions of different accounts of belief bias for the pattern of model parameters consistent with each account. Similarly, the SDT model provides parameters for reasoning performance with response bias controlled for. Measurement models are not free of assumptions—assumptions that need to be tested and supported by the data if the measurement is to be more valid than are analyses of raw data. Batchelder and Riefer (1999) elaborated on the relationship between measurement models and underlying psychological theory and on how to validate measurement models.

One limitation of the SDT model as a measurement tool is that it does not allow one to assess reasoning performance separately for valid and invalid syllogisms. It only provides a measure of the effect of validity, that is, a score for the difference between acceptability levels of valid and invalid syllogisms (on probit-transformed data). In contrast, the multinomial model allows one to assess reasoning performance separately for each of the four

kinds of syllogisms typically administered in studies on belief bias, and the model analyses were thereby able to show that the DRH data are consistent with the account of belief bias by misinterpreted necessity.

Another limitation of the SDT model as a measurement tool was already explained: It does not allow one to detect belief bias taking the form of lowered reasoning accuracy for conflict syllogisms. Yet, according to Klauer et al.'s (2000) theory, belief bias should often take this form.

A final limitation is that the SDT approach is not as well developed as the multinomial-modeling approach in dealing with individual differences between participants and artifacts caused thereby when data are aggregated across participants. Individual differences in reasoning ability, in response bias, and in response style are ubiquitous (e.g., Stanovich & West, 2000). The present analyses as well as DRH's analyses are based on aggregated data. Yet, fitting nonlinear models to aggregated data leads to all of the statistical errors that DRH set out to correct and to additional fallacies (e.g., Klauer, 2006; Rouder & Lu, 2005). Individual data are usually too sparse to allow one to fit the kind of model considered here to each individual's data, but there is a compromise between the extremes of individual-level analyses and analyses of the aggregated data, namely the hierarchical-modeling approach (Raudenbush & Bryk, 2002). Easy-to-use software exists for multinomial models implementing a hierarchical-model extension to safeguard against these fallacies (Stahl & Klauer, 2007), but similar tools are still lacking for signal detection analyses, although they are certainly within reach (Rouder & Lu, 2005).

Summary

In summary, reanalyses of the available published data and of DRH's data suggest (a) that the ROC is linear for binary response formats as employed by the bulk of research on belief bias in syllogistic reasoning; (b) that the use of confidence ratings creates rather than reveals nonlinear ROCs; (c) that Klauer et al.'s (2000) model describes the available data with binary response format significantly better than does DRH's model, and that it describes DRH's data somewhat better than does DRH's model; (d) that DRH's data are consistent with the account by misinterpreted necessity within the limits set by the less-than-optimal goodness of fit of Klauer et al.'s model to DRH's data, whereas DRH's SDT model does not fit DRH's data; and (e) that belief bias is more than a response bias effect.

Model comparisons as presented by DRH are useful even if it turns out, as in the present case, that existing models do a better job. For example, negative results may inspire researchers to probe into the matter more deeply, by identifying and corroborating a flaw or limitation in existing data, by collecting new and diagnostic data, or by developing yet another model that may do an even better job. In pursuing one of these avenues, it may be wise to rely on the binary response format, which, being more constrained, may be less vulnerable to unwanted variance in terms of variations in scale usage than are confidence ratings.

References

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 587–606.

Burnham, K. P., & Anderson, D. R. (2005). *Model selection and multimodel inference* (2nd ed.). New York, NY: Springer.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38,* 191–258.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3,* 186–205.

Dube, C., Rotello, C., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review.*

Egan, J. P., Schulman, A. L., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *The Journal of the Acoustical Society of America, 31,* 768–773.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127,* 83–96.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11,* 295–306.

Evans, J. St. B. T., & Pollard, P. (1990). Belief bias and problem complexity in deductive reasoning. In J.-P. Cavernie, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 131–154). Amsterdam, the Netherlands: Elsevier (North-Holland).

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69,* 192–203.

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 303–309.

Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review, 15,* 889–905.

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika, 71,* 7–31.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75,* 70–98.

Klauer, K. C., & Kellen, D. (in press). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review.*

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107,* 852–884.

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

Malmberg, K. J. (2002). Observations on the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380–387.

McNicol, D. (1972). *A primer in signal detection theory.* London, England: Allen & Unwin.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494.

Newstead, S. E., Pollard, P., Evans, J. St. B. T., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45,* 257–284.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality.* Oxford, England: Oxford University Press.

Onyper, S. V., Zhang, Y., & Howard, M. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General, 139,* 341–362.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

R Development Core Team. (2009). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rosner, B. S., & Kochanski, G. (2009). The law of categorical judgment (corrected) and the interpretation of changes in psychophysical performance. *Psychological Review, 116,* 116–128.

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.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, 105,* 5975–5979.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online,* 8. Available from http://www.dgps.de/fachgruppen/methoden/mpr-online/

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods, 39,* 267–273.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23,* 645–726.

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

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical decision theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review, 107,* 377–383.

Yonelinas, A. P., Dobbins, I. G., Szymanski, M. D., Dhaliwal, H. S., & King, S. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition, 5,* 418–441.

Received March 8, 2010
Revision received June 1, 2010
Accepted June 3, 2010 ■