Psychological Review
2010, Vol. 117, No. 3, 831–863

0033-295X/10/$12.00 DOI: 10.1037/a0019634

Assessing the Belief Bias Effect With ROCs: It's a Response Bias Effect

Chad Dube and Caren M. Rotello
University of Massachusetts

Evan Heit
University of California, Merced

A belief bias effect in syllogistic reasoning (Evans, Barston, & Pollard, 1983) is observed when subjects accept more valid than invalid arguments and more believable than unbelievable conclusions and show greater overall accuracy in judging arguments with unbelievable conclusions. The effect is measured with a contrast of contrasts, comparing the acceptance rates for valid and invalid arguments with believable and unbelievable conclusions. We show that use of this measure entails the assumption of a threshold model, which predicts linear receiver operating characteristics (ROCs). In 3 experiments, subjects made "valid"/"invalid" responses to syllogisms, followed by confidence ratings that allowed the construction of empirical ROCs; ROCs were also constructed from a base-rate manipulation in one experiment. 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. We propose a more appropriate, signal detection–based model of belief bias. We then use that model to develop theoretically sound and empirically justified measures of decision accuracy and response bias; those measures demonstrate that the belief bias effect is simply a response bias effect. Thus, our data and analyses challenge existing theories of belief bias because those theories predict an accuracy effect that our data suggest is a Type I error. Our results also provide support for processing theories of deduction that assume responses are driven by a graded argument-strength variable, such as the probability heuristic model proposed by Chater and Oaksford (1999).

Keywords: belief bias, signal detection, multinomial modeling, syllogistic reasoning, response bias

In a recent ruling by the British Court of Appeals, judges concluded that, contrary to the ruling of a lower court, Pringles are in fact potato chips (Associated Press, 2009). One potential result of this decision is the application of the British value-added tax to Pringles. Much of the argument concerning the chip status of the Pringle hinged on the opinions of high authorities as to whether the percentage of potato content in a Pringles chip (actually less than 50%) was large enough to classify it as a potato chip. The logical structure of the argument can be analyzed by arranging the information as follows.

Some of a Pringle is sliced potatoes.

Some sliced potatoes are potato chips.

____

*A Pringle is a potato chip. (A)

This argument is not logically valid, but the conclusion may be compelling to Pringles fans.

The tendency to accept or reject a conclusion on the basis of its consistency with everyday knowledge, regardless of its logical status, is known as *belief bias* (e.g., Cherubini, Garnham, Oakhill, & Morley, 1998; Evans, Handley, & Harper, 2001; Evans, Newstead, & Byrne, 1993; Markovits & Nantel, 1989; Roberts & Sykes, 2003; Shynkaruk & Thompson, 2006). Belief bias is typically studied using categorical syllogisms, which are similar in structure to the argument in Example A. Syllogisms contain two premises and a conclusion, constructed with three terms: the predicate (X), which is the nonrepeated term of the first premise; the middle term (Y); and the subject (Z), which is the nonrepeated term of the second premise. An allowable conclusion links the subject and predicate terms via their relationship to the middle term; it may be either valid (following necessarily from the premises) or invalid. An abstract example of a valid syllogism is given in Example B:

All X are Y.

No Y are Z.

____

No Z are X. (B)

The arrangement of the premise terms is referred to as the syllogistic *figure*. Four arrangements are possible: Y-X, Z-Y (Syllogistic Figure 1); X-Y, Z-Y (Syllogistic Figure 2); Y-X, Y-Z (Syllogistic Figure 3); X-Y, Y-Z (Syllogistic Figure 4). The argument in Example B is an example of a syllogism in Syllogistic Figure 4. Traditionally, each premise of the syllogism can take one of four quantifiers: "all," "no," "some," and "some . . . are not."

Chad Dube and Caren M. Rotello, Department of Psychology, University of Massachusetts; Evan Heit, School of Social Sciences, Humanities and Arts, University of California, Merced.

This research was supported by National Science Foundation Collaborative Research Grant BCS-0616979 to Evan Heit and Caren M. Rotello. Chad Dube was supported as a trainee on National Institutes of Health Grant T32 MH16745. We thank Lissette Alvarez, Brooklynn Edwards, Efferman Ezell, V. Karina Ibanez, Chanita Intawan, Markie Johnson, Nic Raboy, Haruka Swendsen, and Jonathan Vickrey for their assistance in this research. We also thank Arndt Bröder, Christoph Klauer, and Mike Oaksford for their comments.

Correspondence concerning this article should be addressed to Chad Dube, Department of Psychology, Box 37710, University of Massachusetts, Amherst, MA 01004-7710. E-mail: cdube@psych.umass.edu

The effects of structural factors such as figure and quantification on syllogistic inference have been studied extensively (Begg & Denny, 1969; Dickstein, 1975, 1978, 1981; Johnson-Laird, 1983; Revlis, 1975; Woodworth & Sells, 1935).

Evans, Barston, and Pollard (1983), in an important early study of belief bias, showed that the believability of a given syllogistic conclusion can have drastic effects on the probability that subjects will endorse it. In three experiments, subjects were asked to evaluate the validity of four types of syllogisms, which resulted from crossing the logical status and believability of the conclusions. Their design and results are summarized in Table 1. Consider, for example, the problem in the lower middle cell of Table 1. The conclusion in this case is logically invalid but consistent with everyday knowledge, and 92% of their subjects endorsed this conclusion type as valid. The problem in the lower right cell of Table 1 has logical structure identical to that of the problem in the lower middle, but in this case, the conclusion is not consistent with everyday knowledge. Only 8% of subjects endorsed this conclusion type as valid. Evans et al.'s subjects were not completely insensitive to the logical structure of the arguments, however, as they also accepted more valid than invalid conclusions. A third finding, which has since been replicated in a number of studies, was that the difference in acceptance rates for valid and invalid problems is greater when problems are unbelievable than when they are believable (38% vs. 0% in Table 1). The interaction effect appears to stem from the very low acceptance rate of invalid unbelievable problems, though the precise nature of the Evans et al. result remains unclear. In particular, it is not clear whether the effect is primarily due to a deductive reasoning process that is disrupted by the believability of the conclusion, an evaluation of believability that modulates the reasoning process, or some mixture of the two. Explaining the interaction has been a major goal of extant theories of belief bias.

Rarely discussed in the belief bias literature is the fact that the measurement of the interaction effect and, indeed, any measurement of accuracy in the task necessarily involve some correction for subjects' overall willingness to endorse conclusions. One fairly intuitive method of correcting for response bias effects is to simply subtract erroneous responses, which presumably reflect the tendency to prefer one type of response over another, from correct responses: corrected score = P("Valid"|Valid) − P("Valid"|Invalid). This subtraction can be done separately for believable and unbelievable problems. The interaction effect described by Evans et al. (1983) can thus be estimated by a contrast of contrasts. This measure, often referred to as the *interaction index,* has provided the basic datum for a substantial number of studies that have investigated the Belief × Logic interaction (e.g., Ball, Phillips, Wade, & Quayle, 2006; Evans et al., 1983; Evans & Curtis-Holmes, 2005; Evans, Newstead, Allen, & Pollard, 1994; Morley, Evans, & Handley, 2004; Newstead, Pollard, Evans, & Allen, 1992; Quayle & Ball, 2000; Roberts & Sykes, 2003; Shynkaruk & Thompson, 2006; Stupple & Ball, 2008; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003). Rewriting the index, we can denote P("Valid"|Valid) as *H* (the hit rate), and P("Valid"|Invalid) as *F* (the false-alarm rate). Then, using *B* and *U* to denote believable and unbelievable problems, respectively, a definition of the interaction index is as follows.

$$\text{Interaction index} = (H_U - F_U) - (H_B - F_B). \qquad (1)$$

Positive values of the interaction index are typically observed because the effect of validity is larger for problems with unbelievable conclusions. (For the data in Table 1, the interaction index = [.46 − .08] − [.92 − .92] = .38.)

As much of the research inspired by the results of Evans et al. (1983) has focused on explaining the Belief × Logic interaction, the importance of measuring the effect accurately should not be underestimated. In what follows, we review several theoretical accounts of belief bias, all of which attempt to explain the interaction observed in contrasts of *H* and *F*. We discuss problems arising from assumptions about the relationship between response bias and accuracy that are inherent in the interaction index. We describe the analysis of *receiver operating characteristics* (ROCs), which can be used to assess the appropriateness of various measurement indices such as $H - F$ or d'. Then, in a series of three experiments, we compare the results from analyses of response proportions and ROCs. We propose a new model of belief bias that is justified by the empirical ROCs and show that it fits our data better than an existing model (Klauer, Musch, & Naumer, 2000). We conclude that inappropriate assumptions about the relationship between bias and accuracy, implicit in the interaction contrast, have impeded understanding of the belief bias effect and that all of the existing theories of belief bias are affected. Our model suc-

Table 1
The Design of Evans, Barston, and Pollard (1983, Experiment 1), Example Problems, and Rates of Endorsement

	Conclusion	
Syllogism	Believable	Unbelievable
Valid	No cigarettes are inexpensive. Some addictive things are inexpensive. Therefore, some addictive things are not cigarettes. P("Valid") = 92%	No addictive things are inexpensive. Some cigarettes are inexpensive. Therefore, some cigarettes are not addictive. P("Valid") = 46%
Invalid	No addictive things are inexpensive. Some cigarettes are inexpensive. Therefore, some addictive things are not cigarettes. P("Valid") = 92%	No cigarettes are inexpensive. Some addictive things are inexpensive. Therefore, some cigarettes are not addictive. P("Valid") = 8%

Note. Adapted from "On Belief Bias in Syllogistic Reasoning," by K. C. Klauer, J. Musch, and B. Naumer, 2000, *Psychological Review, 107,* p. 853. Copyright 2000 by the American Psychological Association.

cessfully describes the relationship between accuracy and response bias and provides a more parsimonious explanation of the effect.

Theoretical Accounts of Belief Bias

The first account of the belief bias effect was proposed by Evans et al. (1983) and later termed the *selective scrutiny model.* Selective scrutiny proposes that subjects focus initially on the conclusions of the arguments and accept believable conclusions without any evaluation of their logical validity. When conclusions are not believable, subjects reason through the premises and accept or reject conclusions on the basis of their perceived validity. Selective scrutiny could thus be seen as a process whereby logic-based responding is driven by the (un)believability of conclusions (belief→logic); the Belief × Logic interaction occurs because deductive reasoning is used only for unbelievable conclusions. Recent work employing reaction time and eye-tracking measures (reviewed below), as well as experiments comparing conclusion-production and conclusion-evaluation tasks, does support the idea that conclusion believability has an influence on the processing of premises (e.g., Ball et al., 2006; Morley et al., 2004). However, other data are problematic for the theory: Selective scrutiny predicts that there will be no effect of validity for believable problems, yet such effects are often observed (see Klauer et al., 2000, for a meta-analysis).

A second explanation of the belief bias effect that has gained attention in the literature is the *misinterpreted necessity* model (Dickstein, 1981; Markovits & Nantel, 1989; Newstead et al., 1992). In contrast to selective scrutiny, misinterpreted necessity predicts that subjects will engage in reasoning at the outset and only rely on believability after reaching conclusions that are consistent with, but not necessitated by, the premises. An example of this state of affairs is given by the following abstract problem (Example C):

Some X are Y.

No Z are Y.

____

*Some Z are not X. (C)

When conclusions are indeterminately invalid, subjects may become confused or uncertain and fall back on conclusion believability to make their decisions. Misinterpreted necessity views belief-based responding as an escape-hatch mechanism when deductive reasoning is inconclusive (logic→belief). It provides a sensible explanation of the finding of increased sensitivity to belief on invalid problems because, by definition, only invalid problems can lead to indeterminate conclusions.

Newstead et al. (1992) provided evidence both for and against misinterpreted necessity. Across two initial experiments, they varied whether conclusions were determinately or indeterminately invalid and only obtained the Belief × Logic interaction when problems were of the latter variety. In a third experiment, however, the interaction was not obtained despite the use of indeterminately invalid problems. The reason for this apparent inconsistency will become clear shortly. A further weakness of the misinterpreted necessity model, shared with selective scrutiny, is its inability to account for effects of believability on valid conclusions (Klauer et al., 2000; Newstead et al., 1992).

A third theory that features prominently in the literature on belief bias is situated in the *mental models* framework originally proposed by Johnson-Laird and colleagues (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Steedman, 1978). The mental models account of belief bias (Newstead et al., 1992; Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird, & Garnham, 1989) assumes three basic stages in the processing of syllogisms. First, subjects construct a mental representation that integrates the premises, the terms of which are described essentially as mental tokens. Second, subjects check to see whether a given conclusion is consistent with the model they have constructed. If the conclusion is not consistent, it is rejected; if the conclusion is consistent, then the subject considers its believability. If a conclusion is believable, it is accepted; if a conclusion is unbelievable, a third process is initiated, the goal of which is to construct alternative models of the premises. If the conclusion is consistent with all alternative models, it is accepted, else it is rejected. Thus, mental models theory proposes that responses result from a mixture of belief- and logic-based operations, rather than a single linear relation.

The role of believability in the mental models account is to bias the reasoning process itself, such that construction of alternative models only occurs for problems with unbelievable conclusions, and this manifests itself as a greater effect of logical status when conclusions are unbelievable. The theory groups problems according to the number of possible models of the premises they allow. Therefore, a clear prediction of mental models theory is that the Belief × Logic interaction will only occur for stimuli that allow the generation of alternative models (i.e., multiple-model problems) and will not depend on the type of validity (determinate or indeterminate) of the conclusion. Exactly this prediction was tested by Newstead et al. (1992, Experiment 3): The stimuli were single-model, indeterminately invalid problems, and no interaction was obtained, consistent with the mental models interpretation (but not with misinterpreted necessity).

While mental models theory is compelling, it is important to note that it was originally developed to explain data from the conclusion-production task, in which subjects must generate a valid conclusion from a set of premises. As such, it may not accurately characterize the conclusion-evaluation paradigm used by Evans et al. (1983) and many others. Conclusion-evaluation paradigms seem to require different processes and to inspire different biases. For instance, Morley et al. (2004) evaluated the hypothesis that conclusion production encourages forward reasoning (from premises to conclusion) whereas conclusion evaluation encourages backward reasoning (the conclusion biases construal of the premises). In a series of four experiments, Morley et al. demonstrated structural effects of figure (figural bias) in the absence of belief bias in a conclusion-production task, while the opposite result (belief bias in the absence of figural bias) obtained for the conclusion-evaluation task, consistent with their predictions. The authors suggested that a mental models account in which models of premises are constructed can still apply but that it would need to be modified to allow for effects of conclusion believability on the construction of those models.

Mental models theory also suffers from the fact that belief bias effects have been obtained with valid problems and one-model problems (Gilinsky & Judd, 1994; Klauer et al., 2000; Oakhill et al., 1989). Oakhill et al. (1989) addressed this issue by affixing an ad hoc conclusion-filtering mechanism to their version of the mental models framework. In other words, subjects may be processing syllogisms the way mental models predicts, but in cases where conclusions are unbelievable, subjects may still exhibit response biases that operate secondarily to filter (reject) such conclusions. Even if one were to maintain the conclusion filter, more recent findings from eye tracking (Ball et al., 2006) and response time (Thompson et al., 2003) experiments have converged on the notion that subjects actually spend more time processing believable than unbelievable conclusions, which is inconsistent with an account of the Belief × Logic interaction in which subjects generate more alternative models when conclusions are unbelievable.

Despite these limitations, the detailed characterization of the reasoning process provided by mental models theory has been incorporated into more recent accounts of belief bias. One example is *metacognitive uncertainty* (Quayle & Ball, 2000). This account is similar to misinterpreted necessity in that the belief effect is located in the response stage and is more prominent for invalid conclusions. It differs in that the factor responsible for the greater difficulty of invalid arguments is not the determinacy of the conclusions but an increased working memory load imposed by the larger number of alternatives required in their evaluation (Johnson-Laird, 1983). In other words, when the number of alternative models exceeds working memory capacity, subjects cannot reach a conclusion and must respond on the basis of believability. Consistent with this view, subjects with high working memory spans did not produce a Belief × Logic interaction, while those with lower spans did (Quayle & Ball, 2000). The eye-tracking data reported by Ball et al. (2006), however, are inconsistent with the notion that subjects generate more alternatives when conclusions are invalid.

A fourth account, *verbal reasoning theory* (VRT; Polk & Newell, 1995), bears many similarities to mental model theory, for example, the use of model representations, but replaces the search for alternative models with a process that linguistically encodes and reencodes the premises. Although the verbal reasoning account has been applied to experiments on the belief bias effect showing the interaction between belief and logic (Polk & Newell, 1995; Thompson et al., 2003), Polk and Newell (1995) did not focus on that aspect of the data. Still, their theory predicts an interaction in two different ways. In one variant, reasoners initially make judgments on arguments that are clearly valid or clearly invalid, without relying on prior beliefs. For intermediate arguments, which would tend to be invalid as well, reasoners eventually give up on the reencoding process and instead make a random choice between saying "invalid" and answering in terms of prior beliefs. The most pronounced effect of prior beliefs would be to incorrectly call invalid but believable arguments "valid"; hence, there would be lower accuracy for believable arguments than for unbelievable arguments. In the other variant of the theory, conclusions are self-generated rather than evaluated, and putative conclusions that are not believable are scrutinized more than believable conclusions, again leading to a Belief × Logic interaction.

VRT, which is highly similar to the mental models account, makes similar predictions about response times that are not always supported in the literature (Ball et al., 2006; Thompson et al., 2003). Specifically, the random decision explanation offered by VRT implies longer response times for invalid than valid problems regardless of believability status for conclusion evaluation, and the prediction of greater reencoding for unbelievable problems predicts longer response times for those items relative to believable problems for conclusion generation. Both of these predictions conflict with the response time results reported by Thompson et al. (2003), which led the authors to propose a *modified VRT* (MVRT). The new version adds the novel assumptions that (a) reasoners set a response deadline for the reencoding process that is the same for valid and invalid problems but are more likely to reach a conclusion in time when problems are valid, and (b) reasoners set a longer deadline for believable problems. Unfortunately, the literature does not provide a clear consensus on the issue of processing time differences across the four problem types (Ball et al., 2006; Thompson et al., 2003), an issue we return to in the General Discussion. More research is necessary to decide whether MVRT, which makes the same accuracy predictions as VRT and mental models, is also a more viable account of belief bias.

A Multinomial Processing Tree Model: Klauer, Musch, and Naumer (2000)

An important advancement in research on the belief bias effect was provided by Klauer et al. (2000). The analyses conducted by Klauer et al. were aimed at quantitatively separating the contributions of reasoning-stage effects (valid/invalid discrimination) and response bias effects (willingness to say "valid") to more clearly address the accounts by mental models, selective scrutiny, and misinterpreted necessity. To accomplish this goal, Klauer et al. developed a *multinomial processing tree* (MPT) model of the belief bias effect (see Figure 1) and used it as their primary analytical tool. The model has four processing trees, corresponding to each of the four critical argument types (valid or not, believable or not). Each tree has a reasoning-stage parameter r associated with it, which estimates the probability of correct validity detection for a given stimulus. For example, consider the first tree in Figure 1, which corresponds to those trials on which subjects are presented with valid, believable arguments. With probability r_{vb}, a "valid" detection state (D+) is entered, in which the subject always responds "valid." With probability $1 - r_{vb}$, a nondetection state (D−) is entered, and subjects are forced to guess. The guessing process may be influenced by response bias stemming from conclusion believability. With probability β_b, the subject guesses "valid," and with probability $1 - \beta_b$, the subject guesses "invalid." For invalid, believable problems, the process proceeds in the same manner, except that the detection of an invalid conclusion is governed by the r parameter (r_{ib}). The processing trees for the unbelievable problems are analogous but have different parameters to reflect the unbelievable nature of the conclusion. The MPT model thus allows for effects of believability on both the reasoning stage (the four r parameters), and the response stage (the two β parameters).

Estimating parameters in any model requires that there be at least as many data points as parameter values. In a typical belief bias experiment, there are four response rates (see Table 1) and, thus, too few data points to estimate the six parameters of the MPT model in Figure 1. The solution to this problem is either to increase the number of data points to be fit or to simplify the model by

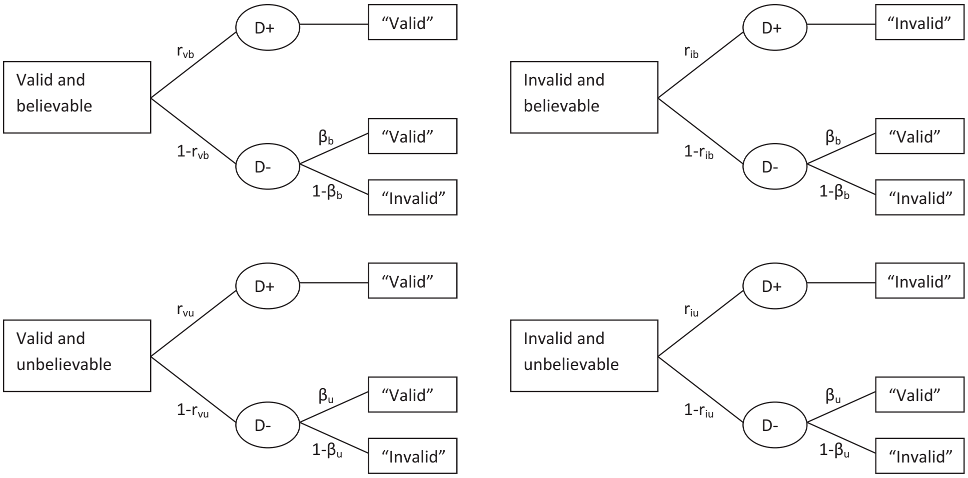

Figure 1. The multinomial processing tree model of belief bias proposed by Klauer, Musch, and Naumer (2000). The model was extended to analyze new data by multiplying each branch corresponding to the believability-defined response bias parameter β by an additional group-defined response bias parameter α_x, where x = low, medium, or high, corresponding to three conditions that differed in the perceived base rates of valid and invalid problems. D+ = detect state; D− = nondetect state; r = reasoning-stage parameter; $_b$ = believable; $_i$ = invalid; $_u$ = unbelievable; $_v$ = valid.

constraining some parameters to be equal. Klauer et al. (2000) adopted the strategy of increasing the number of data points by including a set of problems with neutral conclusions (neither believable nor unbelievable, such as "Some Stoics are not Sophists") and by manipulating across groups the perceived base rate of valid and invalid syllogisms. They also extended each processing tree with an extra response bias parameter, α_x, where x = low, medium, or high, corresponding to the three base-rate groups, and included two extra r parameters for neutral stimuli.

Their approach was to test the predictions of the various theories by setting different parameters equal and measuring the associated changes in model fit. For example, a test of selective scrutiny, which assumes subjects do not assess validity status when problems are believable, would imply that $r_{vb} = r_{ib} = 0$. In a series of eight experiments, Klauer et al. (2000) found that none of the theoretical accounts of belief bias was consistently supported but that there were substantial effects of believability on the reasoning stage, such that the null hypothesis $r_{vb} = r_{vu} = r_{ib} = r_{iu}$ could not be maintained. Particularly low values were observed for r_{vu} and r_{ib} relative to r_{vb} and r_{iu}. Additionally, constraining the guessing parameters to be the same for believable and unbelievable problems, $\beta_b = \beta_u$, consistently had little to no effect on the fit of the model. Together, these results indicate that the observed interaction between logic and belief is due primarily to effects of believability on the reasoning stage. Klauer et al. proposed a new theory of belief bias, the *selective processing* account, which, like other accounts before it, assumes a model-building process that is influenced by conclusion believability. Subjects are assumed to build a single mental model of the premises and to verify believable conclusions relative to that particular model. In the case of unbelievable syllogisms, however, the verification process involves negative hypothesis testing (Klayman & Ha, 1987), in which subjects attempt to integrate the logical negation of the conclusions. As this attempt is most likely to succeed when problems are invalid, the strategy will increase the rejection rate for unbelievable invalid problems relative to the other three problem types, thereby accounting for the observed Belief × Logic interaction.

The study by Klauer et al. (2000) has had a marked influence on subsequent research and has set a new standard for theoretical development in the belief bias literature. The study also provided a definitive statement of the idea that "the reasoning-based effect is what produces the belief by logic interaction noted by Evans et al. (1983) and later researchers," (Morley et al., 2004, p. 671). The Klauer et al. results also lend credence to the interpretation of the interaction by dual process theorists as an index of the contribution of analytical processes to syllogistic reasoning (Evans, 2006, 2007, 2008; Evans & Curtis-Holmes, 2005; but see Shynkaruk & Thompson, 2006). Less frequently discussed, however, are the important points the study raises about the measurement of sensitivity and response bias, which provided some of the motivation for Klauer et al.'s MPT model. Though several methods exist for teasing apart the contributions of sensitivity and bias, the MPT approach being one example, not all methods have been explored or compared in the belief bias literature. We discuss this issue more extensively in the next section.

Measuring Accuracy and Response Bias

As noted previously, one intuitive method of correcting for response bias effects is to simply subtract P("Valid"|Invalid) from P("Valid"|Valid), that is, corrected score = $H - F$, a method that has also been advocated in the recognition memory literature (Snodgrass & Corwin, 1988). This correction, carried out separately for believable and unbelievable problems, is typically used to compute a contrast of contrasts, the interaction index. The basic assumption implied by this statistic is that changes in response bias for a given problem type (e.g., believable problems) lead to equal increments in the acceptance rates for valid and invalid problems when accuracy is constant. An unbiased subject, by this definition, will make no errors ($F = 0$) and will produce correct responses, H, at a rate equivalent to their true sensitivity. Figure 2 provides a visual representation of the relationship between H and F that is implied by $H - F = k$, where k is a constant accuracy level.

Plots of H against F as a function of response bias at constant sensitivity are referred to as isosensitivity curves or ROCs. Figure 2 shows a theoretical ROC implied by the sensitivity statistic $H - F$. Empirical ROC curves, which to our knowledge have never been analyzed in the domain of syllogistic reasoning, have been used extensively by researchers in the areas of memory and perception (Green & Swets, 1966; Macmillan & Creelman, 2005; for a partial review, see Yonelinas & Parks, 2007) and have recently been collected for a variety of inductive reasoning tasks (Heit & Rotello, 2005, 2008, in press; Rotello & Heit, 2009). Empirical ROC data can be obtained in a number of ways, for example, by manipulating bias via payoff matrices, instructions, or the probability of a signal (e.g., valid item) trial (Healy & Jones, 1975; Van Zandt, 2000). In these methods, subjects make a binary (yes/no) response for each stimulus; the operating points are obtained across different experimental conditions or from independent groups of subjects. Alternatively, and more commonly, ROCs may be obtained by requiring subjects to follow their binary decisions (e.g., "valid" or "invalid") with an indication of their confidence in the response on a rating scale. The ROC in Figure 3A was generated from data reported on a 6-point confidence scale, in which a 1 corresponded to *sure valid* and a 6 corresponded to *sure invalid.* Because a rating of 1 corresponds to the most conservative bias (a tendency to avoid "valid" responses in the absence of strong evidence), both the hit and false-alarm rates are lower than at any other point on the function. An important property of ratings ROCs is that they are cumulative, that is, the (F, H) pair at 2 is the sum of hit and false-alarm proportions from Confidence Levels 1 and 2, the (F, H) pair at 3 is the sum of the proportions from 1 to 3, and so forth. The cumulative nature of the 6-point ROC results in a function with 5 points; the 6th point necessarily falls at (1, 1).[1] ROCs that fall higher in the space (toward the upper left) reflect better performance because, in that region, the proportion of correct responses (H) is high relative to the proportion of errors (F).

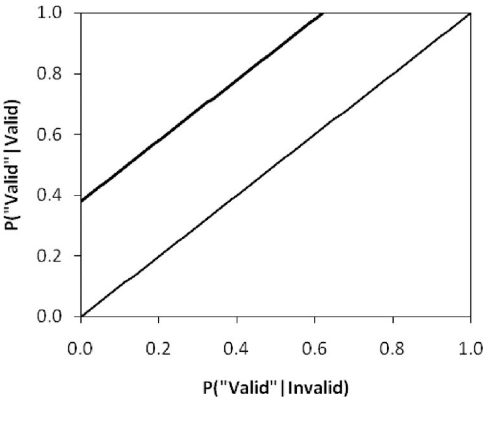

Figure 2. Receiver operating characteristic curve implied by the sensitivity statistic $H - F$.

Figure 2 shows that the use of $H - F$ as a measure of accuracy, as in the interaction index of belief bias, implicitly assumes a linear ROC, an assumption that to our knowledge has never been examined in syllogistic reasoning tasks. In the areas of memory and perception, however, the form of the empirical ROC (both ratings and binary response) is most often found to be curvilinear (Creelman & Donaldson, 1968; Egan, Schulman, & Greenberg, 1959; Emmerich, 1968; Macmillan & Creelman, 2005; Swets, 1986a, 1986b; Tanner, Haller, & Atkinson, 1967; but see Bröder & Schütz, 2009), and the inductive reasoning experiments reported by Heit and Rotello (2005, 2008, in press) and Rotello and Heit (2009) consistently produced curved ROCs similar in form to those observed in other domains.

If ROCs generated in the belief bias task are also curvilinear, what are the implications for conclusions reached in traditional analyses of acceptance rates and the interaction index? A series of simulations by Rotello, Masson, and Verde (2008) demonstrated that when the ROC is truly curvilinear and response bias differs across conditions, statistics based on contrasts of H and F are associated with substantially inflated Type I error rates. That is, even when two conditions differ only in response bias, contrasts of H and F are likely to imply that the conditions differ in accuracy. This problem was exaggerated with the inclusion of additional subjects or more trials per subject and was also increased with the size of the response bias difference across groups. The inflated Type I error rate is a direct result of the fact that $H - F$ is not independent of response bias when increases in response bias do not augment H and F by equal amounts (Macmillan & Creelman, 2005). Thus, changes in response bias caused by, for example, the believability of syllogistic arguments could lead to differences in $H - F$ even if true valid/invalid discrimination is constant across the two problem types (Klauer et al., 2000, made a similar point with respect to the psychometric properties of proportions). To determine whether there is cause for concern, ROCs must be collected in a belief bias task; that was the goal of Experiments 2 and 3.

Without recourse to ROCs, one might assume that estimates of accuracy in the belief bias task are unaffected by response bias differences, as Klauer et al. (2000) failed to demonstrate effects of believability on the response stage in their MPT analysis. However, their MPT model was based on "a simple threshold model" (Klauer et al., 2000, p. 856), and the predictions made by all pure high-threshold models for the effects of bias changes at constant accuracy are the same as those implied by $H - F$: linear ROCs.[2]

[1] The analysis of ROC data is based on a maximum likelihood statistic that uses only the unique, not cumulative, responses at each confidence level; it also takes account of the fact that both H and F are dependent variables measured with noise.

[2] For confidence-rating-based ROCs, this prediction assumes a particular mapping between internal states and response ratings, such that "detect" states necessarily lead to highest confidence responses and "nondetect" states to lower confidence responses. Other state–response mappings are possible that, in effect, make MPT models behave more like signal detection models; these mappings lead to piecewise linear ROCs that are difficult to discriminate from those predicted by signal detection theory (e.g., Bröder & Schütz, 2009; Krantz, 1969; Malmberg, 2002). We revisit this issue in the modeling and discussion sections and in our analysis of Experiment 3.

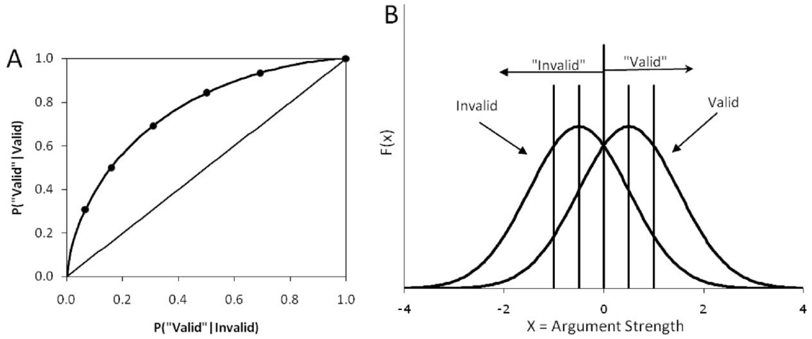

Figure 3. A: Receiver operating characteristic curve generated by the equal-variance signal detection theory model. B: Representation corresponding to Panel A.

In fact, it was the failure of precisely this prediction that led to the abandonment of threshold models by many memory and perception researchers in favor of models based on *signal detection theory* (SDT; Banks, 1970; Green & Swets, 1966; Macmillan & Creelman, 2005). It is to this framework that we now turn.

A Signal Detection Model of Syllogistic Reasoning

In the present study, we evaluate an unequal-variance signal detection model of syllogistic reasoning. SDT, widely used in the study of recognition memory, can be readily extended to belief bias experiments. The conclusion-evaluation task employed by Evans et al. (1983) and others yields four response proportions (see Table 1). If a given stimulus is actually a valid argument, the subject's response is either a hit (a "valid" response) or a miss (an "invalid" response). If a test stimulus is actually an invalid argument, the subject's response is either a correct rejection (an "invalid" response) or a false alarm (a "valid" response).

The signal detection model posits that inferential decisions reflect operations on a single, continuous argument-strength dimension (see Figure 3B). We assume that the argument strengths of invalid stimuli are normally distributed with a mean μ_i and standard deviation σ_i and that the strengths of valid stimuli are normally distributed with mean μ_v and standard deviation σ_v, where $\mu_v > \mu_i$. Subjects' ability to distinguish between valid and invalid stimuli is reflected as increased mean argument strength for valid arguments. In SDT models, discrimination accuracy or sensitivity can be measured in several ways. By far the most commonly used statistic is d', which measures the distance between the means of two distributions (here, for the valid and invalid items) in units of their common standard deviation:

$$d' = z(H) - z(F). \tag{2}$$

If (and only if) the two distributions of evidence strength are normally distributed and have equal variance, d' is independent of response bias.

Response bias (willingness to say "valid") can be measured in a number of ways (see Macmillan & Creelman, 2005, for discussion), but the more common methods are all related by the criterion placement parameter. Criterion placement, *c,* reflects bias relative to the zero-bias point where the valid and invalid distributions intersect. Liberal biases (maximizing hits at the cost of increasing false alarms) produce negative values of *c,* while conservative biases (minimizing false alarms at the cost of a reduced hit rate) produce positive values of *c*.

$$c = -.5[z(H) + z(F)]. \tag{3}$$

In Figure 3B, the area under the valid distribution to the right of the criterion corresponds to the hit rate (*H*), while the area under the invalid distribution to the right of the criterion corresponds to the false-alarm rate (*F*). It is assumed that several criterion positions can be produced by subjects in the reasoning task, thus generating different partitions of the strength axis as illustrated in the figure. The degree of overlap between the distributions is an indication of sensitivity; the greater this area is relative to either distribution, the lower the overall sensitivity, regardless of criterion placement. The areas under the valid and invalid distributions to the left of the criterion correspond to misses (*M*) and correct rejections (*CR*), respectively.

The ROC implied by this signal detection model can be generated by varying response bias in a continuous fashion while holding sensitivity constant. Figure 3A shows that the model yields a curvilinear ROC that is symmetrical about the minor diagonal. Response bias is reflected in the points on the ROC: More conservative biases generate operating points on the left end of the ROC (with low values of both *H* and *F*), and more liberal biases result in operating points on the right (with higher values of both *H* and *F*). Thus, points on the same ROC reflect equal sensitivity but different levels of response bias.

If the ROC is transformed to normal–normal space, the result is a *z*ROC, the slope of which estimates the ratio of the invalid and valid argument standard deviations (σ_i/σ_v). Therefore, the slope of the *z*ROC can be used to make inferences about the variances of strength distributions; in particular, it allows evaluation of the assumption that the variances are equal. In recognition experiments, *z*ROC slope is often less than one, reflecting greater variability in the strengths of studied items relative to lures (Glanzer, Kim, Hilford, & Adams, 1999; Heathcote, 2003; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992). In reasoning tasks, *z*ROC slope is apparently also less than one, reflecting greater variability in the strengths of valid or strong arguments than of invalid or weak arguments (Heit & Rotello, 2005, 2008, in press; Rotello & Heit, 2009). Thus, for both recognition and reasoning tasks, the underlying distributions have unequal variance, which violates an assumption of the measurement statistic d'.

A major concern that applies regardless of the particular tasks or dependent measures one chooses is whether the assumptions of a given model have been met. For example, choosing d' as a measure of sensitivity entails the assumption that the data are equal variance and Gaussian in form. If the slope of the *z*ROC is not one or if the ROC is linear, then that assumption is violated. Similarly, choosing $H - F$ as a sensitivity statistic entails the assumption that the ROC is linear and symmetric, a form that is consistent with underlying equal-variance evidence distributions that are rectangular in form. Empirical ROCs that are curvilinear or asymmetric reflect a violation of that assumption. Violations of this type can dramatically elevate the risk of committing a Type I error (Rotello et al., 2008): Differences in response bias across experimental conditions are easily (and frequently) misinterpreted as differences in accuracy.

In the typical case where the empirical ROCs are curvilinear and asymmetric, indicating that the equal-variance assumption has been violated, one may choose to measure the distance between the distributions (sensitivity) in units of either the stronger (d'_2) or weaker (d'_1) distribution. Alternatively, the measure d_a may be substituted for d', and the corresponding bias measure c_a may be substituted for c. These statistics set the standard deviation of the lure or invalid distribution at 1 (without loss of generality) and the standard deviation of the target or valid distribution at s and measure distances in terms of the root-mean-square standard deviation (for derivation, see Macmillan & Creelman, 2005):

$$d_a = \sqrt{\frac{2}{1+s^2}}\,[z(H) - sz(F)] = \sqrt{\frac{2}{1+s^2}}\,d_2'. \qquad (4)$$

$$c_a = \frac{-\sqrt{2s}}{\sqrt{1+s^2}(1+s)}\,[z(H) + z(F)] \qquad (5)$$

An alternative method for measuring accuracy is simply to estimate the area under the ROC using A_z, which ranges from 0.5 for chance performance to 1.0 for perfect accuracy (Macmillan, Rotello, & Miller, 2004; Swets, 1986a, 1986b):

$$A_z = \Phi\left(\frac{d_a}{\sqrt{2}}\right) \qquad (6)$$

This method is preferable to d_a for two reasons. First, Green and Swets (1966) showed that when performance is unbiased, A_z equals proportion correct in the two-alternative forced-choice task. Second, in a series of simulations comparing A_z and d_a, Macmillan et al. (2004) showed that A_z is a relatively unbiased measure of accuracy and has a smaller standard error than d_a.

Our experiments provide the first ROC data in the syllogistic reasoning task, but ROCs from other reasoning tasks have been curvilinear and asymmetric (Heit & Rotello, 2005, 2008, in press; Rotello & Heit, 2009). Those data suggest two key points: (a) The use of $H - F$ as a measure of reasoning performance in the belief bias task is probably inappropriate, and (b) application of d' would confound sensitivity with response bias. Thus, we adopt the unequal-variance framework in applying the signal detection model to data from the syllogistic reasoning task.

Overview of Current Study

The goal of the present experiments was to evaluate the assumption of ROC linearity in syllogistic reasoning, using novel analytical techniques and the application of quantitative models. If the assumption of linear ROCs is not supported empirically, it is important to know how this violation of assumptions impacts the meaningfulness of the interaction index that is often reported in the reasoning literature. If the interaction index is inappropriate for the data, then theories of belief bias that follow from results obtained with that measure are built on a faulty foundation. To investigate this possibility, we collected ROC data for the conclusion-evaluation task, using both the confidence-rating procedure described previously (Experiments 1–3) and the base-rate method (Experiment 3). We examined the effects of perceived base rates of valid and invalid items (Experiment 1), actual base rates of valid and invalid items (Experiment 3), and conclusion believability (Experiments 2–3). Thus, two experiments included a factor expected to influence response bias but not reasoning accuracy, and two included a belief bias manipulation that might affect either. Across conditions in each study, we compared the results from the interaction index (which relies on $H - F$) with the results from A_z, an SDT-based measure of the area under the ROC.

We fit the MPT model of Klauer et al. (2000) to our base-rate ROCs, comparing its predictions and explanatory power with those of our signal detection model. We also extended Klauer et al.'s model to fit our confidence-rating ROCs. Although we tested only the theoretically motivated SDT model that we have proposed, we included several alternative instantiations of the high-threshold model for the rating ROCs. In all cases, our MPT extensions necessarily produced linear ROCs, as our primary question of interest was whether the ROCs are linear and thus whether use of the $H - F$ interaction contrast is justified. Finally, following Klauer et al.'s example, we conducted specific hypothesis tests for each of the models to determine whether conclusion believability affects accuracy, as assumed in all the theoretical accounts of the belief bias effect, or response bias.

To preview our results, in each experiment, we demonstrated that the assumption of a linear ROC implied by the interaction index is unwarranted. Thus, the highly replicable interaction between validity and believability appears to be a consequence of using an inappropriate measurement statistic ($H - F$) in the face of response bias differences across experimental conditions (i.e., believability). Our data and analyses therefore challenge all existing theories of belief bias because they are built on a flawed measurement statistic: All existing theories are tacitly predicated on the assumption of a linear ROC that is not observed empirically. In addition, our modeling work reveals that the best description of syllogistic reasoning ROCs is provided by the SDT model. Our model suggests a new and more parsimonious account of the locus of the belief bias effect.

Experiment 1

The aim of the present experiment was to assess the form of the empirical ROC in a syllogistic reasoning task. In addition, we manipulated response bias across two conditions so that we could assess the impact of those bias changes on two different accuracy measures, $H - F$ and A_z. Because base-rate manipulations have been shown to affect response bias, but not sensitivity, in perception, memory, and reasoning experiments (Klauer et al., 2000; Rhodes & Jacoby, 2007; Tanner, Rauk, & Atkinson, 1970), we considered a perceived base-rate manipulation to be ideal for testing the robustness of $H - F$ against changes in response bias only. All subjects evaluated a set of abstract syllogisms, half of which were valid and half of which were invalid. In the liberal condition, subjects were instructed that 85% of the problems were valid and 15% were invalid; in the conservative condition, the instructions implied the converse. In both conditions, the actual base rate of valid problems was 50%. Thus, intergroup differences in $H - F$ (i.e., a Base Rate × Validity interaction) could imply either a true (but unexpected) sensitivity effect or a Type I error (Rotello et al., 2008). ROC data allow us to distinguish these alternatives because the area under the ROC provides a theoretically neutral assessment of performance (Swets, 1986a, 1986b). Therefore, we asked subjects to follow each of their validity decisions with a confidence rating; these ratings allowed the construction of ROC curves for each condition.

We also fit and compared the signal detection and MPT models. The purpose of fitting the signal detection model was to assess an account that does not assume linear ROCs and has the potential to provide measures of belief bias that do not depend on that assumption. The purpose of fitting MPT models was to assess variants of this important class of models that assume linear ROCs and to examine what these models would infer about the nature of belief bias. In Experiments 1 and 2, fits of the MPT model were accomplished by extending it to handle confidence ratings. There are several ways in which the states of the MPT model can be mapped onto confidence ratings (e.g., Bröder & Schütz, 2009; Malmberg, 2002); we chose two versions that we considered to be the most faithful to the original model of Klauer et al. (2000). In Experiment 3, we repeated this analytic strategy, but we also fit the published version of the MPT model, using the same method as Klauer et al.

Method

Subjects. Seventy-one undergraduates at the University of California, Merced, participated; they received $5.00 for their participation.

Design. Experiment 1 used a 2 (logical status: valid or invalid) × 2 (perceived base rate: 85% or 15% valid) mixed design. All subjects were asked to evaluate the validity of 32 syllogisms, half of which were actually valid and half of which were actually invalid. Subjects were randomly assigned either to the conservative group ($n = 34$), in which they were told that 85% of the problems were invalid, or to the liberal group ($n = 37$), in which they were told that 15% of the problems were invalid (see Procedure for details).

Stimuli. Sixteen syllogistic problem frames were used (see Appendix D): Eight problem frames controlled for atmosphere (Begg & Denny, 1969), conversion (Dickstein, 1975, 1981; Revlin, Leirer, Yopp, & Yopp, 1980), and figural effects (Dickstein, 1978), and eight controlled for atmosphere and figure, but not conversion. Although the design of this experiment did not require control of these structural effects, this decision allowed use of exactly the same problem frames in Experiments 2 and 3, where conclusion believability was manipulated. Half of the eight problems in each set were valid, and half were invalid. To increase the power of the design, each subject was presented with two versions of each problem, for a total of 32 problems; the repeated problem frames differed in their abstract content. Abstract content was generated by choosing 24 letters of the alphabet (every letter except A and M) and randomly assigning the letters to the positions of the predicate, middle, and subject terms (i.e., X, Y, and Z in Example B). The assignment of content to structures was subject to three constraints. First, no letter appeared more than twice across the 32 problems. Second, no two letters appeared together in more than one problem. Third, no letter occupied more than one position in a given problem. Two different sets of 32 problems were constructed by this method, each of which was presented to approximately half of the subjects in each condition.

There were also five practice problems presented to each group of subjects; these abstract problems did not share any of the same problem frames as the 32 critical problems. The exact proportions of valid and invalid practice syllogisms were varied to roughly parallel the implied base rates in each experimental condition: The conservative group practiced with four invalid problems and one valid problem, and the liberal group practiced with four valid problems and one invalid problem.

Procedure. All subjects were tested individually and were seated approximately 2 ft in front of a computer monitor. The instructions for the present task were similar to the instructions used in other syllogistic reasoning studies (Klauer et al., 2000; Newstead et al., 1992) but contained additional passages relating to the base-rate manipulation. Specifically, subjects were told:

> In this experiment, we are interested in people's reasoning.
>
> For each question, you will be given some information that you should assume to be true. This will appear ABOVE a line. Then you will be asked about a conclusion sentence BELOW the line. If you judge that the conclusion logically follows from the statements, you should answer "Valid", otherwise you should answer "Not Valid". Next, you will be asked how confident you are in this judgment.
>
> You should just answer each question as best as you can, based on the information available.
>
> Please ask the experimenter if you have any questions.
>
> An important thing to remember is that 85% of the problems in this experiment are actually [valid/not valid], and only 15% of them are [invalid/valid]. So, if you need to guess whether a problem is valid or not, you should always guess ["Valid"/"Not Valid"].

Following the instructions, subjects advanced through the five practice trials and then the 32 critical trials. Stimuli within each phase of the experiment were randomized for each subject. Subjects made validity decisions via keypress (J = "valid"; F = "invalid"). After each "valid"/"invalid" response, subjects were asked to rate their confidence on a scale of 1 to 3 (1 = *Not at all confident,* 2 = *Moderately confident,* 3 = *Very confident*). Because the ratings ranged from low to high confidence for "valid" and "invalid" responses, there was a total of six possible response categories. We subsequently recoded the responses with the numbers 1–6, where 1 reflects a high-confidence "valid" judgment, 3 reflects a low-confidence "valid" judgment, 4 reflects a low-confidence "invalid" judgment, and 6 reflects a high-confidence "invalid" judgment.

Results

The proportion of conclusions accepted (summarized in Table 2) was first analyzed using methods similar to those used in the

Table 2
Hit Rate (H), False-Alarm Rate (F), and Contrast Results for Experiments 1–2

Experiment	Condition	H = P (“Valid”|Valid)	F = P (“Valid”|Invalid)	$H - F$
1	Liberal	.79	.67	.12
	Conservative	.55	.31	.24
2	Believable	.86	.61	.25
	Unbelievable	.68	.32	.36

reasoning literature. A 2 × 2 mixed analysis of variance (ANOVA) was conducted with validity as a within-subjects factor and perceived base rate (low vs. high) as a between-subjects factor. The ANOVA confirmed that subjects accepted more valid than invalid conclusions ($H - F > 0$), $F(1, 69) = 69.964$, $MSE = .017$, $p < .001$, $\eta^2 = .503$. Our bias manipulation also had the intended effect: Subjects' overall acceptance rates were higher in the liberal than in the conservative condition, $F(1, 69) = 57.353$, $MSE = .056$, $p < .001$, $\eta^2 = .454$. Finally, these two factors interacted: $H - F$ was greater for subjects in the conservative condition, $F(1, 69) = 7.713$, $MSE = .017$, $p < .01$, $\eta^2 = .101$, implying that accuracy was higher for that group than for the liberal group.

ROCs for the conservative and liberal conditions are plotted in Figure 4A. As can be seen in the figure, the operating points for the liberal group (circles) are displaced upward and rightward relative to the corresponding points for the conservative group (squares), indicating a main effect of condition on response bias that is consistent with the ANOVA results. The height of the two ROCs in the space is similar, however, suggesting that there is little difference in accuracy across the groups. We confirmed this observation using Metz's ROCKIT software (Metz, 1998) to compare A_z for the two conditions: There was no significant difference in accuracy ($z = 0.50$, $p = .60$).

Although the interaction between condition and validity that was observed in the acceptance rates ($H - F$) appears to contradict the conclusion that there is no difference in accuracy (A_z) between the two conditions, this apparent discrepancy is readily explained by the forms of the implied and empirical ROCs for the two conditions. As noted previously, ROCs implied by the accuracy measure $H - F$ are linear with unit slope. The *y*-intercept is obtained by subtracting F from H for each group and equals .24 and .12 for the conservative and liberal groups, respectively (see Table 2). The corresponding implied ROCs are superimposed on the observed ROCs in Figure 4A. Two points are evident in the figure. First, the intercept for the conservative condition is higher than that for the liberal condition, consistent with the conclusion based on acceptance rates that subjects in the conservative

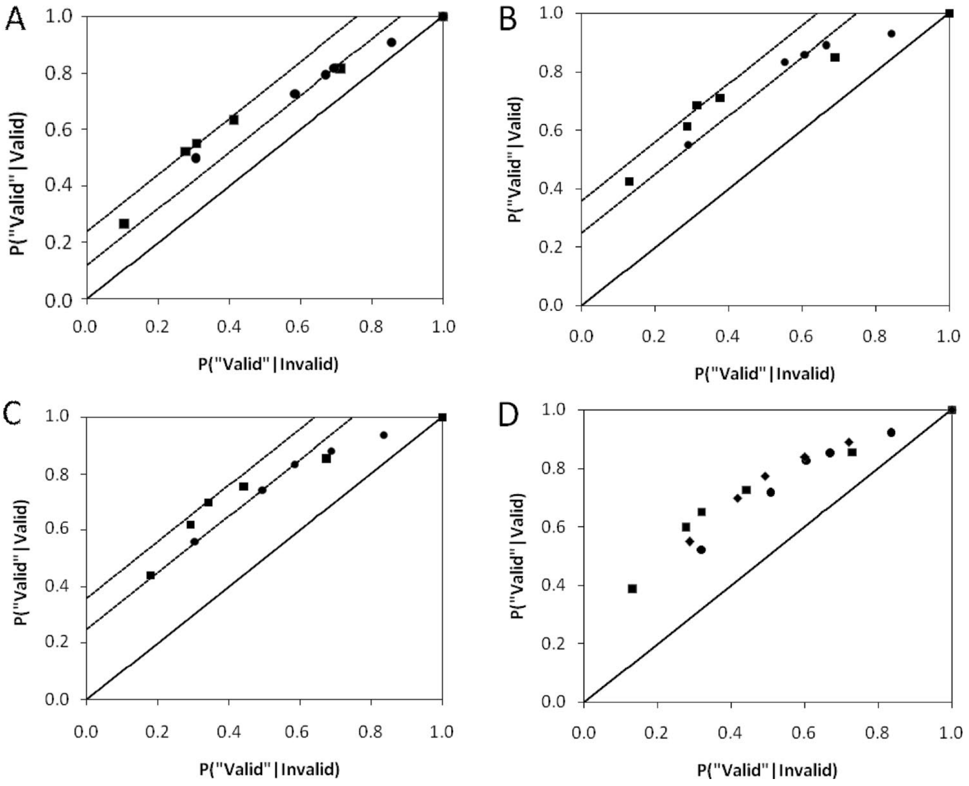

Figure 4. Empirical receiver operating characteristics (ROCs) for Experiments 1–3. ROCs implied by $H - F$ are superimposed on the observed data (dashed lines). A: ROCs for the conservative (squares) and liberal (circles) conditions of Experiment 1. B: ROCs for unbelievable (squares) and believable (circles) problems, Experiment 2. C: ROCs for unbelievable (squares) and believable (circles) problems, Experiment 3. D: ROCs for the conservative (squares), liberal (circles), and neutral (diamonds) conditions of Experiment 3.

condition were better able to detect the difference between valid and invalid problems. Second, the ROCs implied by $H - F$ do not fit the data well, suggesting that this accuracy measure is not appropriate for the data.

Models fit. We fit three different models to these data: two MPT models (MPT1 and MPT2) and our signal detection model. MPT fits were accomplished by extending Klauer et al.'s (2000) model to allow confidence ratings. (See Malmberg, 2002, for related efforts to fit rating data with multinomial models.) In the most basic, eight parameter version of the model depicted in Figure 5, we assumed that high-confidence responses follow directly from entry into the detect state (D+), yielding a "1" response for valid detection and a "6" response for invalid detection. The probability of a high-confidence error—(P("1"|Invalid) or P("6"|Valid)—was set to an arbitrarily small, nonzero, value (.0000001). This version of the model, which we call MPT1, follows Klauer et al. in making the strong assumption that subjects virtually never produce high-confidence errors. We assumed that lower confidence responses, Ratings 2–5, are made only from nondetect states. When the subject has guessed, with probability β, that the item is valid, then the response "2" is given with probability α_2 and the response "3" with probability $1 - \alpha_2$. When the item is guessed to be invalid, then the response "4" is given with probability α_4 and the response "5" with probability $1 - \alpha_4$. This model requires two reasoning parameters (r_v, r_i) and one response bias parameter (β) for each condition, as well as two parameters that map internal states onto confidence ratings (α_2, α_4).

According to this strict version of the model, entry into State D+ always leads to a high-confidence, correct response; invalid items essentially never result in a high-confidence error. As noted by Klauer et al. (2000), it may be inappropriate to assume that reasoners never make errors "which they then hold with high confidence" (p. 857), in which case the model must be modified. In discussing this issue, the authors wrote, "False positive decisions are, in fact, an important topic of the present theoretical and empirical argument. The new theory of belief bias developed later allows one to specify the conditions under which they arise frequently" (Klauer et al., 2000, pp. 857–858). The conditions in question appear to necessitate the use of one-model problems, which were not used in the present experiments. For this reason, the MPT extension in Figure 5 is the ratings version that is most faithful to Klauer et al.'s description. Nonetheless, high-confidence errors were observed in all of the previous studies reported by Heit and Rotello (2005, 2008, in press) and Rotello and Heit (2009), and the same is clearly true in Experiment 1 (see Figure 4A). To account for high-confidence errors, we also implemented a different mapping of internal states to confidence ratings, MPT2. Rather than assuming high-confidence errors occurred with probability .0000001, in MPT2 that probability was a free parameter, ε. Those high-confidence errors are assumed to arise from decision noise (e.g., random responding) or other non-reasoning-based processes.

As suggested by a reviewer, we also fit an MPT model (MPT-R) in which any confidence rating (R), including "sure valid," could be given from the uncertain state (D−). This model is capable of producing ROCs that approach the curved functions generated by SDT models, and thus, this analysis does not speak at all to the measurement issue under examination in this study. The conclusions reached with MPT-R agreed substantially with those of our other MPT models, and this model consistently failed to outperform the SDT model. Thus, we do not discuss this model any

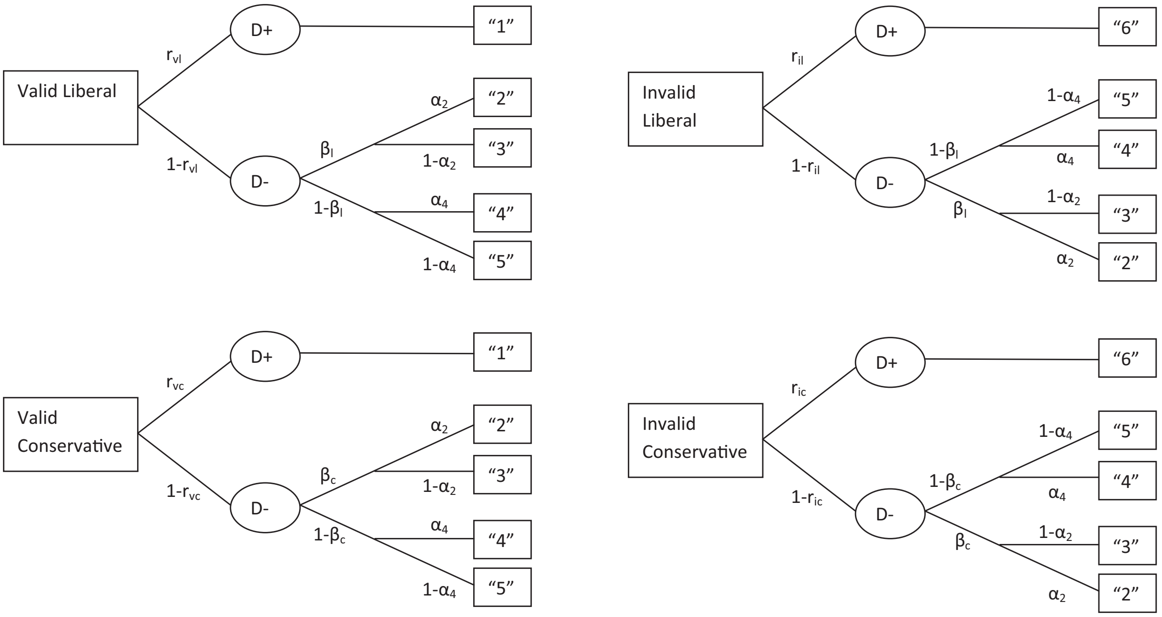

Figure 5. MPT1: the multinomial processing tree (MPT) model proposed by Klauer, Musch, and Naumer (2000), extended to the ratings design of Experiment 1. Slightly modified versions of this basic framework were applied to data from Experiments 1–3, as described in the text. In MPT1, the probability of a high-confidence error—P("1"|Invalid) or P("6"|Valid)—was set to an arbitrarily low value (.0000001). D+ = detect state; D− = nondetect state; r = reasoning-stage parameter; $_c$ = conservative; $_i$ = invalid; $_l$ = liberal; $_v$ = valid.

further, though we have reported the statistics for all MPT-R fits in Appendix B. The equations for the MPT1, MPT2, and MPT-R models are provided in Appendix A.

Finally, we also fit an unequal-variance signal detection model in which responses were presumed to be made along a single dimension of argument strength. To fit each condition, this model requires two reasoning parameters (μ_v, σ_v) and five decision criteria (c_1–c_5). Equations are provided in Appendix A.

Best fitting parameter values were estimated for each model by using Excel's Solver routine to minimize G^2 (Dodson, Prinzmetal, & Shimamura, 1998). Model selection was accomplished by comparing values of the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978) for each condition (both conditions lead to the same conclusion). These statistics combine a goodness-of-fit statistic, ln(*L*), the log-likelihood of the data given the model, with a penalty for the number of free parameters (*k*):

$$\mathrm{AIC} = -2\ln(L) + 2k, \tag{7}$$

$$\mathrm{BIC} = -2\ln(L) + k\ln(n). \tag{8}$$

BIC weights the parameters as a function of the sample size, *n*. The resulting values of both AIC and BIC are smaller for more parsimonious models, assuming equivalent goodness of fit (Myung, Pitt, & Kim, 2003). The parameter penalty in BIC is more severe; thus, both measures are usually reported.

Modeling results. Consistent with the visually poor fit of the ROCs implied by the measure $H - F$, shown in Figure 4A, the fit of the simple MPT1 model to the ROCs was very poor in all comparisons. As such, we report only the augmented version of the MPT model (MPT2), which allowed for the possibility of high-confidence errors. Fit statistics for MPT1 can be found in Appendix B. MPT2 provided a much better fit to the ROCs than did MPT1, but a very similar pattern of parameter values was obtained.

The best fitting parameter values for each model are shown in Table 3, and several fit statistics are presented in Table 4. All of the fit statistics indicate that the SDT model provides the better fit to the data. The ROCs generated with the best fitting parameter values of each model are shown in Figure 6, separately for the liberal and conservative conditions. As is clear in the figure, MPT2 generates ROCs that fail to adequately capture the observed functions: The shapes of the ROCs are too linear, and the predicted operating points (shown as crosses) do not correspond well to the observed points (shown as circles). In contrast, the shape of the SDT-generated ROC is closer to that of the data, and each predicted operating point falls near the corresponding observed point in the data. The fact that the SDT model provides a better account of the data confirms that the ROCs are not linear, supporting our claim that A_z is a better measure of accuracy than $H - F$ for these results.

Table 3
Best Fitting Parameter Values for MPT2 and SDT, Experiment 1

Model	Parameter	Liberal condition	Conservative condition
MPT2	r_v	.54	.32
	r_i	.21	.32
	β	.68	.42
	α_2	.82	
	α_4	.24	
	ϵ	.17	
SDT	μ_v	0.50	0.56
	σ_v	1.16	1.21
	c_1	0.51	1.30
	c_2	−0.21	0.55
	c_3	−0.44	0.46
	c_4	−0.53	0.19
	c_5	−1.05	−0.54
	d_a	0.46	0.50

Note. In the SDT, d_a is computed from μ_v and σ_v (see Equation 4 in the text) and is not an extra free parameter. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model.

Table 4
Fit Statistics for MPT2 and SDT in Experiment 1

	MPT2			SDT		
Condition	AIC	BIC	$G^2_{4\mathrm{df}}$	AIC	BIC	$G^2_{3\mathrm{df}}$
Liberal	3,661.65	3,962.11	106.24	3,559.23	3,594.77	1.83
Conservative	3,579.69	3,609.64	56.67	3,530.44	3,565.38	5.42

Note. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model; AIC = Akaike information criterion; BIC = Bayesian information criterion.

The parameter values of the models (see Table 3) make clear that both models find a substantial effect of condition on response bias: The β parameter of MPT2 is larger for the liberal condition than the conservative condition, reflecting an increased tendency to guess "valid," and there are more negatively valued (i.e., liberally placed) criteria in the liberal condition according to SDT, reflecting greater willingness to say "valid." To evaluate the significance of these response bias differences, the bias parameters for each model were constrained to be equal across conditions, and G^2 values for these restricted models were computed. If the restriction significantly reduces the fit of the model, then we can infer that different bias parameters are needed. The restricted version of each model is nested within the corresponding full model, and the difference in the fit ($G^2_{\mathrm{restricted}} - G^2_{\mathrm{full}}$) is distributed as a chi-square with degrees of freedom (*df*) equal to the difference in the number of free parameters. For MPT2, the restricted model assumes $\beta_l = \beta_c$, and the full model does not, for a difference of one parameter and thus one degree of freedom. For SDT, the restricted model assumes a single set of criteria ($c_1 \ldots c_5$), whereas the full model estimates separate criteria for each condition, for a difference of five degrees of freedom. The chi-square test indicated the restriction of equal response bias parameters significantly reduced the fit of both models (MPT2: $\Delta G^2_{1\mathrm{df}} = 84.508$, $p < .001$; SDT: $\Delta G^2_{5\mathrm{df}} = 188.402$, $p < .001$). Therefore, both models conclude that response bias differs between conditions, an unsurprising result given the design of the experiment.

A more interesting question asks what the models conclude about the reasoning process: Do both models predict equal accuracy across conditions, consistent with the empirical ROCs? It can be seen that, in fact, the MPT2 and SDT models differ in their assessment of the degree to which accuracy differs by condition. Table 3 suggests that there is an interaction between condition and validity in the reasoning-stage parameters of MPT2, with $r_{vl} > r_{vc}$ and $r_{il} < r_{ic}$. As can be seen in Figure 6, this amounts to differ-

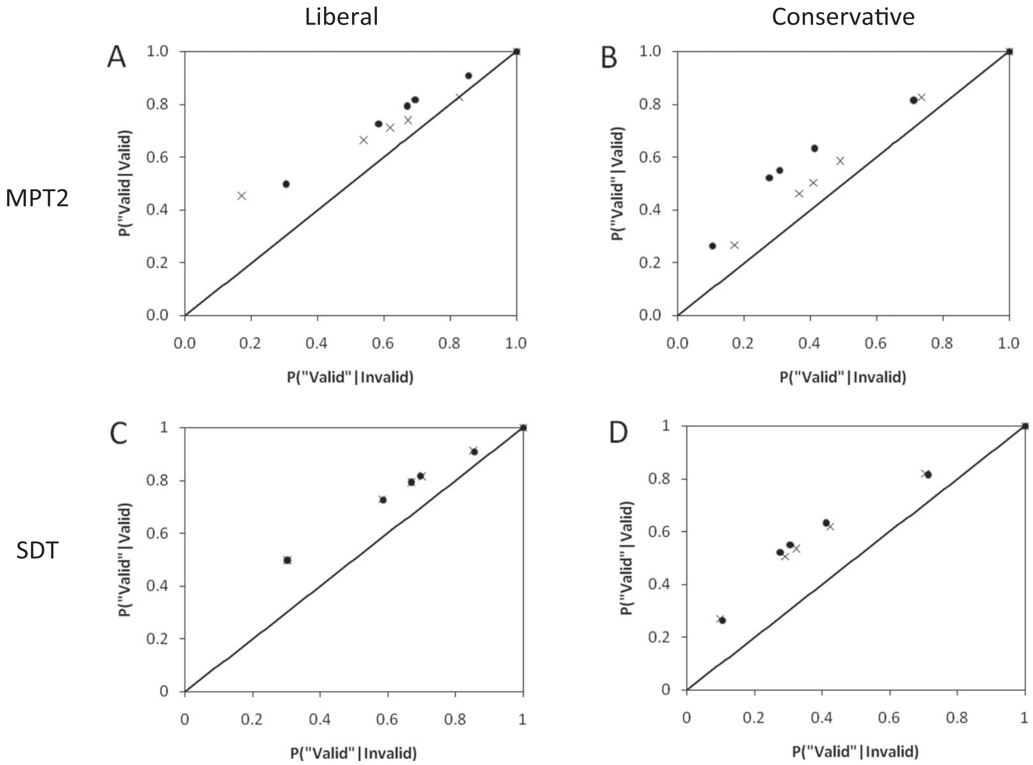

Figure 6. A: Observed (circles) and MPT2-predicted (crosses) receiver operating characteristics (ROCs) for the liberal condition of Experiment 1. B: Observed and MPT2-predicted ROCs for the conservative condition. C: Observed (circles) and SDT-predicted (crosses) ROCs for the liberal condition of Experiment 1. D: Observed and SDT-predicted ROCs for the conservative condition. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model.

ences in ROC height and slope across the conditions. To confirm this pattern, the *r* parameters were constrained to be equal across the groups, at each level of validity (i.e., the constraint $r_{vl} = r_{vc}$, $r_{il} = r_{ic}$, was imposed), thus freeing two degrees of freedom for the chi-square test. This restriction resulted in a significant reduction in fit for MPT2 ($\Delta G^2_{2\mathrm{df}} = 62.939$, $p < .001$). An analogous test of the reasoning stage in the SDT model implies the restriction $d_{a,\,c} = d_{a,\,l}$. In Equation 4, it can be seen that the sensitivity parameter d_a depends on both d'_2 (μ_v/σ_v) and s ($1/\sigma_v$). Thus, d_a is calculated from the free parameters of the model, rather than serving as an additional parameter, and imposing the restriction frees two parameters for the chi-square test. The restriction of equal sensitivity and slope for the liberal and conservative groups had little effect on the fit of the SDT model ($\Delta G^2_{2\mathrm{df}} = 0.42$, $p > .750$).[3] Thus, only SDT infers that subjects in the two groups perform at similar accuracy levels,[4] consistent with our expectations: Subjects were randomly assigned to experimental conditions, and the perceived base-rate manipulation was expected to affect response bias but not accuracy (Klauer et al., 2000).

Discussion. The results from Experiment 1 indicated our manipulation of perceived base rate had the desired effect on response bias, though the presence of an apparent accuracy effect was shown to depend on whether or not the ROCs were assumed to be linear. A test of ROC area (A_z) indicated, contrary to the results in $H - F$ (see Table 2), that there was no difference in accuracy across the two conditions. The MPT2 model indicated effects of condition on both the reasoning and response stages of the model, consistent with the results in $H - F$, whereas the SDT model indicated effects on response bias only, consistent with A_z. The SDT model provided the best fit to the data, confirming what is apparent in Figure 4: The ROCs are not linear, and there is no difference in accuracy across the groups. We conclude that application of the measure $H - F$ produced a Type I error in the analyses of the acceptance rates shown in Table 2, consistent with the predictions of Rotello et al. (2008).

[3] We also fit an equal-variance version of the SDT model in which the distributions of valid and invalid problems were assumed to have the same variance; this model fit significantly worse than the model with free variance ($\Delta G^2_{2\mathrm{df}} = 13.14$, $p < .01$).

[4] We also considered whether artifacts from averaging over subjects could account for the success of the SDT model. Although we did not have enough data to model individual-subject ROCs, we did fit the ROCs for a group of 15 high- and 15 medium-accuracy subjects (chance-level ROCs are necessarily linear, rendering low-accuracy groups less informative). For both the medium- and high-performing groups, the SDT model fit significantly better than the MPT model (SDT $G^2_{3\mathrm{df}} = 1.37$ and 4.01, respectively; MPT $G^2_{4\mathrm{df}} = 26.48$ and 41.16, respectively). We reached analogous conclusions in analyses of subgroups in Experiment 2 and 3.

In summary, Experiment 1 revealed that a response bias manipulation in a syllogistic reasoning task produced an artifactual accuracy effect as measured with $H - F$. Because accuracy effects in the belief bias task are typically measured with a contrast involving $H - F$, one implication is that the apparent accuracy effects in the belief bias task could be the result of response bias differences across the believability of the conclusion. The goal of Experiment 2 was to investigate the locus of belief bias effects using ROC analysis and model-based comparisons.

Experiment 2

Experiment 2 used a design similar to that of Experiment 1 but with a manipulation of conclusion believability rather than perceived base rate. As in typical studies of the belief bias effect, subjects in Experiment 2 evaluated the conclusions of four kinds of syllogisms: those with believable and unbelievable conclusions, each of which was equally likely to be logically valid or invalid. As in Experiment 1, we compared summary statistics based on $H - F$ (including the interaction index), ROC area (A_z), and quantitative models (SDT and MPT2) to more thoroughly assess accuracy effects. Believability-based differences in $H - F$ (the Belief × Logic interaction) in this case could imply a true sensitivity effect or a Type I error (Rotello et al., 2008) depending on the nature of the empirical ROC functions and the conclusions of the best fitting model.

Method

Subjects. Thirty-eight undergraduates at the University of Massachusetts (Amherst, MA) participated; they received extra credit in their psychology courses in exchange for their participation.

Design. Experiment 2 used a 2 (valid or invalid problem) × 2 (believable or unbelievable conclusion) within-subjects design. All subjects were asked to evaluate the validity of syllogisms that differed in their logical status and conclusion believability.

Stimuli. All subjects were presented with the same 16 syllogistic problem frames as in Experiment 1. The abstract content used in Experiment 1 was replaced with meaningful content; each problem frame was assigned content that led to one believable conclusion and one unbelievable conclusion, thus providing a total of 32 problems. All sets of content were randomly assigned to the 32 problem structures.

Meaningful content for 13 problems was taken from a previous study (Morley et al., 2004); new content was generated for the remaining 19 problems. For the new content, conclusion believability was rated in a separate study. Fifty-nine psychology undergraduates at the University of Massachusetts rated the believability of a large set of potential conclusion statements, presented in isolation, on a 1–5 scale (1 = *unbelievable,* 3 = *neutral,* 5 = *believable*). Items that elicited the most extreme ratings, on average, were selected to serve as conclusions in Experiment 2. (See Appendix C for a detailed list of the conclusions and ratings.) All content was chosen such that the conclusions described a category–exemplar relationship between the subject and predicate terms. To minimize the effects of premise believability, subject and predicate terms were linked via an esoteric middle term. For example:

No sculptors are Hammerkops.

Some Hammerkops are not artists.

____

*Some artists are not sculptors. (D)

The semantic content was counterbalanced across subjects such that it appeared in both believable and unbelievable forms and both valid and invalid structures. Between subjects, modulation of believability was accomplished by reversing the order of assignment of words to the subject and predicate positions. In other words, for each subject who received the conclusion “Some spiders are not insects,” an equal number received the conclusion “Some insects are not spiders,” and no subject received both conclusions. Furthermore, for each of the 16 structures, the actual believable or unbelievable content was also varied so that, for example, the structure in Example D received one set of content (e.g., artists/sculptors/Hammerkops) for one subset but a different set of content (e.g., insects/spiders/metazoans) for another subset. Counterbalancing the content by believability and validity thus yielded four subsets of 32 problems, with each structure containing a unique set of content in each subset.

As we described in the Method section of Experiment 1, half of the problem structures allowed illicit conversion, which has previously been described as a structural confound in belief bias experiments (Evans et al., 1983; Revlis, 1975; Revlin et al., 1980). Although the premise quantifiers in half of our problems were convertible, conversion of each problem would lead to the same response; thus, these particular problems should not influence the belief bias effect.

Finally, all subjects received three valid and two invalid practice problems with different structures, selected from the pool of abstract stimuli used in the practice phase of Experiment 1.

Procedure. All subjects were tested individually. The procedure was identical to that of Experiment 1 except that the instructions did not mention any base-rate manipulation.

Results

The proportion of conclusions accepted is reported in Table 2 as a function of the validity and believability of the conclusion. Similar to Experiment 1, we analyzed these responses using a 2 × 2 ANOVA with logical status and believability as within-subjects factors. Subjects were more likely to accept valid than invalid conclusions, $F(1, 37) = 59.55$, $MSE = .061$, $p < .001$, $\eta^2 = .617$, and they were more likely to accept believable than unbelievable conclusions, $F(1, 37) = 38.24$, $MSE = .054$, $p < .001$, $\eta^2 = .508$. Importantly, these factors interacted: There was a greater effect of logical status for unbelievable than believable problems, $F(1, 37) = 6.50$, $MSE = .020$, $p < .05$, $\eta^2 = .150$. This pattern of responses replicates previous studies of belief bias (e.g., Evans et al., 1983).

The increased tendency to accept problems with believable conclusions may indicate a form of response bias, akin to the response bias displayed by subjects in the liberal condition of Experiment 1. To evaluate that possibility, we used the confidence ratings to generate ROCs for believable and unbelievable problems; the results are plotted in Figure 4B. As expected from the values of H and F, the operating points for believable problems are displaced upward and rightward relative to the corresponding points for unbelievable problems, indicating a main effect of believability on response bias: Subjects responded "valid" more often to problems with believable conclusions. The points for both believable and unbelievable problems appear to lie on a single ROC, however, indicating that subjects showed little to no difference in accuracy when judging conclusion validity. A comparison of A_z for believable and unbelievable ROCs was consistent with this interpretation, indicating there was no difference in the area under the ROC as a function of problem type ($z = 0.62$, $p = .54$).

As in Experiment 1, the source of the measurement discrepancy is apparent upon consideration of the implied and observed ROCs. In Figure 4B, the linear functions implied by $H - F$ are superimposed on the observed responses to believable and unbelievable problems. The effect of the believability manipulation in the present experiment is similar to the effect of perceived base rates in Experiment 1: A change in response bias resulted in a change in the y-intercepts of the ROCs implied by $H - F$.

Models fit. As with Experiment 1, we fit two versions of a multinomial model to the data, though the fit of the model corresponding to MPT1 was again very poor in all comparisons and is not considered further. Minor modifications were made to MPT2 to account for the believability factor in the design. First, rather than four reasoning-stage parameters to measure detection of valid (v) and invalid (i) problems for the conservative (c) and liberal (l) groups of Experiment 1, there were four parameters for detection of valid and invalid problems with unbelievable (u) and believable (b) conclusions. In other words, the parameters r_{vl}, r_{vc}, r_{il}, and r_{ic} were replaced with the four reasoning-stage parameters (r_{vb}, r_{vu}, r_{ib}, r_{iu}) of Klauer et al.'s (2000) model. Second, the two condition-defined response-stage parameters used to fit the data of Experiment 1 (β_l and β_c) were replaced with the two believability-defined response-stage parameters of Klauer et al.'s model (β_b and β_u). These changes resulted in a full (unconstrained) model with nine parameters. Thus, MPT2 as applied to these data is a direct extension of Klauer et al.'s model to a confidence-rating design.

As with Experiment 1, we also fit a signal detection model to the data. For the SDT model, $H - F$ is an inappropriate accuracy measure; if the SDT model fits the data better than the multinomial model, then conclusions based on $H - F$ are likely to be erroneous (Rotello et al., 2008).

Modeling results. The best fitting parameter values for each model are presented in Table 5, and fit statistics are presented in Table 6. As in Experiment 1, the SDT model consistently provided a better fit than the MPT2 model. Using the parameter values in Table 5, we generated predicted ROCs for each condition; these are presented in Figure 7 along with the observed data. As is clear in the figure, the MPT2 model produces ROCs that do not adequately describe the observed ROCs: The form of each predicted ROC is too linear, and the predicted operating points do not correspond very well to the observed points. In contrast, the parameters of the SDT model generate ROCs that are more similar in form to the empirical ROCs, and each point predicted by the model falls near the corresponding observed point. The fact that the SDT model again provides a better description of the data confirms the nonlinearity that is visually apparent in Figure 4B and adds further support for our claim that A_z is a better measure of accuracy for conclusion-evaluation data than $H - F$.

Table 5
Best Fitting Parameter Values for MPT2 and SDT, Experiment 2

Model	Parameter	Believable	Unbelievable
MPT2	r_v	.59	.50
	r_i	.22	.35
	β	.67	.45
	α_2	.83	
	α_4	.21	
	ϵ	.16	
SDT	μ_v	0.73	0.93
	σ_v	1.06	1.32
	c_1	0.58	1.17
	c_2	−0.19	0.57
	c_3	−0.33	0.40
	c_4	−0.48	0.26
	c_5	−0.96	−0.48
	d_a	0.71	0.79

Note. In the SDT, d_a is computed from μ_v and σ_v (see Equation 4 in the text) and is not an extra free parameter. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model.

The parameter values obtained with MPT2 are similar to those reported by Klauer et al. (2000). The reasoning-stage parameters r_{vu} and r_{ib} are lower than the values for r_{vb} and r_{iu}, respectively, indicating that reasoning is reduced when there is a conflict between logical status and believability at a given level of validity. There also appears to be an effect of believability on the response bias parameter β, with $\beta_b > \beta_u$, suggesting the observed interaction between logical status and believability is a product of effects on both the reasoning stage and response stage. The pattern of parameters in the SDT model suggests a different story, however, with marked effects on the response criteria (more negative criteria for believable than for unbelievable problems) and relatively smaller effects on the sensitivity measure d_a, which is derived from the free parameters of the model.

To evaluate the significance of the parameter differences across levels of believability, hypothesis tests were conducted for both models by constraining certain parameter values to be equal for believable and unbelievable problems. For the response stage in MPT2, a null effect of believability suggests a restricted model in which $\beta_b = \beta_u$. Imposing this restriction significantly reduced the fit of the multinomial model ($\Delta G^2_{1\text{df}} = 28.73$, $p < .001$), indicating an effect of believability on the response stage. An analogous test of response bias in the SDT model implies a single set of criteria ($c_1 \ldots c_5$) for believable and unbelievable problems. Imposing this restriction also reduced the fit of the SDT model ($\Delta G^2_{5\text{df}} = 64.70$, $p < .001$). Thus, the models concur that one effect of believability is to shift response bias.

A more critical question is whether the models attribute performance differences for believable and unbelievable problems to a difference in the accuracy with which the validity of the conclusions is judged. In the multinomial model, we addressed this question by considering the reasoning-stage parameters, r_{vb}, r_{vu}, r_{ib}, and r_{iu}. If believability affects accuracy, then constraining the model parameters so that reasoning is equally successful for believable and unbelievable problems at each level of validity (i.e., the restrictions $r_{vb} = r_{vu}$ and $r_{ib} = r_{iu}$) should worsen the fit of the model. Imposing this restriction markedly reduced the fit of the model ($\Delta G^2_{2\text{df}} = 14.63$, $p < .001$), indicating an interaction at the

Table 6
Fit Statistics for MPT2 and SDT in Experiment 2

	MPT2			SDT		
Conclusion	AIC	BIC	$G^2_{4\text{df}}$	AIC	BIC	$G^2_{3\text{df}}$
Believable	1,820.33	1,846.79	83.60	1,748.68	1,779.55	9.95
Unbelievable	1,918.82	1,945.28	33.34	1,900.47	1,931.34	12.99

Note. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model; AIC = Akaike information criterion; BIC = Bayesian information criterion.

level of the reasoning-stage parameters as previously reported by Klauer et al. (2000). An analogous test of the reasoning stage in the SDT model implies the restriction $d_{a,\ b} = d_{a,\ u}$. As d_a depends on both the distance between the valid and invalid distributions (μ_v) and the ratio of the valid and invalid standard deviations ($1/\sigma_v$; see Equation 4), a test of equal sensitivity implies equal slope and mean separation for believable and unbelievable problems, which frees two degrees of freedom for the chi-square test. Imposing this constraint had little effect on the fit of the SDT model ($\Delta G^2_{2\text{df}} = 2.39$, $p > .25$), indicating a negligible effect of believability on accuracy.[5]

To summarize, SDT outperforms MPT2[6] and suggests there is no effect of conclusion believability on sensitivity to logical status. The results for MPT2 replicated the pattern in the reasoning-stage parameters observed in Klauer et al. (2000) but failed to capture the form of the observed ROCs and thus provided a worse fit. The MPT2 model indicated an interaction at the level of the reasoning-stage parameters, consistent with the results in $H - F$ (see Table 2), a statistic that is not appropriate for these data because the observed ROCs are not linear.

Discussion. These data replicated the standard belief bias effect in all respects: Subjects accepted significantly more valid than invalid and more believable than unbelievable conclusions. In addition, the key interaction between believability and validity in the conclusion acceptance rates was observed. Interestingly, in a pattern of results similar to that of Experiment 1, accuracy, as measured by the area under the ROC (A_z), did not differ with conclusion believability. Furthermore, the SDT model, which assumes a nonlinear ROC, consistently outperformed the MPT2 model in describing the data. Hypothesis tests carried out on both models revealed opposing conclusions regarding the locus of the belief bias effect, with MPT2 locating the effect in both the reasoning and response stages and the SDT model inferring an effect solely on the response stage. The interaction between believability and validity and the notion (supported by the MPT2 model) that believability acts on the reasoning stage appear to be the result of inflated Type I error associated with the linearity assumption (Rotello et al., 2008).

These data present a pattern of acceptance rates and MPT model parameters similar to those described by Klauer et al. (2000). However, our design did not provide enough data for a between-subjects analysis of group-averaged acceptance rates analogous to that reported by Klauer et al. (i.e., the design provided too few data points relative to the number of free parameters in the original model). As such, it is possible that our novel extension of the MPT model to ROC data (i.e., the proposed mapping of the internal states to confidence ratings) produced conclusions that would not have been reached had the group analysis been possible. To address this issue, Experiment 3 included a base-rate manipulation in addition to the belief bias manipulation of Experiment 2. Thus, the data collected in Experiment 3 allowed us to apply the same model-fitting and ROC-based analysis strategy as in Experiments 1 and 2, while also allowing us to employ Klauer et al.'s model-fitting strategy.

Experiment 3

In Experiment 3, three separate groups of subjects performed the belief bias task from Experiment 2. The groups differed only with respect to the actual base rates of valid and invalid syllogisms that were presented. In the liberal group, subjects were instructed that 60% of the problems would be valid and 40% would be invalid; they were then presented with 32 valid and 22 invalid problems. In the conservative group, the subjects were instructed that 40% of the problems would be valid and 60% would be invalid; they were then presented with 22 valid and 32 invalid problems. In the neutral group, subjects were told that 50% of the problems would be valid and 50% would be invalid; they were then presented with 27 valid and 27 invalid problems.

The addition of the base-rate manipulation to the design of Experiment 2 allowed us to fit the MPT model to the average acceptance rates for three groups differing only in response bias, as in Klauer et al. (2000). For this analysis, the original model used by the authors was applied; no additional parameterization was necessary. The model, which we refer to as MPTK, is illustrated in Figure 1. As in Klauer et al., each tree in Figure 1 was extended with an extra response bias parameter α_x, where x = low, medium, or high, corresponding to the base-rate conditions; the baseline model for acceptance rates thus contains nine parameters. The equations for MPTK can be found in Appendix A.

As in Experiment 2, Klauer et al.'s (2000) MPT model was also extended to the confidence-rating data. The model was fit to ROCs for the conservative, liberal, and neutral groups and for believable and unbelievable arguments. The fit to ROCs allowed comparison of the parameter values with those obtained in the fit to average acceptance rates, as well as comparison with the SDT model. To examine the locus of effects stemming from base rates and conclusion believability, hypothesis tests were conducted by constraining certain parameter values to be constant across conditions. The ROC data were again used to supplement an analysis using $H - F$, to provide a more thorough examination of believability and base-rate effects indicated in the observed response proportions.

The present design also included a greater number of trials per subject, to increase the number of observations contributing to the operating points of the ROCs and to increase our power to detect the effects of the base-rate manipulation.

[5] As in Experiment 1, constraining the variances of the valid and invalid distributions to be equal impaired the fit ($\Delta G^2_{2\text{df}} = 8.73$, $p < .02$).

[6] As in Experiment 1, we also fit the ROCs for a group of 15 high- and 15 medium-accuracy subjects to evaluate whether averaging over subjects changed the model selection results. For both groups, the SDT model fit significantly better than the MPT model (SDT medium-accuracy $G^2_{3\text{df}} = 1.16$ and high-accuracy 14.19; MPT $G^2_{4\text{df}} = 42.50$ and 20.77).

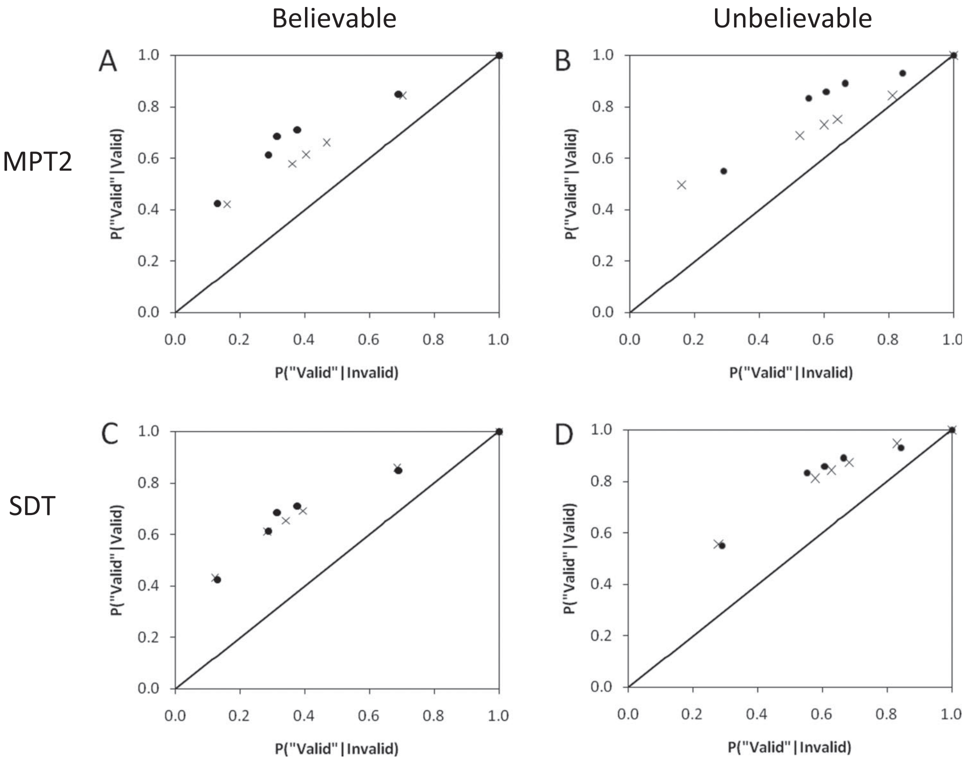

Figure 7. A: Observed (circles) and MPT2-predicted (crosses) receiver operating characteristics (ROCs) for believable problems, Experiment 2. B: Observed and MPT2-predicted ROCs for unbelievable problems. C: Observed (circles) and SDT-predicted (crosses) ROCs for believable problems in Experiment 2. D: Observed and SDT-predicted ROCs for unbelievable problems. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model.

Method

Subjects. Seventy-two undergraduates at the University of Massachusetts volunteered to participate in exchange for extra credit in their psychology courses.

Design. Experiment 3 used a 2 (validity) × 2 (believability) × 3 (base rate) mixed design. All subjects were asked to evaluate the validity of 54 syllogisms differing in logical status and conclusion believability. Subjects were randomly divided into three groups differing in the proportion of valid and invalid syllogisms. Subjects in the conservative group received 11 believable valid, 11 unbelievable valid, 16 believable invalid, and 16 unbelievable invalid syllogisms. Subjects in the liberal group received 16 believable valid, 16 unbelievable valid, 11 believable invalid, and 11 unbelievable invalid syllogisms. In the neutral group, subjects received equal numbers of valid, invalid, believable, and unbelievable syllogisms.

Stimuli. All subjects evaluated 54 syllogisms, created using the four subsets of 32 stimuli from Experiment 2. Subjects received two blocks of stimuli (although the blocking was not apparent to them). The same set of structures contributed to both blocks, the main difference being in the actual content used to construct the stimuli for a given subject in each block. In the first block, subjects received the believable (or unbelievable) version of each set of content (e.g., "Some skyscrapers are not buildings"), and in the second block, they received the unbelievable (or believable) version of the same set (e.g., "Some buildings are not skyscrapers"). The version in the second block was always assigned to a different structure with the same validity status as in the first block. The assignment of content sets to validity status was also counterbalanced between subjects, so that, for example, the set containing skyscrapers/buildings appeared in both valid and invalid structures.

For the conservative group, subjects evaluated 22 valid syllogisms, randomly sampled from the sets corresponding to Blocks 1 and 2, with the constraint that equal numbers of believable and unbelievable stimuli appeared in the sample. In the first block, subjects received either five believable (and six unbelievable) valid syllogisms or six believable (and five unbelievable) valid syllogisms. The numbers for each subject were reversed in the second block, so that each subject evaluated a total of 11 believable valid and 11 unbelievable valid syllogisms in the conservative group. All 32 invalid stimuli (16 per block) were presented in the conservative condition; half of these were believable, and half were unbelievable. The same logic was applied to the liberal group; counterbalancing thus meant each subject in the liberal group evaluated 11 believable invalid and 11 unbelievable invalid syllogisms; all 32 valid stimuli were presented in the liberal condition. In the neutral group, subjects received equal numbers of valid and invalid stimuli. As 54 stimuli were used, with the constraint that equal numbers of valid, invalid, believable, and unbelievable problems were presented, it was necessary to counterbalance the exact number of problems in each of the four cells defined by problem type. Thus, subjects received six stimuli from one of the four problem types in Block 1 (and seven from each of the other three types in the same block) and six from the type in Block 2 that differed in both believability and validity status from the set of six stimuli in Block 1 (with seven from each of the other three types). So, for example, a subject who received six believable valid problems in Block 1 would receive six unbelievable invalid problems in Block 2. Counterbalancing by problem type yielded four subgroups, each with a total of 27 problems at each level of validity and believability.

Finally, all subjects received practice problems, selected from the pool of stimuli used in the practice phase of Experiment 1. For the practice phase, the conservative group received two valid and four invalid problems; the liberal group received four valid and two invalid problems; the neutral group received three valid and three invalid problems.

Procedure. The procedure was similar to that of Experiment 1, the only exceptions being the use of different numbers of practice (six) and critical trials (54), the specific percentages of valid and invalid problems indicated in the instructions (60%/40%, 40%/60%, or 50%/50%), and the fact that problem presentation was accomplished using blocked randomization.

Results

The observed response proportions for the four problem types and the three groups are summarized in Table 7. Table 7 also shows the corresponding logic, belief, and interaction contrasts. The proportion of conclusions accepted was analyzed using a 2 × 2 × 2 × 3 mixed ANOVA with validity, believability, and test block as within-subjects factors and base rate (low, medium, high) as a between-subjects factor. The ANOVA confirmed that subjects accepted more valid than invalid conclusions, $F(1, 69) = 126.480$, $MSE = .087$, $p < .001$, $\eta^2 = .647$, and more believable than unbelievable conclusions, $F(1, 69) = 70.509$, $MSE = .074$, $p < .001$, $\eta^2 = .505$. As in Experiment 2, believability and validity interacted: There was a greater effect of validity for unbelievable than for believable problems, $F(1, 69) = 15.224$, $MSE = .029$, $p < .001$, $\eta^2 = .181$, as in typical belief bias experiments. There was also a main effect of condition, $F(2, 69) = 23.857$, $MSE = .099$, $p < .001$, $\eta^2 = .409$, indicating differences in response bias as a function of actual base rate. Pairwise comparisons confirmed that all three groups differed from one another in overall acceptance rates; the liberal group accepted more conclusions than the neutral group, $t(70) = 6.91$, $p < .001$, and the neutral group accepted more conclusions than the conservative group, $t(70) = 4.50$, $p < .001$. Unlike the results of Experiment 1, however, base rate and logical status did not interact, $F(2, 69) = 1.602$, $MSE = .087$, $p = .209$, $\eta^2 = .044$, though planned comparisons revealed that $H - F$ was marginally larger in the conservative than the liberal group (.33 vs. .22), $t(70) = 1.784$, $p = .079$. Finally, there was a marginal interaction between validity and list half, indicating that $H - F$ was slightly higher in the first block (.32 vs. .28), $F(1, 69) = 3.173$, $MSE = .024$, $p = .079$, $\eta^2 = .044$. This likely reflects a fatigue effect due to the relatively large number of trials in the present experiment. No other effects reached or approached significance.

The greater acceptance rates for believable conclusions suggest a response bias effect. To examine this possibility, we used subjects' confidence ratings to generate ROCs for believable and unbelievable problems; the results are plotted in Figure 4C. Consistent with the results in the acceptance rates, the operating points for believable problems are displaced upward and rightward relative to the corresponding points for unbelievable problems, indicating a main effect of believability on response bias. As in Experiment 2, however, the points for both believable and unbelievable problems appear to fall on a single ROC, indicating that subjects were equally accurate in evaluating believable and unbelievable problems. A comparison of A_z for believable and unbelievable ROCs was consistent with that interpretation, revealing no difference in area under the ROC for the two problem types ($z = 1.26$, $p = .21$).

As in Experiments 1 and 2, the source of the measurement discrepancy is apparent upon inspection of the implied and observed ROCs. In Figure 4C, the functions implied by $H - F$ are superimposed on the observed ROCs for believable and unbelievable problems, collapsed across base-rate condition. The effect of the believability manipulation in the present experiment replicates the response bias effects of the previous experiments: A change in response bias resulted in an artifactual change in the *y*-intercepts of the ROCs implied by $H - F$.

As can be seen in Figure 4D, the effects of the base-rate manipulation are also apparent in the ROCs; the operating points on the ROC for the liberal condition are displaced upward and rightward relative to the corresponding points on the conservative

Table 7
Observed and Model-Predicted Hit Rates (H), False-Alarm Rates (F), and Contrast Results for Experiment 3

		H = P("Valid"|Valid)			F = P("Valid"|Invalid)			
Condition	Conclusion	Observed	MPTK	SDT	Observed	MPTK	SDT	$H - F$ observed
Liberal	Believable	.91	.91	.91	.76	.76	.76	.15
	Unbelievable	.74	.76	.75	.45	.45	.44	.29
Conservative	Believable	.72	.71	.72	.43	.40	.43	.29
	Unbelievable	.58	.60	.59	.21	.24	.21	.37
Neutral	Believable	.82	.83	.82	.60	.62	.60	.22
	Unbelievable	.72	.70	.72	.38	.36	.39	.34

Note. MPTK = multinomial processing tree model of Klauer, Musch, and Naumer (2000); SDT = signal detection theory model.

and neutral ROCs. The effect on response bias is also apparent in the comparison of the neutral and conservative ROCs: The operating points for the conservative condition are clustered nearer to the origin than the corresponding points on the neutral ROC. Though the data are consistent with our previous analyses in that there were no significant effects of response bias on A_z in any of these comparisons, there were marginal effects indicating lower accuracy for the liberal relative to the conservative condition (z = 1.83, p = .07) and for the liberal relative to the neutral condition (z = 1.82, p = .07). Although the cause of these marginal effects is unclear, it is important to remember that, despite what our results may suggest, not all effects on $H - F$ necessarily imply null effects in ROCs. Additionally, effects of actual base-rate manipulations on the form of the ROC, although not always obtained, have been reported for other experimental tasks (Van Zandt, 2000; see Benjamin, Diaz, & Wee, 2009, for review).

Models fit. In the present experiment, three versions of the MPT model were fit. As we elaborate shortly, the MPTK model was fit to the data in Table 7, using the methods of Klauer et al. (2000). MPT2 was fit to the believability ROCs, exactly as in Experiment 2 (i.e., collapsing across the base-rate conditions). A third model, MPT3, was fit to the ROCs across base-rate conditions (i.e., collapsing across believability).

The model used by Klauer et al. (2000), here termed MPTK, was fit to the group-averaged hit and false-alarm rates (i.e., the data in Table 7). This model, described in Figure 1, contains a distinct reasoning parameter for each problem type: r_{vb}, r_{vu}, r_{ib}, and r_{iu}. Following Klauer et al. and Bröder and Schütz (2009), these parameters were held constant across the base-rate conditions. There are also two guessing (bias) parameters to account for believability, one each for believable and unbelievable problems (β_b and β_u; these were constant across base-rate conditions), and there are three bias parameters to account for the base-rate manipulation (α_x, where x=low, medium, or high). Because this model describes only the overall acceptance rates for the four problem types, there are no parameters to map internal states onto confidence ratings.

The other new model for Experiment 3 is MPT3, which accounts for the base-rate factor in the ROCs but not the believability factor. This model has six reasoning parameters (r_{vc}, r_{vl}, r_{vn}, r_{ic}, r_{il}, r_{in}), corresponding to valid and invalid detection across the three base-rate conditions. Like the Klauer et al. (2000) model, MPT3 includes a response bias parameter for each base-rate condition (β_x, where x=low, medium, or high). Like the MPT2 model, MPT3 includes two bias parameters that map internal states onto confidence ratings (α_2, α_4) and a parameter (ε) that captures high-confidence errors.

As in Experiments 1 and 2, we also fit a signal detection model to the data in each base-rate condition and for both believable and unbelievable conclusions. For the SDT model, $H - F$ is an inappropriate accuracy measure. Therefore, if the SDT model outperforms the MPT models, then conclusions based on $H - F$ are likely to be erroneous (Rotello et al., 2008).

Modeling results.

Belief bias ROC modeling. The MPTK model of Klauer et al. (2000) provided a good fit to the data ($G^2_{3\mathrm{df}} = 5.63$, $p > .05$). The best fitting parameter values can be found in Table 8, and the predicted and observed response proportions can be found in Table 7. The pattern across the α parameters shows that the model is sensitive to the differences in the base rate of valid problems across conditions. Imposing the restriction that base rate had no effect (i.e., $\alpha_l = \alpha_m = \alpha_h$) significantly reduced the quality of fit ($\Delta G^2_{2\mathrm{df}} = 163.50$, $p < .001$), confirming that the groups differed in their overall willingness to say "valid."

The reasoning-stage parameters show the same pattern as in Klauer et al. (2000): $r_{vu} < \mathrm{r}_{vb}$ and $r_{ib} < \mathrm{r}_{iu}$; in Experiment 2, using an extension of this model to the ratings design (MPT2), we also found the same pattern. These parameters imply that reasoning is less accurate when believability conflicts with validity, with particularly low accuracy for invalid believable problems. Consistent with this implication, we found that while imposing the constraint $r_{vb} = r_{vu} = r_{iu}$ did not affect the fit of the model ($\Delta G^2_{2\mathrm{df}} = 4.42$, $p > .05$), imposing the constraint $r_{vb} = r_{vu} = r_{ib} = r_{iu}$ led to a significant loss in goodness of fit ($\Delta G^2_{3\mathrm{df}} = 29.85$, $p < .001$). The values for the believability-defined bias parameters (β_b and β_u) indicate belief bias in the response stage as well, with a higher value for β_b than β_u. Setting the β parameters equal reduced the fit of the model ($\Delta G^2_{1\mathrm{df}} = 10.61$, $p < .01$). These findings replicate those of our previous experiments and indicate that our results are not due to the novel extension of the MPT model to confidence ratings. According to MPTK, the interaction between validity and believability that is observed in the interaction index of $H - F$ is due to effects on the reasoning and response stages.

We also addressed the issue of whether the use of confidence-rating data to generate empirical ROCs distorted the results in favor of the SDT model over the multinomial models. Bröder and Schütz (2009) recently argued that multinomial and SDT models are more fairly compared when the operating points on the empirical ROC are collected independently of one another via a manipulation of the base rate of target items or a payoff scheme. Therefore, we also fit a signal detection model to the group data, using a model like the one in Figure 3B. The accuracy parameter d_{ax}, which combines μ_{vx} and σ_{vx} (where x = believable or unbelievable; see Equation 4), was fixed across base-rate conditions; only the six decision criteria (c_{yx}, one per base-rate condition, y,

Table 8
Best Fitting Parameter Values for MPTK and SDT for the Base-Rate Conditions of Experiment 3

Model	Parameter	Liberal	Conservative	Neutral
MPTK	r_{vb}		.49	
	r_{ib}		.07	
	r_{vu}		.43	
	r_{iu}		.22	
	β_b		.99	
	β_u		.69	
	α	.82	.44	.67
SDT	μ_{vb}		0.85	
	μ_{vu}		1.13	
	σ_{vb}		1.18	
	σ_{vu}		1.49	
	C_b	−0.72	0.18	−0.25
	C_u	0.14	0.81	0.28
	$d_{a,\ b}$		0.78	
	$d_{a,\ u}$		0.89	

Note. In the SDT, d_a is computed from μ_v and σ_v (see Equation 4 in the text) and is not an extra free parameter. MPTK = multinomial processing tree model of Klauer, Musch, and Naumer (2000); SDT = signal detection theory model.

and believability value, x) were allowed to vary. If our analyses of the confidence-rating data were biased in favor of the SDT model, then a comparison of MPTK and SDT fits to the response proportions in each base-rate condition should result in a conclusion that differs from those reached in the ratings comparisons.

The best fitting parameter values of each model are shown in Table 8, the predicted response proportions are shown in Table 7, and the fit statistics G^2, AIC, and BIC are shown in Table 9. While both models fit the data, the SDT model fit better in terms of G^2 and AIC, though MPTK had a slight edge in BIC. As in our other analyses, we found that constraining the SDT accuracy parameters to be equal for the believable and unbelievable problems did not harm the fit ($\Delta G^2_{2\text{df}} = 0.77$, $p > .6$), but equating the criteria for believable and unbelievable problems ($c_\text{b} = c_\text{u}$ in each base-rate condition) did ($\Delta G^2_{3\text{df}} = 150.05$, $p < .001$). Thus, the SDT model fits well and implies that conclusion believability only affects response bias. This result is consistent with our conclusions based on the confidence-rating ROCs and suggests that our analyses have not distorted the data.

Having confirmed that our acceptance-rate data (see Table 7) and basic modeling efforts replicate the Klauer et al. (2000) findings, we turned our attention to fitting the observed ratings ROCs with the remaining models (MPT2, MPT3, and SDT). The best fitting parameter values for each model are shown in Tables 10 and 11, and the corresponding fit statistics are reported in Table 12.

As in Experiments 1 and 2, the SDT model consistently provided a better fit to the data than did the multinomial model applied to either the ROCs as a function of base-rate condition (MPT3) or believability (MPT2). In Figure 8, ROCs generated with the best fitting parameter values of MPT2 and SDT are shown along with the observed data for believable and unbelievable problems. As in Experiment 2, the MPT2 model generates linear ROCs that are inconsistent with the form of the observed ROCs. In contrast, the SDT model produces ROCs that are much closer in form to the observed data, with each predicted point falling close to the corresponding observed point. The fact that the SDT model provides a better description of the data than does MPT2 confirms what is visually apparent in Figures 4C and 4D: The ROCs are not linear. This further supports our claim that A_z is a better accuracy measure than $H - F$ or the interaction index for these data.

The parameter values for the reasoning stage in the MPT2 model (see Table 10) show the same pattern as the one obtained with the MPTK model: $r_{vu} < r_{vb}$ and $r_{ib} < r_{iu}$, indicating the interaction between validity and believability observed in the acceptance rates is due to an effect on reasoning processes. We tested this hypothesis by imposing the restrictions that $r_{vb} = r_{vu}$ and $r_{ib} = r_{iu}$. The fit of MPT2 was significantly reduced by that restriction ($\Delta G^2_{2\text{df}} = 52.57$, $p < .001$), supporting the conclusion that there are substantial effects of believability on the reasoning stage. However, the believability-defined response bias parameters, β_b and β_u, also differ from one another, replicating the effect that we observed in Experiment 2. Indeed, imposing the restriction that the bias parameters were equal ($\beta_b = \beta_u$) significantly reduced the fit of the model ($\Delta G^2_{1\text{df}} = 36.95$, $p < .001$). Thus, MPT2 concludes that believability affects both the reasoning stage and the decision stage of processing.

The parameter values for SDT (see Table 10) suggest a different conclusion. While there are large effects of believability on response bias (i.e., more negatively valued criteria for believable than for unbelievable problems), the effects on the accuracy measure d_a (derived from μ and σ) are quite small. Model-based hypothesis tests supported these interpretations. Imposing the con-

Table 9
Fit Statistics for MPTK and SDT in Experiment 3

Model	AIC	BIC	G^2	*df*
MPTK	4,488.68	4,545.07	5.61	3
SDT	4,485.38	4,548.04	0.34	2

Note. MPTK = multinomial processing tree model of Klauer, Musch, and Naumer (2000); SDT = signal detection theory model; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table 10
Best Fitting Parameter Values for MPT2 and SDT for Problem Types in Experiment 3

Model	Parameter	Believable	Unbelievable
MPT2	r_v	.60	.52
	r_i	.24	.40
	β	.61	.47
	α_2	.68	
	α_4	.37	
	ϵ	.17	
SDT	μ_v	0.68	0.81
	σ_v	1.03	1.14
	c_1	0.52	0.96
	c_2	0.01	0.52
	c_3	−0.26	0.32
	c_4	−0.51	0.09
	c_5	−0.95	−0.43
	d_a	0.67	0.76

Note. In the SDT, d_a is computed from μ_v and σ_v (see Equation 4 in the text) and is not an extra free parameter. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model.

Table 11
Best Fitting Parameter Values for MPT3 and SDT for the Base-Rate Conditions of Experiment 3

Model	Parameter	Liberal	Conservative	Neutral
MPT3	r_v	.57	.45	.62
	r_i	.24	.31	.39
	β	.67	.40	.55
	α_2		.68	
	α_4		.37	
	ϵ		.17	
SDT	μ_v	0.54	0.81	0.71
	σ_v	1.02	1.27	1.01
	c_1	0.48	1.15	0.58
	c_2	−0.03	0.57	0.20
	c_3	−0.34	0.42	−0.01
	c_4	−0.48	0.11	−0.28
	c_5	−0.94	−0.60	−0.57
	d_a	0.53	0.71	0.71

Note. In the SDT, d_a is computed from μ_v and σ_v (see Equation 4 in the text) and is not an extra free parameter. MPT3 = Multinomial Processing Tree 3 model; SDT = signal detection theory model.

Table 12
Fit Statistics for MPT2 (Conclusion Believability), MPT3 (Base Rate), and SDT in Experiment 3

Condition/ conclusion	MPT2, MPT3			SDT		
	AIC	BIC	$G^2_{4\text{df}}$	AIC	BIC	$G^2_{3\text{df}}$
Believable	6,168.47	6,201.91	245.59	5,936.29	5,975.30	11.41
Unbelievable	6,286.35	6,319.79	92.09	6,231.67	6,270.68	35.41
Liberal	4,064.99	4,095.99	160.66	3,920.78	3,956.95	14.45
Conservative	3,772.00	3,802.20	65.68	3,719.03	3,754.26	10.71
Neutral	4,583.58	4,615.29	124.04	4,467.33	4,504.32	5.79

Note. MPT2 = Multinomial Processing Tree 2 model; MPT3 = Multinomial Processing Tree 3 model; SDT = signal detection theory model; AIC = Akaike information criterion; BIC = Bayesian information criterion.

straint of a single set of five criteria in the SDT model (i.e., no differences as a function of believability) led to a marked loss in goodness of fit ($\Delta G^2_{5\text{df}} = 223.41$, $p < .001$), but constraining accuracy to be equal for believable and unbelievable problems (i.e., constraining the parameters for mean separation, $\mu_{vb} = \mu_{vu}$, and slope, $\sigma_{vb} = \sigma_{vu}$) had no significant effect on the fit[7] ($\Delta G^2_{2\text{df}} = 2.71$, $p > .25$). According to SDT, believability affects response bias but not accuracy.

These results are consistent with those of Experiment 2: The signal detection model provides a better description of the data and locates the belief bias effect in the response stage.[8] The MPTK and MPT2 models imply linear ROCs that are inconsistent with the observed ROCs; the models also conclude there are effects of believability on both the reasoning and response stages. The erroneous conclusions suggested by the MPTK and MPT2 analyses arise from inappropriate assumptions about the form of the ROC, assumptions that are shared by the statistic $H - F$.

Base-rate ROC modeling. We also evaluated the SDT and multinomial models using the ratings ROCs for each base-rate condition, collapsing over the believability factor. Both models should show changes in their response bias parameters across conditions, but because all conditions included the same problems, we would not expect the model parameters to indicate accuracy differences across the base-rate factor. This analysis involved the MPT3 and SDT models; Figure 9 shows the best fitting ROCs. The parameter values of both the MPT3 and SDT models confirm that the response bias manipulation was effective: The β parameters of MPT3 increase with the proportion of valid problems, as does the number of negatively valued (i.e., liberally placed) decision criteria in the SDT model. Constraining these parameters to be equal across base-rate conditions significantly worsened the fit of both models (for MPT3, $\Delta G^2_{2\text{df}} = 85.51$, $p < .001$, and for SDT, $\Delta G^2_{10\text{df}} = 215.853$, $p < .001$). Thus, both models detect substantial effects of condition on willingness to say "valid," consistent with the results of Experiment 1.

Inspection of the parameter values in Table 11 reveals differences in the reasoning-stage parameters as a function of base rate in MPT3, indicating effects of condition on valid and invalid detection that are similar to those of Experiment 1. Specifically, the pattern $r_{il} < r_{ic} < r_{in}$, $r_{vc} < r_{vl} < r_{vn}$, was obtained, indicating accuracy differences among the three groups, with reasoning success being particularly high in the unbiased (50% valid) condition. Tests of the hypotheses $r_{vn} = r_{vl}$, $r_{in} = r_{il}$, and $r_{vl} = r_{vc}$, $r_{il} = r_{ic}$, indicated both sets of constraints significantly reduced the fit of MPT3 ($\Delta G^2_{2\text{df}} = 25.50$, $p < .001$, and $\Delta G^2_{2\text{df}} = 18.80$, $p < .001$, respectively). The constraint $r_{vn} = r_{vc}$, $r_{in} = r_{ic}$, also reduced the fit of MPT3 ($\Delta G^2_{2\text{df}} = 35.38$, $p < .001$). The MPT3 model indicates the base-rate manipulation significantly affected sensitivity to validity status in these three groups, with accuracy being particularly high in the neutral group. An analogous test in the SDT model implies the constraint $d_{a,\ c} = d_{a,\ l} = d_{a,\ n}$, which can be accomplished by setting the parameters for mean separation (μ_v) and slope ($1/\sigma_v$) equal across the three conditions. Imposing this constraint led to a significant loss in fit for SDT ($\Delta G^2_{4\text{df}} = 11.05$, $p < .05$). Unlike the results for MPT3, however, the difference in accuracy was driven by the liberal condition; a test of the constraint $d_{a,\ c} = d_{a,\ n}$ indicated no difference in accuracy between the conservative and neutral conditions ($\Delta G^2_{2\text{df}} = 5.15$, $p > .05$). Thus, while the results for MPT3 indicated differences in accuracy among all three groups, with particularly high accuracy in the neutral condition, the results for SDT were consistent with the results for A_z that indicated slightly lower sensitivity for the liberal group but no difference in sensitivity between the conservative and neutral groups.

Discussion. Experiment 3 replicated the main results of Klauer et al. (2000) and of Experiment 2. A standard belief bias effect was observed in the acceptance rates, and an analysis using multinomial models, including MPTK (the original model of Klauer et al., 2000), indicated that these results were due to effects of believability on response bias and sensitivity to logical status. In contrast, the signal detection model, which provided the best fit to the data in every condition, indicated no effect of believability on accuracy but large response bias effects. As in Experiments 1 and 2, Experiment 3 showed that the accuracy measure $H - F$, which is implicitly or explicitly used in most analyses of the belief bias effect, is influenced by changes in response bias. Conclusions based on $H - F$ were similar to those suggested by the Klauer et al. and MPT2 models, which make assumptions about the form of the ROC that are similar to the assumptions of $H - F$. An analysis of A_z, which is theoretically and empirically justified by the form of the ROCs and the fit of the SDT model, suggested no difference in accuracy as a function of believability. Because the SDT model provided the best account of the data, A_z is preferred over $H - F$ in the analysis of data from belief bias experiments using the conclusion-evaluation task. Use of $H - F$ produced a Type I error in measuring accuracy across conclusion believability, consistent with the results of the simulations in Rotello et al. (2008).

General Discussion

In a series of three experiments, changes in response bias in the syllogistic reasoning task produced changes in accuracy as measured by $H - F$, though no such changes were apparent in

[7] As in Experiment 1, constraining the variances of the valid and invalid distributions to be equal impaired the fit ($\Delta G^2_{2\text{df}} = 6.02$, $p < .05$).

[8] As in Experiments 1 and 2, we also fit the ROCs for a group of 15 high- and 15 medium-accuracy subjects to evaluate whether averaging over subjects changed the model selection results. For both groups, the SDT model fit significantly better than the MPT model (SDT medium-accuracy $G^2_{4\text{df}} = 4.70$ and high-accuracy 25.87; MPT $G^2_{4\text{df}} = 46.41$ and 213.23).

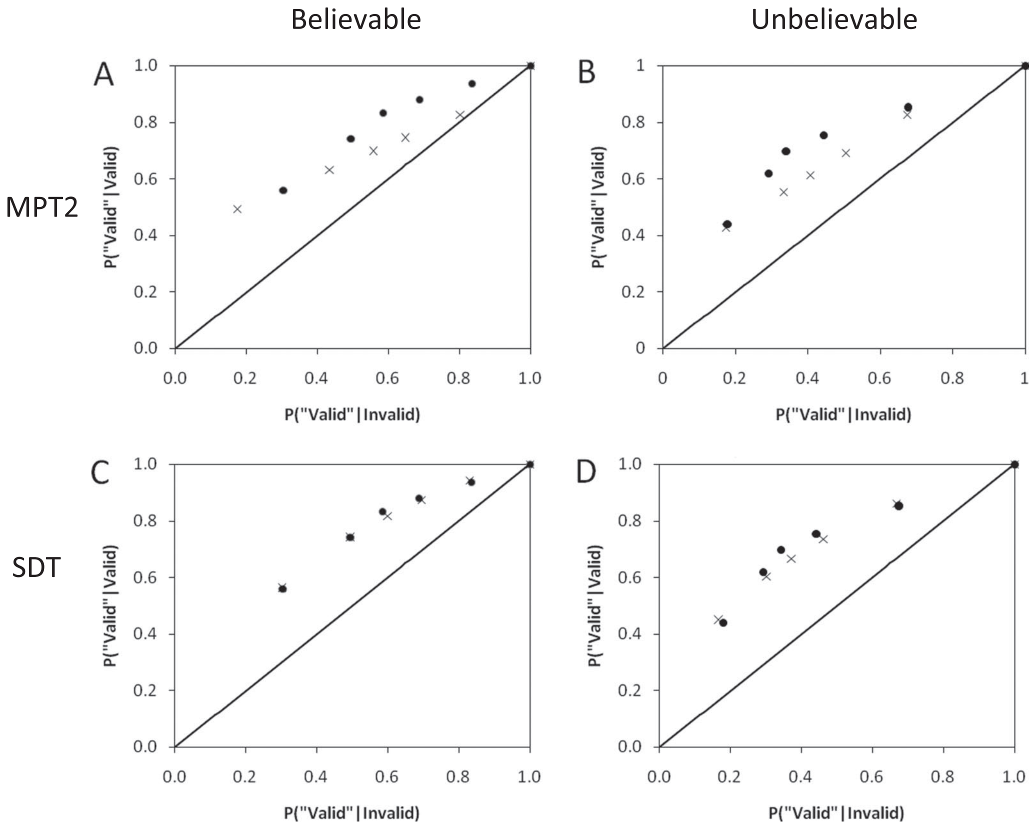

Figure 8. A: Observed (circles) and MPT2-predicted (crosses) receiver operating characteristics (ROCs) for believable problems, Experiment 3. B: Observed and MPT2-predicted ROCs for unbelievable problems. C: Observed (circles) and SDT-predicted (crosses) ROCs for believable problems in Experiment 3. D: Observed and SDT-predicted ROCs for unbelievable problems. MPT2 = Multinomial Processing Tree 2 model; SDT = signal detection theory model.

estimates of ROC area (A_z) or the SDT-based sensitivity parameter d_a. The same pattern of results was obtained whether bias was manipulated by changing the perceived base rates of valid and invalid problems (Experiment 1), conclusion believability (Experiments 2–3), or actual base rates (Experiment 3). A comparison of empirical and implied ROC curves indicated that the measurement discrepancy was due to the assumption of ROC linearity inherent in $H - F$. In all three experiments, the linearity assumption was directly evaluated by comparing the fit of multinomial models, which produce linear ROCs, and the signal detection model, which produces curvilinear ROCs. In every comparison, the SDT model outperformed the MPT in describing the ROCs, indicating that the observed ROCs are not linear and that contrasts of hits and false alarms confound sensitivity and response bias. The fact that we consistently obtained a Type I error in $H - F$ is consistent with the simulation results reported by Rotello et al. (2008) and provides a strong argument in favor of alternative measures of accuracy in the conclusion-evaluation task. One such alternative, the measure A_z, produced results that were in line with conclusions based on visual inspection of the observed ROCs and direct tests using quantitative models. As such, for researchers wishing to avoid measurement error associated with response bias in the conclusion-evaluation task, we recommend the use of A_z.

A major goal of the present study was to examine the locus of the belief bias effect. Although the results reported by Klauer et al. (2000) were consistent with previous theories of belief bias in suggesting that conclusion believability affects accuracy via effects on the reasoning stage, with little to no effect on the response stage, the authors used an MPT analysis to reach this conclusion. In the present study, we extended the MPT model to ROCs and found that the parameters indicated effects of our response bias manipulations on accuracy and ROC slope. In Experiment 3, we also fit the original MPT model of Klauer et al. to data from the belief bias task and replicated their finding that the apparent accuracy difference favoring unbelievable problems in hits and false alarms was modeled in the reasoning stage, just as in our extended MPT analyses. In contrast, the SDT model indicated that the effects of conclusion believability are limited to the response stage, with minimal effects on sensitivity or ROC slope. The superior fit of the SDT model supports the response bias interpretation and confirms what is visually apparent in the ROCs: The Belief × Logic interaction obtained in $H - F$ is a Type I error. These data are inconsistent with selective scrutiny theory, misinterpreted necessity, mental models theory, metacognitive uncertainty, VRT, MVRT, and the selective processing theory advanced by Klauer et al., all of which attempt to explain the Belief × Logic

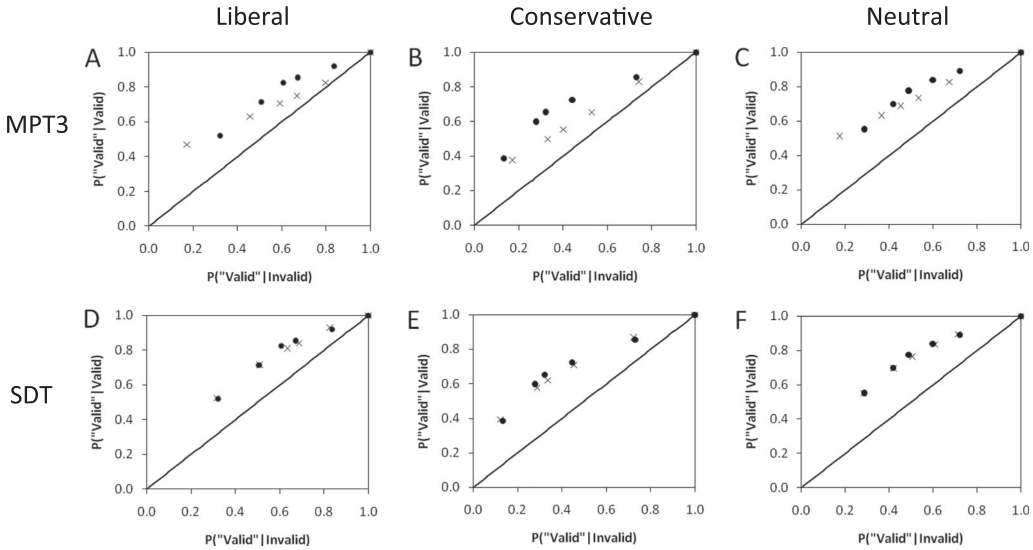

Figure 9. A: Observed (circles) and MPT3-predicted (crosses) receiver operating characteristics (ROCs) for the liberal condition of Experiment 3. B: Observed and MPT3-predicted ROCs for the conservative condition. C: Observed and MPT3-predicted ROCs for the neutral condition. D: Observed (circles) and SDT-predicted (crosses) ROCs for the liberal condition of Experiment 3. E: Observed and SDT-predicted ROCs for the conservative condition. F: Observed and SDT-predicted ROCs for the neutral condition. MPT3 = Multinomial Processing Tree 3 model; SDT = signal detection theory model.

interaction and specifically predict higher accuracy for unbelievable problems. The data also suggest a reinterpretation of results relevant to the heuristic-analytic theory of deduction (e.g., Evans & Curtis-Holmes, 2005; Shynkaruk & Thompson, 2006), which is taken up below.

In accounting for the present results, the only adequate theory of belief bias that we know of is provided by SDT. The SDT model, applied to syllogistic reasoning, assumes two Gaussian distributions of argument strength corresponding to valid and invalid problems, with higher mean strength assigned to the former. The strength axis is partitioned with a response criterion such that "valid" responses occur only for items that fall to the right of the criterion. The interpretation of the belief bias effect provided by SDT is simple: Believable conclusions produce a shift in response criteria to favor the "valid" response, and unbelievable conclusions produce a shift in response criteria to favor the "invalid" response. Valid/invalid discrimination, modeled as the mean separation of strength distributions corresponding to valid and invalid arguments, is unaffected by conclusion believability.

The criterion shift account is parsimonious, but it may also be oversimplified. For instance, it is possible that conclusion believability does not shift response criteria but instead shifts the strength distributions. In recognition memory, SDT-based explanations assuming criterion shifts and distribution shifts have been notoriously difficult to discriminate (Rotello & Macmillan, 2008). For the belief bias task, the alternative SDT-based explanation would amount to a model with a single criterion and four distributions ordered, in terms of mean strength, $\mu_{vb} > \mu_{vu} > \mu_{ib} > \mu_{iu}$, where $\mu_{vb} - \mu_{ib} = \mu_{vu} - \mu_{iu}$. Though research designed to discriminate between the two accounts of belief bias has yet to be conducted, it should be noted that in the present study, markedly similar effects of perceived base rate (Experiment 1) and conclusion believability (Experiment 2) were obtained (see Table 2 and Figures 4A and 4B), indicating that the most parsimonious interpretation of the data set as a whole is that the same process produces both effects. Since it is not clear why a change in the perceived base rate of abstract arguments would affect their strength, the criterion shift interpretation is to be preferred.

A more serious criticism of SDT is that it does not explicitly model the reasoning process or provide an account of how subjects determine whether arguments are valid or invalid. Though the criticism may be warranted, the purpose of the SDT modeling was not to provide a processing account but to determine whether conclusion believability affects accuracy and/or response bias (a goal shared by Klauer et al., 2000). It is conceivable that conclusion believability has subtle effects on how subjects process syllogistic arguments. For instance, Thompson et al. (2003) found longer response times for believable invalid than for unbelievable invalid problems but no difference between believable and unbelievable valid problems. Ball et al. (2006) used eye tracking to measure inspection times for syllogistic premises, following the first viewing of conclusions. Though premise inspection times did not differ before conclusion viewing, the authors found longer postconclusion premise times for believable invalid and unbelievable valid problems relative to believable valid and unbelievable invalid problems. Although the results of these two studies are not consistent in the specific patterns obtained across the four problem types, they at least converge in showing longer processing times for believable than unbelievable problems. Processing differences do not necessitate accuracy differences, however. As the aim of the SDT analysis was to assess accuracy differences, we do not see evidence of processing differences, in and of itself, as being incompatible with a criterion shift account.

An advantage of the SDT approach we have adopted in the current study is that it offers a thorough account of the decision process involved in syllogistic reasoning. As pointed out by Rotello and Heit (2009), one potentially fruitful research strategy would be to complement future processing accounts of reasoning with SDT analyses of decision making, allowing an understanding of both the accumulation of evidence and the way in which decision processes act on the evidence once it is accumulated. In general, SDT analyses can be used to constrain the development of more detailed processing accounts. In addition, processing accounts could be used to make more detailed predictions, for example, about where particular arguments would fall on a scale of argument strength, that could then be used by an SDT account. For the time being, however, it must be concluded that SDT provides the only explanation of the belief bias effect that is consistent with our results.

The SDT-based account also provides an explanation for anomalies related to the behavior of the interaction index that has implications for dual process theories of deduction (Evans, 2006, 2007; Shynkaruk & Thompson, 2006). Dual process theories of reasoning generally posit two reasoning systems: a fast-acting, error-prone heuristic system, which is more heavily influenced by prior knowledge, and a slower, less error-prone analytic system, which more closely follows the rules of logic. In support of this distinction, Evans and Curtis-Holmes (2005) showed that, when subjects were given only 10 s to evaluate a conclusion in the belief bias task, the effect of believability was increased, and the effect of logic reduced, relative to a condition in which no time constraint was imposed. Interestingly, the interaction index was also reduced in the 10-s group, consistent with the idea that the interaction effect reflects the operation of analytical processes that were curtailed by the short deadline. A number of results argue against this interpretation of the interaction, however. First, a pair of studies by Newstead et al. (1992, Experiment 5) and Evans et al. (1994, Experiment 3) demonstrated that when relatively complex instructions stressing the correct interpretation of syllogistic quantifiers or the concept of logical necessity are given, the main effect of belief and the Belief × Logic interaction are greatly reduced. Second, a study by Quayle and Ball (2000) showed that the Belief × Logic interaction is reduced in subjects with relatively high working memory spans. Third, a study by Shynkaruk and Thompson (2006), using a deadline procedure similar to that of Evans and Curtis-Holmes, observed the Belief × Logic interaction (also assessed with $H - F$) in subjects given both a short (10-s) and a long (60-s) response deadline in which speeded responses could be reconsidered and possibly corrected. The size of the interaction effect did not differ by response opportunity, although it was only obtained for subjects with relatively high reasoning accuracy: Poor reasoners showed no interaction at either response opportunity. In reference to the assumption that the interaction reflects analytical processing, the authors wrote,

> under the more probable assumption that the short deadline was sufficient to curtail extensive analysis, one must conclude that the interaction is not due to formal reasoning processes but, rather, arises from the application of fast and simple heuristics, which can be applied in about 10 sec. (Shynkaruk & Thompson, 2006, p. 630)

Although this statement is consistent with the bulk of the results cited above, it still does not explain why the interaction was not observed in the 10-s condition of the study by Evans and Curtis-Holmes, or why it would be obtained only for subjects in the higher accuracy group, who would, intuitively at least, seem less likely to rely on heuristic responding.

An answer is provided once one considers the relationship between accuracy and ROC form. In the conditions of both studies in which the interaction was not obtained, accuracy was at (Shynkaruk & Thompson, 2006) or very near (Evans & Curtis-Holmes, 2005) chance levels. Although the area measure A_z cannot be estimated from the available data, it is likely that these conditions would have yielded ROCs falling at or near the chance line. As chance performance necessarily produces a straight line with unit slope, the isosensitivity relationship for near-chance responding is expected to yield roughly equal estimates of $H - F$ at every level of response bias. Thus, the interaction index is reduced as performance approaches the floor.

Still, at a more general level, our findings are compatible with dual process theories of deduction, if not specific versions such as selective scrutiny or selective processing. In terms of the signal detection account illustrated in Figure 3B, the analytic system may act to determine the position of arguments along the axis of argument strength, while the heuristic system might use knowledge-based cues such as believability to determine the location of the response criterion. Furthermore, our data do not rule out alternative, more complex, signal detection interpretations of dual process theory (e.g., the multidimensional account in Heit & Rotello, 2008, in press; Rotello & Heit, 2009), but these are not required to account for the data from Experiments 1–3.

Finally, we do not feel that our results pose problems for the MPT framework in general but only for those models that generate linear ROCs. It is possible that an MPT approach that differs from the one proposed by Klauer et al. (2000) will eventually provide an adequate description of our data. In the recognition literature, for instance, it has been pointed out that high-threshold models that assume more complex mappings of internal states to rating responses can be used to generate ROCs that more closely approximate the curvilinear functions typically observed for the ratings task (Bröder & Schütz, 2009; Erdfelder & Buchner, 1998; Klauer & Kellen, in press; Malmberg, 2002). Though we did consider one such alternative mapping, in which the nondetect states could lead to high- and low-confidence responses (i.e., MPT-R), we still found that a superior fit was provided by SDT. In any case, the main purpose of the present study was not to document the flexibility of the MPT framework but to evaluate the assumption of ROC linearity implied by $H - F$. For this reason, the mapping used in the present study was chosen to differ minimally from the structure and logic of the double high threshold model proposed by Klauer et al., in which high-confidence errors were assumed to not occur. The close correspondence between the results for MPTK and MPT2 in Experiment 3 suggests that the mapping we adopted was appropriate in that respect. We suspect that, regardless of the framework one applies to the present results, models that provide a good account of the ROC data will agree with the conclusions reached in our SDT model: Changes in bias do not augment H and F by equal amounts. It is for this reason that analyses based on $H - F$ are likely to result in Type I errors, as indicated by Rotello et al. (2008). We have shown that the Belief × Logic interaction is one such error and that several theories of belief bias assuming an interpretation of the data stemming from $H - F$ are affected.

In sum, we feel that alternative models of belief bias should be constructed and compared in future work, whether they involve an SDT or an MPT framework. Alternatives to ratings-based ROCs should also be evaluated as tests of the assumptions made by processing and measurement models (e.g., base-rate ROCs of the sort collected in Experiment 3). We hope that our work will have an effect similar to that of the seminal study by Klauer et al. (2000), supplying an impetus for the construction of quantitative models that will provide an improvement over our SDT model.

Related Findings in the Memory Literature

One's choice of accuracy measure should not be made arbitrarily. As we have shown for the belief bias task, the interpretation of the data depends critically on which measure is used. Following Swets (1986a, 1986b), Rotello et al. (2008) showed that the use of accuracy statistics such as $H - F$, d', A' (Pollack & Norman, 1964), percent correct, and the Goodman-Kruskal gamma correlation (Goodman & Kruskal, 1954; Nelson, 1984) necessarily entails assumptions about the structure of the underlying data. For example, d' assumes that the underlying evidence distributions are equal-variance Gaussian in form; $H - F$ and percent correct are consistent with equal-variance rectangular distributions. When those assumptions are violated, bias differences across experimental conditions are very likely to be misinterpreted as accuracy differences. Worse, Rotello et al.'s simulations showed that the probability of making such a Type I error increases with sample size (i.e., running more subjects or more trials exaggerates the problem). This type of problem is at the heart of previous misinterpretations of the belief bias effect: Use of a measure, $H - F$, whose assumptions are inconsistent with the form of the data, led to the erroneous claim of an accuracy effect.

Type I errors of the sort uncovered in this investigation are not limited to the belief bias literature. A number of similar cases have been identified recently in the memory literature. For example, recognition memory for negatively valenced stimuli is often thought to be better than recognition of neutral stimuli, but ROC-based analyses of both younger and older adults' memory for emotional stimuli have shown that the effect is one of response bias and not accuracy (Dougal & Rotello, 2007; Kapucu, Rotello, Ready, & Seidl, 2008). Similarly, researchers have argued that when subjects claim to remember the experience of studying an item, accuracy is high relative to trials in which subjects claim only to know an item was presented (see Yonelinas, 2002, for a summary). However, modeling work has demonstrated that the difference between remember and know judgments is only one of response bias (e.g., Cohen, Rotello, & Macmillan, 2008; Rotello, Macmillan, Hicks, & Hautus, 2006; Rotello & Zeng, 2008; Verde, Macmillan, & Rotello, 2006).

Another example is the *revelation effect* (Watkins & Peynircioglu, 1990), in which subjects make more "old" responses to recognition memory probes that follow an unrelated revelation task than to probes that are not preceded by such a task; d' is also found to be higher in the revelation condition than the control condition. The revelation task itself can be almost anything (anagram solution, math problems, etc.), which made the effect quite puzzling theoretically. Using ROC analyses, Verde and Rotello (2003, 2004) showed that the use of d' was not justified by the data, which were consistent with unequal-variance Gaussian evidence distributions. They concluded that the revelation effect is usually just a response bias effect: Subjects respond more liberally in the revelation condition.

Finally, in the domain of metacognition, Weaver and Kelemen (2003) rejected their preferred theory on the basis of results from a test using gamma correlations. Masson and Rotello (2009) showed that the gamma statistic has unfortunate properties that render it a poor measure of performance. In particular, although gamma has been marketed as a nonparametric statistic (Nelson, 1984, 1986), its value is not independent of response bias. An ROC-based reanalysis of Weaver and Kelemen's data indicated that the previously rejected theory was actually well supported by the data.

Broader Implications for Reasoning and Related Areas of Cognitive Psychology

Because the belief bias effect on syllogistic reasoning is so well documented, it has been investigated more broadly in the reasoning literature, in application to other issues. For example, Stanovich and West (1998) examined the belief bias effect along with performance on several standardized reasoning tests and found that the likelihood of avoiding belief bias is associated with individual differences in general cognitive ability. Norenzayan, Smith, Jun Kim, and Nisbett (2002) used belief bias measures to study cultural differences in reasoning, namely, whether Western thinkers are more likely than others to favor formal reasoning and hence are less likely to show a belief bias. In the context of the Norenzayan et al. study, Unsworth and Medin (2005) raised the question of whether apparent cross-cultural differences might simply reflect a response bias rather than true differences in the use of logic. More generally, having a better understanding of the belief bias effect in syllogistic reasoning, including better measures of the effect, would help to support applications such as the study of individual differences or cross-cultural differences. For example, knowing whether belief bias truly corresponds to lower accuracy on some materials than other materials or is a response bias by nature has implications for the interpretation of whatever results are found.

The belief bias effect on syllogistic reasoning has gained much of its importance because it is a well-studied example of the influence of top-down knowledge, a topic that pervades cognitive psychology. In syllogistic reasoning, specifically, following one's beliefs rather than the rules of logic is an error, but in the more general case of reasoning under uncertainty, it is normative to consider other knowledge (Heit, Hahn, & Feeney, 2005; Skyrms, 2000). Indeed, studies of inductive or probabilistic reasoning have documented widespread and systematic effects of beliefs. For example, experts in a given domain typically use knowledge specific to that domain to override default patterns of inductive reasoning that would be used by nonexperts (see Hayes, Heit, & Swendsen, in press, and Shafto, Coley, & Vitkin, 2007, for reviews). Potentially, once better analytical tools and better theoretical accounts are developed to address belief bias in syllogistic reasoning, these could be applied to effects of beliefs on other forms of reasoning. We place priority on developing theoretical accounts of reasoning that span different reasoning tasks and paradigms, rather than just narrowly addressing syllogistic reasoning. For instance, recent work by Heit and Rotello (2005, 2008, in press) and Rotello and Heit (2009) has applied signal detection models to inductive reasoning to examine the effects of related variables that differentially cue the application of real-world knowledge.

The effects of prior beliefs are also pervasive in tasks that do not explicitly seek to investigate reasoning (e.g., memory; Heit, 1993). On this point, Heit (1997) argued that models of reasoning ideally will dovetail with models of other cognitive abilities such as memory and categorization. For instance, there is a tradition in memory research going back at least to Bartlett (1932) that concerns itself with the influences of prior beliefs (or schemas) on memory. Heit (1993) conducted simulations to examine the effects of prior beliefs on recognition memory by comparing recognition of stereotype-congruent versus stereotype-incongruent stimuli (person descriptions). Although the methodology employed by Heit differed from that of the present study, he concluded that an apparent advantage in recognition accuracy for incongruent stimuli was not due to differential processing of congruent and incongruent stimuli (i.e., selective weighting, or distortion, of one type of stimuli or the other). Instead, his favored explanation of the effect of prior beliefs on recognition was that it was a simple response bias effect, with memory traces corresponding to prior beliefs having a fixed, positive effect on the familiarity of stereotype-congruent stimuli, leading to an increased level of false alarms on those stimuli.

In categorization research, there is also an important area of study investigating the influences of wider beliefs on concept learning (Murphy & Medin, 1985). Heit (1994, 2001) assessed several experiments showing effects of prior beliefs on categorization and again concluded that these were best explained as a kind of a response bias, with memory traces corresponding to prior beliefs having a fixed effect on category representations, rather than as differential processing of category members depending on whether or not they fit prior beliefs. Thus, prior beliefs may have similar effects in reasoning, memory, and categorization, and people do not necessarily process syllogistic arguments, stimuli to be memorized, or category members to be learned differently depending on whether they fit prior beliefs. The work of Heit (1993, 1994, 2001) did not apply multidimensional signal detection models, but such models (without the prior belief component) have been used to account for both memory (e.g., Banks, 2000; Hautus, Macmillan, & Rotello, 2008; Rotello, Macmillan, & Reeder, 2004) and categorization phenomena (e.g., Ashby & Gott, 1988; Maddox & Dodd, 2003). Future work could extend these models to include an explicit role for prior belief in these domains.

Beyond the specific issue of the belief bias effect, our findings have wider implications for models of reasoning. One important research program (Oaksford & Chater, 2007) has used an account based on Bayesian probability theory to address several tasks (syllogistic reasoning, conditional reasoning, and the selection task) that have traditionally been the subject of theories of deduction. In effect, this work extends standard logic to probabilities, where the premises and conclusions of arguments can have some level of uncertainty. Chater and Oaksford (1999) presented the probability heuristics model (PHM) of syllogistic reasoning, which has at its center the notion of probabilistic validity. By taking this position, the PHM is able to address not only the use of traditional syllogistic quantifiers ("all," "no," "some," "some . . . are not") but also generalized quantifiers ("most," "few") by treating these as probabilistic statements. In a similar way, the PHM assumes that a conclusion derived from a given set of premises has a probability attached to it. Thus, the PHM differs markedly from the other accounts of syllogistic reasoning that we have reviewed because it characterizes the output of the reasoning process as a continuously distributed value (a probability) rather than as a discrete (valid or invalid) judgment or mental state.[9] Although the PHM differs from our own SDT account in many ways, the models share the central notion that the reasoning process results in a continuous argument-strength value. Our own analyses (e.g., the finding of curvilinear rather than linear ROCs; see also Rotello & Heit, 2009) strongly support this idea, and thus, they also support Chater and Oaksford's PHM.

We have not applied the PHM to our own results because it does not address the belief bias effect directly. Chater and Oaksford (1999) suggested that belief bias is a matter of everyday reasoning strategies rather than something intrinsic to syllogistic reasoning itself. Although we think it is desirable to model the everyday reasoning processes that lead to the belief bias effect, our finding that belief bias can be explained as a form of response bias applied to argument-strength values is roughly compatible with the PHM approach.

Finally, our work suggests new tools for assessing the calibration of confidence and accuracy in deductive reasoning (Prowse-Turner & Thompson, 2009; Shynkaruk & Thompson, 2006). For instance, Prowse-Turner and Thompson (2009) showed that while reasoners tend to be poorly calibrated (overconfident) for the syllogistic reasoning task, calibration can be much improved with training that involves trial-by-trial accuracy feedback and instruction in the concept of logical necessity. Although the analytical approach taken by the authors was quite rigorous, accuracy was assessed in some cases by using percent correct. As we have shown, ROC curves can be used to obtain more appropriate measures of performance such as A_z and are readily available when confidence ratings are collected. ROCs based on metacognitive assessments of performance (rating confidence in the correctness of a response; Type II ROCs) can also be used to provide area-based measures that estimate monitoring accuracy (Galvin, Podd, Drga, & Whitmore, 2003; Higham, Perfect, & Bruno, 2009). We hope future work that expands on the foundation laid by Prowse-Turner and Thompson will benefit from supplementary measures of reasoning and monitoring accuracy that can be obtained with ROCs.

Summary

In sum, the present study has demonstrated for the first time that ROCs for the belief bias task are not linear, meaning that accuracy is best measured using A_z. Traditional analyses that contrast acceptance rates for valid and invalid problems at each level of believability are likely to produce Type I errors, such as the artifactual interaction between validity and believability that we

[9] The PHM suggests that the weakest conclusions are those associated with the "some . . . are not" quantifier (Chater & Oaksford, 1999). Thus, the relatively low accuracy levels observed in our experiments may reflect our use of that quantifier in the conclusion (see Appendix C).

demonstrated in two experiments. Our results suggest that none of the current theories of belief bias are satisfactory because all of them explain the interaction between validity and believability in terms of an accuracy difference rather than the bias difference that we have documented. We advance a new model that provides the best quantitative fit to the data in each experiment; it characterizes belief bias as a response bias effect that operates on the positioning of subjects' response criteria. Though some readers may find this conclusion to be unbelievable, we feel confident that a thorough consideration of the argument we have presented will nonetheless compel its endorsement.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Associated Press. (2009, May 20). Pringles are potato chips, UK taxman decides. *USA Today.* Retrieved from http://www.usatoday.com/money/industries/food/2009–05-20-pringles_N.htm

Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology, 53,* 77–86.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11,* 267–273.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, England: Cambridge University Press.

Begg, I., & Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology, 81,* 351–354.

Benjamin, A. S., Diaz, M., & 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.

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

Cherubini, P., Garnham, A., Oakhill, J., & Morley, E. (1998). Can any ostrich fly? Some new data on belief bias in syllogistic reasoning. *Cognition, 69,* 179–218.

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.

Creelman, C. D., & Donaldson, W. (1968). ROC curves for discrimination of linear extent. *Journal of Experimental Psychology, 77,* 514–516.

Dickstein, L. S. (1975). Effects of instructions and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 376–384.

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory & Cognition, 6,* 76–83.

Dickstein, L. S. (1981). Conversion and possibility in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18,* 229–232.

Dodson, C. S., Prinzmetal, W., & Shimamura, A. P. (1998). Using Excel to estimate parameters from observed data: An example from source memory data. *Behavior Research Methods, Instruments, & Computers, 30,* 517–526.

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

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

Emmerich, D. S. (1968). ROCs obtained with two signal intensities presented in random order, and a comparison between yes-no and rating ROCs. *Perception & Psychophysics, 3,* 35–40.

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. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13,* 378–395.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement.* New York, NY: Psychology Press.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

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., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11,* 382–389.

Evans, J. St. B. T., Handley, S. J., & Harper, C. N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 54*(A), 935–958.

Evans, J. St. B. T., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology, 6,* 263–285.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hillsdale, NJ: Erlbaum.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10,* 843–876.

Gilinsky, A. S., & Judd, B. B. (1994). Working memory and bias in reasoning across the life span. *Psychology and Aging, 9,* 356–371.

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.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–764.

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

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.

Hayes, B. K., Heit, E., & Swendsen, H. (in press). Inductive reasoning. *Cognitive Science.*

Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task? *Memory & Cognition, 3,* 233–238.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1210–1230.

Heit, E. (1993). Modeling the effects of expectations on recognition memory. *Psychological Science, 4,* 244–252.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1264–1282.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D.

Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7–41). London, England: Psychology Press.

Heit, E. (2001). Background knowledge and models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 155–178). Oxford, England: Oxford University Press.

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W. Ahn, R. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 87–99). Washington, DC: American Psychological Association.

Heit, E., & Rotello, C. M. (2005). Are there two kinds of reasoning? In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society* (pp. 923–928). Hillsdale, NJ: Erlbaum.

Heit, E., & Rotello, C. M. (2008). Modeling two kinds of reasoning. In B. C. Love, K. McRae, & V. M. Sloutsky, *Proceedings of the Thirtieth Annual Meeting of the Cognitive Science Society* (pp. 1831–1836). Austin, TX: Cognitive Science Society.

Heit, E., & Rotello, C. M. (in press). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using Type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 57–80.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16,* 1–61.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology, 10,* 64–99.

Kapucu, A., Rotello, C. M., Ready, R. E., & Seidl, K. N. (2008). Response bias in "remembering" emotional stimuli: A new perspective on age differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 703–711.

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.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Krantz, D. (1969). Threshold theories of signal detection. *Psychological Review, 76,* 308–324.

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

Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of Gaussian ROC statistics. *Perception & Psychophysics, 66,* 406–421.

Maddox, W. T., & Dodd, J. L. (2003). Separating perceptual and decisional attention processes in the identification and categorization of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 467–480.

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

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17,* 11–17.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509–527.

Metz, C. E. (1998). ROCKIT computer program. Chicago, IL: University of Chicago, Department of Radiology. Available from http://xray.bsd.uchicago.edu/cgi-bin/roc_software.cgi

Morley, N. J., Evans, J. St. B. T., & Handley, S. J. (2004). Belief bias and figural bias in syllogistic reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 57*(A), 666–692.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Myung, I. J., Pitt, M. A., & Kim, W. (2003). Model evaluation, testing, and selection. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 422–436). London, England: Sage.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100,* 128–132.

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

Norenzayan, A., Smith, E. E., Jun Kim, B., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science, 26,* 653–684.

Oakhill, J., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31,* 117–140.

Oakhill, J. V., & Johnson-Laird, P. (1985). The effects of belief on the spontaneous production of syllogistic conclusions. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 37*(A), 553–569.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford, England: Oxford University Press.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review, 102,* 533–566.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science, 1,* 125–126.

Prowse-Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning, 15,* 69–100.

Quayle, J., & Ball, L. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 53*(A), 1202–1223.

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

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

Revlin, R., Leirer, V., Yopp, H., & Yopp, R. (1980). The belief bias effect in formal reasoning: The influence of knowledge on logic. *Memory & Cognition, 8,* 584–592.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior, 14,* 180–195.

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 305–320.

Roberts, M. J., & Sykes, E. D. A. (2003). Belief bias and relational reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 56*(A), 131–154.

Rotello, C. M., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1317–1330.

Rotello, C. M., & Macmillan, N. A. (2008). Response bias in recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use* (pp. 61–94). San Diego, CA: Elsevier.

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

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review, 111,* 588–616.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70,* 389–401.

Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember-know paradigm. *Psychonomic Bulletin & Review, 15,* 825–832.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Shafto, P., Coley, J. D., & Vitkin, A. (2007). Availability in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 114–136). Cambridge, England: Cambridge University Press.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34,* 619–632.

Skyrms, B. (2000). *Choice and chance: An introduction to inductive logic* (4th ed.). Belmont, CA: Wadsworth.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188.

Stupple, E. J. N., & Ball, L. J. (2008). Belief-logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning, 14,* 168–181.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Tanner, T. A. J., Haller, R. W., & Atkinson, R. C. (1967). Signal recognition as influenced by presentation schedules. *Perception & Psychophysics, 2,* 349–358.

Tanner, T. A. J., Rauk, J. A., & Atkinson, R. C. (1970). Signal recognition as influenced by information feedback. *Journal of Mathematical Psychology, 7,* 259–274.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review, 10,* 184–189.

Unsworth, S. J., & Medin, D. L. (2005). Cross cultural differences in belief bias with deductive reasoning? *Cognitive Science, 29,* 525–529.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics, 68,* 643–654.

Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 739–746.

Verde, M. F., & Rotello, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review, 11,* 560–566.

Watkins, M. J., & Peynircioglu, Z. F. (1990). The revelation effect: When disguising test items induces recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1012–1020.

Weaver, C. A., III, & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1058–1065.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18,* 451–460.

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.

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

(*Appendices follow*)

Appendix A

Model Equations for Experiments 1–3

MPTK Model

Valid Problems

$$P(\text{“Valid”}|\text{Believable}) = r_{vb} + (1 - r_{vb})\beta_b\alpha_x.$$

$$P(\text{“Valid”}|\text{Unbelievable}) = r_{vu} + (1 - r_{vu})\beta_u\alpha_x.$$

Invalid Problems

$$P(\text{“Valid”}|\text{Believable}) = (1 - r_{ib})\beta_b\alpha_x.$$

$$P(\text{“Valid”}|\text{Unbelievable}) = (1 - r_{iu})\beta_u\alpha_x.$$

Note: The parameter α_x is a response bias parameter estimating P(“Valid”) where x is defined by the base-rate condition in Experiment 3 (x = low, medium, or high). With the exception of the x subscript, the four equations are identical across the three conditions.

MPT-R Model

Valid Problems

$$P(\text{“1”}) = r_{vy} + (1 - r_{vy})\beta_y\alpha_{y1}.$$

$$P(\text{“2”}) = (1 - r_{vy})\beta_y\alpha_{y2}.$$

$$P(\text{“3”}) = (1 - r_{vy})\beta_y(1 - \alpha_{y1} - \alpha_{y2}).$$

$$P(\text{“4”}) = (1 - r_{vy})(1 - \beta_y)(1 - \alpha_{y5} - \alpha_{y6}).$$

$$P(\text{“5”}) = (1 - r_{vy})(1 - \beta_y)\alpha_{y5}.$$

$$P(\text{“6”}) = (1 - r_{vy})(1 - \beta_y)\alpha_{y6}.$$

Invalid Problems

$$P(\text{“1”}) = (1 - r_{iy})\beta_y\alpha_{y1}.$$

$$P(\text{“2”}) = (1 - r_{iy})\beta_y\alpha_{y2}.$$

$$P(\text{“3”}) = (1 - r_{iy})\beta_y(1 - \alpha_{y1} - \alpha_{y2}).$$

$$P(\text{“4”}) = (1 - r_{iy})(1 - \beta_y)(1 - \alpha_{y5} - \alpha_{y6}).$$

$$P(\text{“5”}) = (1 - r_{iy})(1 - \beta_y)\alpha_{y5}.$$

$$P(\text{“6”}) = r_{iy} + (1 - r_{iy})(1 - \beta_y)\alpha_{y6}.$$

Note: The subscript y varies with the believability of the problem conclusion (b = believable, u = unbelievable). These equations correspond to MPT-R applied to the belief bias data from Experiments 2 and 3. MPT-R applied to the data from Experiment 1 uses the equations above, except that the subscripts b and u are replaced with the subscripts l and c, corresponding to the liberal and conservative conditions of that experiment. MPT-R applied to base-rate data from Experiment 3 is the same as in Experiment 1 but uses three levels of bias (l, c, n) corresponding to the liberal, conservative, and neutral conditions of that experiment.

MPT1 Model

Valid Problems

$$P(\text{“1”}) = r_{vy}[1 - P(\text{“6”})].$$

$$P(\text{“2”}) = (1 - r_{vy})\beta_y\alpha_2[1 - P(\text{“6”})].$$

$$P(\text{“3”}) = (1 - r_{vy})\beta_y(1 - \alpha_2)[1 - P(\text{“6”})].$$

$$P(\text{“4”}) = (1 - r_{vy})(1 - \beta_y)\alpha_4[1 - P(\text{“6”})].$$

$$P(\text{“5”}) = (1 - r_{vy})(1 - \beta_y)(1 - \alpha_4)[1 - P(\text{“6”})].$$

$$P(\text{“6”}) = .0000001.$$

Invalid Problems

$$P(\text{“1”}) = .0000001.$$

$$P(\text{“2”}) = (1 - r_{iy})\beta_y\alpha_2[1 - P(\text{“1”})].$$

$$P(\text{“3”}) = (1 - r_{iy})\beta_y(1 - \alpha_2)[1 - P(\text{“1”})].$$

$$P(\text{“4”}) = (1 - r_{iy})(1 - \beta_y)\alpha_4[1 - P(\text{“1”})].$$

$$P(\text{“5”}) = (1 - r_{iy})(1 - \beta_y)(1 - \alpha_4)[1 - P(\text{“1”})].$$

$$P(\text{“6”}) = r_{iy}[1 - P(\text{“1”})].$$

Note: The subscript y varies with the believability of the problem conclusion (b = believable, u = unbelievable). These equations correspond to MPT1 applied to the belief bias data from Experiments 2 and 3. MPT1 applied to the data from Experiment 1 uses the equations above, except that the subscripts b and u are replaced with the subscripts l and c, corresponding to the liberal and conservative conditions of that experiment. MPT1 applied to base-rate data from Experiment 3 is the same as MPT3 (detailed below), except that the parameter ε is constrained to equal 0 in the equations and $P(\text{“6”}|\text{Valid}) = P(\text{“1”}|\text{Invalid}) = .0000001$.

MPT2 Model

Valid Problems

$$P(\text{“1”}) = r_{vy}(1 - \varepsilon).$$

$$P(\text{“2”}) = (1 - r_{vy})\beta_y\alpha_2(1 - \varepsilon).$$

$$P(\text{“3”}) = (1 - r_{vy})\beta_y(1 - \alpha_2)(1 - \varepsilon).$$

$$P(\text{“4”}) = (1 - r_{vy})(1 - \beta_y)\alpha_4(1 - \varepsilon).$$

$$P(\text{“5”}) = (1 - r_{vy})(1 - \beta_y)(1 - \alpha_4)(1 - \varepsilon).$$

$$P(\text{“6”}) = \varepsilon.$$

Invalid Problems

$P(“1”) = \varepsilon.$

$P(“2”) = (1 - r_{iy})\beta_y\alpha_2(1 - \varepsilon).$

$P(“3”) = (1 - r_{iy})\beta_y(1 - \alpha_2)(1 - \varepsilon).$

$P(“4”) = (1 - r_{iy})(1 - \beta_y)\alpha_4(1 - \varepsilon).$

$P(“5”) = (1 - r_{iy})(1 - \beta_y)(1 - \alpha_4)(1 - \varepsilon).$

$P(“6”) = r_{iy}(1 - \varepsilon).$

Note: The subscript *y* varies with the believability of the problem conclusion (*b* = believable, *u* = unbelievable). These equations correspond to MPT2 applied to the belief bias data from Experiments 2 and 3. MPT2 applied to the data from Experiment 1 uses the equations above, except that the subscripts *b* and *u* are replaced with the subscripts *l* and *c,* corresponding to the liberal and conservative conditions of that experiment.

MPT3 Model

Valid Problems

$P(“1”) = r_{vx}(1 - \varepsilon).$

$P(“2”) = (1 - r_{vx})\beta_x\alpha_2(1 - \varepsilon).$

$P(“3”) = (1 - r_{vx})\beta_x(1 - \alpha_2)(1 - \varepsilon).$

$P(“4”) = (1 - r_{vx})(1 - \beta_x)\alpha_4(1 - \varepsilon).$

$P(“5”) = (1 - r_{vx})(1 - \beta_x)(1 - \alpha_4)(1 - \varepsilon).$

$P(“6”) = \varepsilon.$

Invalid Problems

$P(“1”) = \varepsilon.$

$P(“2”) = (1 - r_{ix})\beta_x\alpha_2(1 - \varepsilon).$

$P(“3”) = (1 - r_{ix})\beta_x(1 - \alpha_2)(1 - \varepsilon).$

$P(“4”) = (1 - r_{ix})(1 - \beta_x)\alpha_4(1 - \varepsilon).$

$P(“5”) = (1 - r_{ix})(1 - \beta_x)(1 - \alpha_4)(1 - \varepsilon).$

$P(“6”) = r_{ix}(1 - \varepsilon).$

Note: The subscript *x* indexes the base-rate condition in Experiment 3 (*x* = low, medium, or high).

SDT Model

Valid Believable Problems

$P(“1”) = \Phi[(\mu_{vb} - c_{1b})/\sigma_{vb}].$

$P(“2”) = \Phi[(\mu_{vb} - c_{2b})/\sigma_{vb}] - \Phi[(\mu_{vb} - c_{1b})/\sigma_{vb}].$

$P(“3”) = \Phi[(\mu_{vb} - c_{3b})/\sigma_{vb}] - \Phi[(\mu_{vb} - c_{2b})/\sigma_{vb}].$

$P(“4”) = \Phi[(\mu_{vb} - c_{4b})/\sigma_{vb}] - \Phi[(\mu_{vb} - c_{3b})/\sigma_{vb}].$

$P(“5”) = \Phi[(\mu_{vb} - c_{5b})/\sigma_{vb}] - \Phi[(\mu_{vb} - c_{4b})/\sigma_{vb}].$

$P(“6”) = \Phi[(c_{5b} - \mu_{vb})/\sigma_{vb}].$

Invalid Believable Problems

$P(“1”) = \Phi(-c_{1b}).$

$P(“2”) = \Phi(-c_{2b}) - \Phi(-c_{1b}).$

$P(“3”) = \Phi(-c_{3b}) - \Phi(-c_{2b}).$

$P(“4”) = \Phi(-c_{4b}) - \Phi(-c_{3b}).$

$P(“5”) = \Phi(-c_{5b}) - \Phi(-c_{4b}).$

$P(“6”) = \Phi(c_{5b}).$

Note: In the signal detection theory equations, $\Phi(z)$ returns a value $P(z)$ on the inverse standard normal cumulative distribution function for a value *z.* The equations for unbelievable arguments are the same but are subscripted with *u* instead of *b.* The same set of equations applies to the data of all three experiments, the only differences being whether parameters are subscripted according to conclusion believability (Experiments 2 and 3) or bias condition (Experiments 1 and 3). In the case of Experiment 3, an extra set of 12 equations with subscript *n* was used to fit data from the neutral condition.

(*Appendices continue*)

Appendix B

Fit Statistics for the MPT1 and MPT-R Models in Experiments 1–3

Table B1

		MPT1			MPT-R		
Experiment	Condition	$G^2_{5\text{df}}$	AIC	BIC	$G^2_{3\text{df}}$	AIC	BIC
1	Liberal	6,467.39	10,020.91	10,046.29	60.22	3,617.63	3,653.16
	Conservative	4,271.39	7,792.41	7,817.37	100.38	3,625.40	3,660.35
2	Believable	3,023.75	4,758.48	4,780.53	23.41	1,762.14	1,793.01
	Unbelievable	2,279.59	4,163.07	4,185.12	26.16	1,913.64	1,944.51
3	Believable	9,869.44	15,790.33	15,818.19	35.63	5,960.51	5,999.52
	Unbelievable	8,667.65	14,859.91	14,887.77	83.96	6,280.22	6,319.23
	Liberal	6,392.14	10,294.47	10,320.30	28.82	3,935.15	3,971.32
	Conservative	4,175.09	7,879.41	7,904.58	38.74	3,747.06	3,782.30
	Neutral	7,982.56	12,440.09	12,466.51	21.84	4,483.37	4,520.37

Note. MPT-R allows responses from the nondetect state to map onto every level of confidence, thus producing curved receiver operating characteristics that approximate those generated by SDT. This model is therefore a major departure from the MPT model proposed by Klauer, Musch, and Naumer (2000) and is not useful for testing the linearity assumption. Nonetheless, we reached the same conclusions with this model as we did in our other MPT analyses and consistently obtained poorer fits for MPT-R than for SDT. For comparison, fit statistics for SDT can be found in Tables 4, 6, and 12 in the main text. MPT = multinomial processing tree; MPT1 = Multinomial Processing Tree 1 model; MPT-R = multinomial processing tree model with confidence rating; SDT = signal detection theory model; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Appendix C

Conclusion Ratings for Materials Used in Experiments 2 and 3

Table C1

Believable	*M*	*SD*	Unbelievable	*M*	*SD*
Some animals are not llamas.	4.55	1.21	Some llamas are not animals.	1.00	0.00
Some bears are not grizzlies.	4.75	0.84	Some grizzlies are not bears.	1.52	1.21
Some birds are not parrots.	4.68	1.06	Some parrots are not birds.	1.19	0.79
Some boats are not canoes.	4.35	1.31	Some canoes are not boats.	1.86	1.56
Some cars are not oldsmobiles.	4.19	1.56	Some oldsmobiles are not cars.	1.43	0.96
Some criminals are not robbers.	4.61	1.05	Some robbers are not criminals.	2.11	1.59
Some dances are not tangos.	4.68	0.90	Some tangos are not dances.	1.65	1.23
Some drinks are not beers.	4.82	0.77	Some beers are not drinks.	1.58	1.36
Some horses are not ponies.	3.68	1.63	Some ponies are not horses.	2.42	1.78
Some insects are not spiders.	4.58	1.09	Some spiders are not insects.	2.07	1.56
Some killers are not assassins.	3.96	1.69	Some assassins are not killers.	1.32	0.79
Some plants are not weeds.	4.52	1.15	Some weeds are not plants.	2.29	1.58
Some relatives are not uncles.	4.84	0.73	Some uncles are not relatives.	2.29	1.67
Some reptiles are not lizards.	4.39	1.29	Some lizards are not reptiles.	1.48	1.06
Some storms are not blizzards.	4.86	0.76	Some blizzards are not storms.	1.55	1.15
Some trees are not oaks.	4.55	1.23	Some oaks are not trees.	1.96	1.50
Some weapons are not cannons.	4.61	1.17	Some cannons are not weapons.	2.61	1.73
Some words are not verbs.	4.86	0.76	Some verbs are not words.	1.55	1.36
Some writers are not novelists.	4.79	0.79	Some novelists are not writers.	1.84	1.49

Appendix D

Problem Structures Used in Experiments 1–3

A

Set A		Set B	
Valid	Invalid	Valid	Invalid
EI2_O1	EI2_O2	OA2_O2	OE2_O2
EI3_O1	EI3_O2	AO2_O1	EO2_O1
EI4_O1	EI4_O2	OA3_O1	OE3_O1
IE4_O2	IE4_O1	AO3_O2	EO3_O2

B

No X are Y.
Some Z are Y.
____
Some Z are not X.

Figure D1. A: Set A includes problems that minimize figure, atmosphere, and conversion effects. Set B includes problems that minimize figure and atmosphere effects. Structures are identified by quantifiers used in the premises, where A = all, E = no, I = some, and O = some . . . are not. The first letter corresponds to the quantifier of the first premise, and the third letter corresponds to the conclusion. Following the quantifiers for the two premises is a number corresponding to figure, and following the quantifier for the conclusion is a number corresponding to the ordering of conclusion terms. A 1 indicates a conclusion in the Z-X direction and a 2 indicates a conclusion in the X-Z direction. B: Using this notation, the above example would be syllogism EI2_O1.

Received August 21, 2009
Revision received February 26, 2010
Accepted February 27, 2010 ■