Cognition 131 (2014) 75–91

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Traditional difference-score analyses of reasoning are flawed

Evan Heit [a,*], Caren M. Rotello [b]

[a] University of California, Merced, United States
[b] University of Massachusetts Amherst, United States

ARTICLE INFO

Article history:
Received 2 April 2013
Revised 10 December 2013
Accepted 10 December 2013

Keywords:
Reasoning
Deduction
Syllogistic reasoning
Conditional reasoning
Belief bias
Signal detection theory

ABSTRACT

Studies of the belief bias effect in syllogistic reasoning have relied on three traditional difference score measures: the logic index, belief index, and interaction index. Dube, Rotello, and Heit (2010, 2011) argued that the interaction index incorrectly assumes a linear receiver operating characteristic (ROC). Here, all three measures are addressed. Simulations indicated that traditional analyses of reasoning experiments are likely to lead to incorrect conclusions. Two new experiments examined the role of instructional manipulations on the belief bias effect. The form of the ROCs violated assumptions of traditional measures. In comparison, signal detection theory (SDT) model-based analyses were a better match for the form of the ROCs, and implied that belief bias and instructional manipulations are predominantly response bias effects. Finally, reanalyses of previous studies of conditional reasoning also showed non-linear ROCs, violating assumptions of traditional analyses. Overall, reasoning research using traditional measures is at risk of drawing incorrect conclusions.

Published by Elsevier B.V.

1. Introduction

One of the central research issues in cognition is how prior beliefs are put together with new observations. For example, this issue arises in perception (e.g., Schyns & Oliva, 1999), memory (e.g., Bartlett, 1932), comprehension (e.g., Bransford & Johnson, 1972), categorization (e.g., Heit & Bott, 2000), social cognition (e.g., Sherman et al., 2008) and contingency judgment by humans as well as animals (e.g., Alloy & Tabachnik, 1984). Here our focus is reasoning. Broadly speaking, when reasoning is uncertain, it is normative to take account of prior beliefs, indeed any knowledge, in an effort to improve inferences (Skyrms, 2000; see also Heit, Hahn, & Feeney, 2005). However, when the task is to reason according to standard rules of logic, it is normative to focus on the form of an argument only, and not how it connects with other knowledge. For example, in typical studies of syllogistic reasoning, participants are explicitly instructed to focus on whether the conclusion logically follows from the premises. Researchers can then measure how prior beliefs, despite instructions, influence reasoning (e.g., Evans, Barston, & Pollard, 1983; Oakhill & Johnson-Laird, 1985).

One result of this research strategy is the *belief bias effect*, which is the tendency for conclusions of syllogisms to be accepted when they are consistent with prior beliefs, regardless of their validity. For example, Evans et al. (1983) found that syllogisms with invalid, but believable conclusions, like

No addictive things are inexpensive.
Some cigarettes are inexpensive.
*Therefore, some addictive things are not cigarettes. (1)

were judged to be "valid" 71% of the time. In contrast, structurally identical invalid problems with unbelievable conclusions, such as

* Corresponding author. Address: School of Social Sciences, Humanities and Arts, University of California, Merced, 5200 North Lake Road, Merced, CA 95343, United States. Tel.: +1 2092284334.
E-mail address: eheit@ucmerced.edu (E. Heit).

0010-0277/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.cognition.2013.12.003

No cigarettes are inexpensive.
Some addictive things are inexpensive. (2)
*Therefore, some cigarettes are not addictive.

were accepted only 10% of the time. Evans et al. also observed a smaller discrepancy in the acceptance rates for logically valid problems with believable and unbelievable conclusions (89% and 56%, respectively). The different sizes of the belief effect for valid and invalid problems resulted in a statistically reliable interaction between the validity of the conclusion and its believability. The three basic effects—higher acceptance rates for valid than invalid conclusions, higher acceptance rates for believable than unbelievable conclusions, and an interaction between validity and believability—have been studied extensively. Evans et al. referred to these three key measures as the logic index, the belief index, and the interaction index.

Most researchers have measured belief bias effects using a 2×2 ANOVA on the raw scores. A convincing belief bias effect is observed whenever both main effects and the interaction are statistically significant. As reviewed by Dube, Rotello, and Heit (2010), this work has served as the empirical basis for three decades of research on reasoning. Historically, the interaction effect has been taken to support important theories of reasoning (e.g., dual-process theory, Evans & Curtis-Holmes, 2005; mental models theory, Oakhill & Johnson-Laird, 1985; see also Evans et al., 1983; Klauer, Musch, & Naumer, 2000; Newstead, Pollard, Evans, & Allen, 1992; Polk & Newell, 1995; Quayle & Ball, 2000). Assessment of the interaction index continues to be a key point in recent studies of belief bias on syllogistic reasoning in a variety of arenas (e.g., neuroscience, Stollstorff, Bean, Anderson, Devaney, & Vaidya, 2013; emotion and cognition, Blanchette & Campbell, 2012; Eliades, Mansell, Stewart, & Blanchette, 2012; Goel & Vartanian, 2011; individual differences, Stupple, Ball, Evans, & Kamal-Smith, 2011; informal argumentation, Thompson & Evans, 2012).

The logic effect is also a matter of extensive interest in reasoning research, beyond syllogistic reasoning tasks. For example, Pollard and Evans (1987) proposed that the logic index based on raw difference scores (logically correct answers minus logically incorrect answers) should be used to analyze performance on the Wason (1968) selection task. This proposal has been influential (e.g., Griggs, 1989; Platt & Griggs, 1993; Stanovich & West, 2008). The logic index has also been used extensively in studies of conditional reasoning (e.g., Evans, Legrenzi, & Girotto, 1999; Sellen, Oaksford, & Gray, 2005). Therefore, the critiques in this paper of the logic index apply not only to syllogistic reasoning but to the selection task and conditional reasoning as well. Similarly, the belief index has also been used to study belief bias effects in conditional reasoning (e.g., Evans, Handley, & Bacon, 2009; Handley, Capon, Beveridge, Dennis, & Evans, 2004).

What the aforementioned studies have in common is that they rely on analyses of simple difference scores and interactions. A few experiments have used these measures to investigate the important topic of whether the belief bias effect can be reduced or eliminated intentionally (Evans, Newstead, Allen, & Pollard, 1994; Newstead et al., 1992). In other words, can an experimenter's instructions lead a participant to avoid using prior beliefs when evaluating logical validity? This is an important theoretical question because it addresses a core issue in dual-process accounts of reasoning, namely whether automatic processes can be inhibited or substituted with more controlled processes (referred to as an *intervention* by Evans, 2008, and an *override* by Stanovich, 2009). In an experiment with syllogisms, Newstead et al. found that highly detailed instructions eliminated both the belief effect and the interaction effect. In contrast, two of the three experiments on syllogisms reported by Evans et al. (1994) found no reduction in the belief effect or the interaction effect. Their favored explanation for the inconsistent results focused on stimulus and instruction effects. But another possibility is that the traditional measures they considered have a tendency to lead to distorted or unreliable conclusions.

Dube et al. (2010) raised a related concern. We showed that the theoretical *receiver operating characteristic* (ROC) curves—which plot correct response rates (*hits*, H) against error response rates (*false alarms*, F) as a function of changing response bias but constant accuracy level—implied by traditional measures are linear (see also Macmillan & Creelman, 2005, p. 13; Swets, 1986, p. 111). In contrast, the empirical ROCs obtained in reasoning tasks, including both syllogistic belief bias and inductive reasoning, are curved and therefore inconsistent with the assumptions of the raw score approach (Dube, Rotello, & Heit, 2011; Dube et al., 2010; Heit & Rotello, 2005; Heit & Rotello, 2008; Heit & Rotello, 2010; Heit & Rotello, 2012; Heit, Rotello, & Hayes, 2012; Rotello & Heit, 2009; Trippas, Handley, & Verde, 2013). Note that in Dube et al. (2010, 2011) we focused on the interaction index and did not consider potential problems with the logic index or the belief index that are addressed here for the first time.

Applying a measurement statistic, like a difference between acceptance rates, has been shown to result in a high probability of the data being misinterpreted if the assumptions of that measure are not met. This point has been made often in memory research (e.g., Evans, Rotello, Li, & Rayner, 2009; Masson & Rotello, 2009; Verde & Rotello, 2003; Wixted & Mickes, 2012). For example, what are actually response bias differences between two experimental conditions may be falsely interpreted as accuracy differences. That negative consequence of violated assumptions cannot be overcome by collecting larger sample sizes, which often, insidiously, worsen the problem (Rotello, Masson, & Verde, 2008).

Dube et al. (2010) applied a signal detection (SDT) model of belief bias to our curved ROC data, and concluded that the belief effect and the interaction effect could be fully accounted for by a simple response bias shift for believable and unbelievable problems: Reasoning accuracy did not differ with believability, though subjects' willingness to say "valid" did. In contrast, traditional analyses had indicated that reasoning accuracy was greater for unbelievable arguments than for believable arguments. Accuracy differences are often used to justify theoretical claims of differential or extra processing for some argument types; for example Evans et al. (1983) concluded that when

reasoners are faced with an unbelievable conclusion, they undertake additional processing to scrutinize an argument's premises. If belief bias is simply a tendency to respond more positively to some arguments than others, these extra processes are unnecessary.

In Dube et al. (2010), difference score analyses led to the usual conclusion that there is an interaction between logic and belief, but SDT analyses concluded that there was no interaction. Indeed, in one experiment we eliminated the belief content of the syllogisms, replacing the content words with letters and imposing a between-subjects manipulation of response bias. The resulting response rates were analyzed using a standard 2×2 ANOVA and revealed significant main effects of logic and bias, and a significant interaction. In other words, despite being presented with identical problems, participants in the conservative condition appeared to reason more consistently with the rules of logic than those in the liberal condition. In contrast, SDT model fits led to the conclusion that only the response bias parameters need to be free to vary to account well for the data; accuracy parameters do not differ.

To see how these vastly different interpretations of the data are possible, consider Fig. 1, which shows hypothetical data that might be observed in an experiment on belief bias. A typical result is represented by points B and D, where point B reflects more conservative responding to unbelievable problems, and point D reflects more liberal responding to believable problems. Notice that point B falls on a linear ROC implied by a higher value of the traditional difference score measure, *H–F*, relative to point D. In a traditional analysis, these points would be interpreted as showing an interaction between validity and believability, in which reasoning accuracy is higher for unbelievable conclusions. Points A and C would also be interpreted as showing an interaction, although for that pair of points higher reasoning accuracy would be inferred for the believable problems (point C). If Points B and C were observed empirically, the experiment might be deemed a failure: They fall on the same *H–F* ROC and therefore difference score analysis would conclude that the interaction index was zero. The signal detection interpretation of these points, shown as the smooth curve, is that they all reflect the same reasoning accuracy, differing only in response bias.

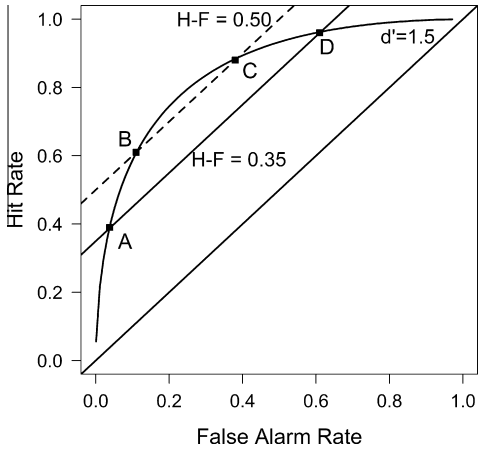

Fig. 1. Hypothetical data from an experiment on belief bias and corresponding ROCs implied by difference score measures and SDT.

We emphasize that linear ROCs are a necessary assumption of traditional difference-score analyses of reasoning. The difference score approach subtracts the positive response rate to one type of stimulus (say, invalid problems) from that to another stimulus type (valid problems). This assumes that the difference, *H–F*, measures accuracy; Snodgrass and Corwin (1988) called this measure P_r. If only bias changes across conditions, then response rates to both types of stimuli will increase or decrease, but P_r should be constant. Indeed, because ROCs are isosensitivity curves, connecting the data points from conditions that differ in response bias but not accuracy defines the theoretical ROC for that accuracy measure. In the case of P_r, we note that $P_r = H - F$, or, equivalently, that $H = P_r + F$. Because P_r is necessarily constant in an ROC, this simple equation shows that the hit rate is a linear function of the false alarm rate, with intercept equal to P_r and a slope of 1. All points that have equal P_r must fall on the same line (Swets, 1986, p. 111). For example, consider one experimental condition that results in a correct response rate of 0.61 and an error rate of 0.11, yielding $P_r = 0.50$ (this is point B in Fig. 1). If another condition produces a liberal response bias shift, the false alarm rate might increase, say to 0.39. If accuracy is unchanged, then the hit rate must be 0.89, because $H = P_r + F = 0.50 + 0.39$ (this is point C in Fig. 1). Similarly, if responding is very conservative, so that the false alarm rate is 0, then the hit rate must be 0.50 if accuracy is constant; these data would appear at the *y*-intercept of the $H - F = 0.50$ line in Fig. 1. Data points that do not fall on that line reflect different values of P_r and therefore different accuracy levels.

The current state of the field leaves researchers in a muddle. The belief bias effect is a central phenomenon in reasoning research, but there are differing views on the nature of the belief bias effect, how to measure it, and how to explain it. Traditional raw score measures suggest a dramatically different interpretation of the data than the newly-proposed signal detection measures and model. Traditional measures have implied that believability influences reasoning itself, whereas signal detection measures based on ROCs have generally suggested that believability influences only participants' willingness to say "valid." In this paper, we further contrast traditional analyses of reasoning tasks and SDT model-based analyses. We first show via simulation that traditional approaches can easily lead to different conclusions than SDT model-based analyses. Next, in two experiments, we investigate the important issue of how instructions affect the three belief bias effects; these data show that the assumptions of the traditional analyses (but not SDT analyses) are violated. Finally, we present ROC data from other deductive reasoning tasks, such as conditional reasoning, that are also consistent with the assumptions of SDT and inconsistent with traditional analytic approaches.

1.1. Simulations

We first present new simulations that demonstrate that all three traditional difference-score measures for

reasoning experiments are at risk. We take as a starting point the observation that all extant empirical ROC curves for reasoning experiments are curved (Dube et al., 2010; Dube et al., 2011; Heit & Rotello, 2005; Heit & Rotello, 2008; Heit & Rotello, 2010; Heit & Rotello, 2012; Heit et al., 2012; Rotello & Heit, 2009; Trippas et al., 2013), and are consistent with arguments that vary in strength according to Gaussian distributions. We generate simulated data for hypothetical experiments containing an instructional manipulation that affects response bias and not reasoning accuracy. We show that traditional raw score analyses will tend to incorrectly imply that there is a difference in the size of the logic effect across instructional conditions when there is none, and will likewise tend to incorrectly imply that the interaction between logic and belief differs from one instructional condition to the other. Then we address the case in which instructions affect reasoning accuracy rather than response bias. Although traditional measures will correctly pick up the difference in the logic effect, they may incorrectly imply that the belief effect differs across conditions when it does not.

To generate the simulated data, we sampled evidence values from Gaussian distributions like those shown in Fig. 2a. Evidence values sampled from the valid distribution had a higher mean than those sampled from the invalid distribution, and the magnitude of the difference in mean strength was varied over several levels (see Table 1 for details). In this set of simulations, we assumed that the believability of an argument's conclusion influenced only response bias (i.e., criterion location), with believable problems yielding a more liberal bias (see Fig. 2a). Although there was a logic effect, it was the same size for both believable and unbelievable problems, and thus there was no interaction between belief and logic. The magnitude of the bias shift was also varied over several levels (Table 1). Layered on top of the bias effect attributable to believability, we assumed in this simulation that the instructional manipulation itself affected response bias. For simplicity, the two bias effects were assumed to be additive. For each simulated trial, the sampled evidence value was compared to the appropriate decision criterion; values above the criterion led to "valid" responses, and those below it led to "invalid" decisions. The number of simulated trials per subject was varied parametrically, as was the number of simulated subjects per condition in each experiment.

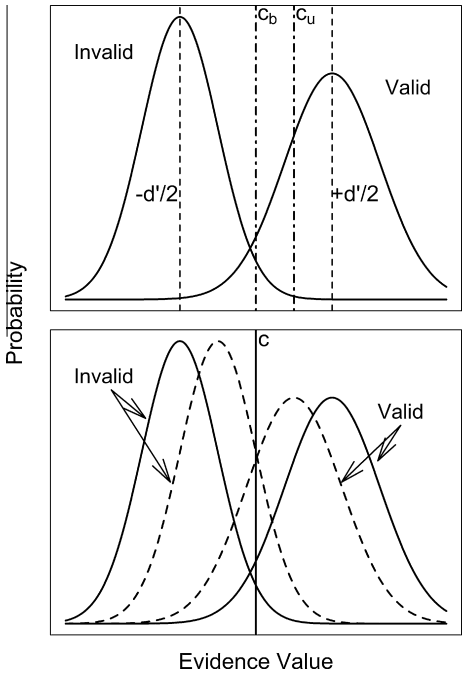

Fig. 2. Signal detection model assuming (A) Believability affects only criterion location, with a more conservative criterion for unbelievable conclusions (c_u) than for believable conclusions (c_b); instructions also affect only criterion location. (B) Instructions affect reasoning accuracy (solid distributions show higher accuracy condition; dashed distributions show lower accuracy condition); believability affects only response bias, as in panel A.

For each combination of parameter values, we simulated 1000 experiments, computing the logic, belief, and interaction indices in both instructional conditions. These values were then compared using two-sample *t*-tests. Because the distance between the valid and invalid distributions did not vary with instructions or conclusion believability, significant *t*-tests for the logic effect or the interaction effect each represent Type I errors in which the two instructional conditions are erroneously concluded to yield different accuracy or an interaction. Although there was a belief effect within each condition as a result of the response bias shift for believable problems, the magnitudes of the effect were identical, so significant *t*-tests for the belief effect also represent Type I errors in this simulation.

The results for the logic effect are shown in the upper row of Fig. 3, for a representative set of parameters (overall d' = 1, *z*ROC slope = 0.8, the most liberal criterion placed at the mean of the invalid distribution, and instruction effect of 0.4 standard deviations). The left panel shows the probability that the two instruction conditions are falsely declared to yield a different validity effect when there are 20 simulated subjects per condition, and the right panel shows the results for 60 simulated subjects. This simulation shows that there is a substantial risk of erroneously inferring that there is a different logic effect across instruction conditions. Moreover, increasing the number of subjects (left vs. right panel) or the number of sampled trials (x-axis values within a panel) increases the probability of drawing the wrong conclusion. This probability increases as the magnitude of the belief bias shift increases (functions within a panel). The probability of this error also increases as the bias difference between conditions increases (not shown). In sum, the circumstances that would usually lead to a more powerful experiment—more subjects, more data per subject, larger effect size—actually lead to more incorrect conclusions for the traditional difference-score measure of the logic effect.

The middle row of Fig. 3 presents the results for the interaction index. The same general patterns are revealed, indicating that the traditional measures of the interaction index often leads to the incorrect conclusion that there is

Table 1
Parameter values used in simulations of experiments in which believability affected response bias and instructions affected either response bias or accuracy.

Parameter	Values simulated
N = number of simulated subjects	20, 40, 60, 80
Trials = number of trials of each problem type per subject	2, 4, 8, 16, 32
Slope of zROC	0.6, 0.8, 1, 1.2
Accuracy = distance between distribution means in units of invalid distribution standard deviation	0, 0.5, 1, 1.5
Criterion position for believable problems	c = 0, 0.5, 1, 1.5
Criterion increment for unbelievable problems	0, 0.2, 0.4, 0.6

an interaction present. Finally, the bottom row of Fig. 3 shows that the belief effect had a Type I error rate near 0. The result obtains for the simple reason that the bias shifts were always the same size in both conditions.

In a second set of simulations, analogous to our first set, we assumed that the effect of instructional condition was to influence overall reasoning accuracy and not response bias. In brief, traditional measures on the simulated data showed belief and interaction effects that were not really there, and were more likely to do so when sample sizes were larger. Finally, note that both sets of simulations are independent of the nature of the reasoning task, e.g., they would apply equally well to syllogistic reasoning and conditional reasoning.

1.2. Overview of experiments

Having shown the potential for traditional analyses to draw incorrect conclusions from simulated data, we next turn to two new experiments. Dube et al. (2010) argued that the traditional interaction measure is inappropriate and can lead to incorrect inferences. Here, we extend our arguments to the logic index and the belief index as well. In our simulations, we have shown that in experiments designed to measure the influence of instructions on belief bias, or, indeed, with any experimental manipulation that could influence either response bias or reasoning accuracy, all three traditional measures will often lead to incorrect conclusions. In our new experiments, we revisit the issue of how instructions affect belief bias in syllogistic reasoning (Evans et al., 1994; Newstead et al., 1992).

Evans et al. (1994) varied the level of detail provided in instructions to participants. Some conditions used elaborate instructions that emphasized the nature of logical necessity; other instruction sets were long, complex, and included the logical definition of the quantifier "some." These complex instructions were found to reduce the size of the belief effect and to render the interaction non-significant. Evans et al. interpreted those data as indicating that the believability of a conclusion strongly influences the accuracy of a participant's reasoning processes. In contrast to previous work, we compared the three traditional measures of reasoning performance to measures derived from SDT analyses. If the resulting ROCs are curved and consistent with SDT, then the SDT analysis will more accurately identify the contributing effects.

2. Experiment 1

Two sets of instructions were used, standard and augmented. The standard instructions were based on Evans et al. (1994, Experiment 1) as well as Newstead et al. (1992, Experiment 5). The augmented instructions, also based on those experiments, included additional

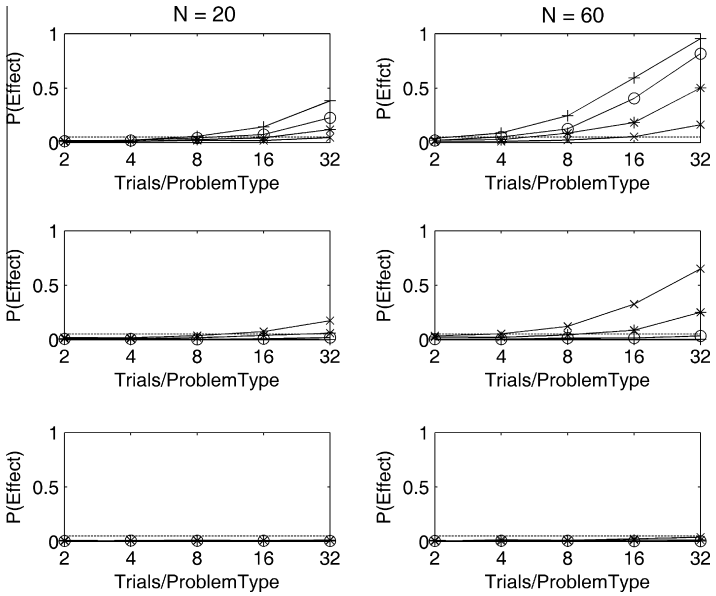

Fig. 3. Representative results of simulation in which both believability and instructions affect only response bias. Left panels: N = 20; Right panels: N = 60. Upper row shows the logic effect, middle row shows the interaction effect, bottom row shows the belief effect. The dashed line at P(Effect) = .05 represents the nominal α level; functions within each panel reflect various belief effects, with higher functions invariably reflecting larger effects.

reminders to "only endorse a conclusion if it definitely follows from the information given." We focused on whether the augmented instructions led participants to adhere more closely to the rules of logic (thus increasing the logic index) and to show less of a belief bias effect in terms of the belief index and the interaction index. Due to the requirements of modeling, we used a larger set of syllogisms compared to the previous studies, and also collected confidence ratings after each judgment of validity.

2.1. Method

2.1.1. Participants

Eighty-nine undergraduates from the University of California, Merced, participated in exchange for extra credit. Approximately half were randomly assigned to each instruction condition ($n = 45$ in standard condition, $n = 44$ in augmented condition).

2.1.2. Design

This experiment used a 2 (standard or augmented instructions) × 2 (valid or invalid problem) × 2 (believable or unbelievable conclusion) mixed design, with instructions as a between-subject variable and the other two variables within-subjects. All participants were asked to evaluate the validity of syllogisms that differed in their logical status and conclusion believability.

2.1.3. Materials

Aside from the instructions, the stimuli were identical to those of Dube et al. (2010, Experiment 2). All participants first received 3 valid and 2 invalid practice problems with abstract materials. Then each participant saw 32 syllogisms, derived from 16 syllogistic problem frames, 8 corresponding to valid arguments and 8 corresponding to invalid arguments. Each problem frame was assigned content that led to one believable conclusion and one unbelievable conclusion. All sets of content were randomly assigned to the 32 problem structures. The meaningful content was derived from a study by Morley, Evans, and Handley (2004) as well as pretesting by Dube et al. (2010). 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. (3)
Some artists are not sculptors.

The instructions were as follows (standard instructions are shown; additional material for the augmented instructions only is indicated in bold).

> This experiment is designed to find how people solve logical problems. Your task is to decide whether each conclusion follows logically from the information given in that problem. The premises—the information given—appear above the line and the conclusion appears below the line.
>
> You must assume that all the information you are given is true; this is very important. If, and only if, you judge that a given conclusion logically follows from the information given you should answer ''Valid.'' If you think that the given conclusion does not necessarily follow from the information given you should answer ''Not Valid.'' Also, you will be asked how confident you are in this judgment.
>
> **Please note that according to the rules of deductive reasoning, you can only endorse a conclusion if it definitely follows from the information given. A conclusion that is merely possible, but not necessitated by the premises is not acceptable. Thus, if you judge that the information is insufficient and you are not absolutely sure that the conclusion follows you must reject it and answer ''Not Valid.''**
>
> Please take your time and be certain that you have the logically correct answer.
>
> **REMEMBER, IF AND ONLY IF YOU JUDGE THAT A GIVEN CONCLUSION LOGICALLY FOLLOWS FROM THE INFORMATION GIVEN YOU SHOULD ANSWER ''Valid,'' OTHERWISE ''Not Valid.''**

2.1.4. Procedure

All participants were tested individually using a computer program. The instructions were presented once on the computer screen and also on a sheet of paper that participants were allowed to refer to throughout the experiment.

Following the instructions, participants advanced through the 5 practice trials and then the 32 critical trials. Participants made validity decisions via keypress (J = "valid"; F = "invalid"). After each "valid"/"invalid" response, participants were asked to rate their confidence on a scale of 1 to 3 (1 = not at all confident, 2 = moderately confident, 3 = very confident). We subsequently recoded the responses with the values 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.

2.2. Results and discussion

The proportions of "valid" responses given to each stimulus type in each condition are shown in Table 2.

2.2.1. Traditional analysis

As would be usual in the belief bias literature, the data in Table 2 were subjected to a 2 (validity status: valid, invalid) × 2 (belief status: believable, unbelievable) × 2 (instruction condition: standard, augmented) ANOVA. Subjects gave more "valid" responses to valid problems ($F(1,87) = 77.16$, MSe = 0.06, $p < .001$), showing a significant logic effect. They also gave more "valid" responses to problems with believable conclusions ($F(1,87) = 52.03$, MSe = 0.06, $p < .001$), showing a reliable belief effect. The third component of the standard belief bias effect, a

significant interaction of belief and validity, was also observed ($F(1,87) = 13.67$, MSe = 0.02, $p < .001$). The interaction index has typically been interpreted as implying that subjects reason more accurately about problems with unbelievable conclusions than those with believable conclusions.

There was no main effect of instruction condition ($F(1,87) = 0.001$), however the logic effect was slightly larger with the standard instructions (0.28 vs. 0.18); the condition × validity interaction was marginally significant ($F(1,87) = 3.51$, MSe = 0.06, $p = .064$). This result suggests that reasoning is more accurate overall with standard, rather than augmented, instructions. The 3-way interaction of condition, validity, and belief status was also reliable ($F(1,87) = 10.32$, MSe = 0.02, $p < .01$), presumably because the belief × validity interaction was larger in the standard instruction condition (0.18) than the augmented condition (0.01). Evans et al. (1994, Exp. 3) also observed a near-zero interaction effect for their complex instruction condition. A traditional interpretation of the data for either experiment would suggest that the longer instructions either eliminated the reasoning benefit bestowed by unbelievable conclusions or encouraged more careful reasoning about the believable problems. No other effects were statistically significant, including the condition × belief interaction. That is, the main effect of prior beliefs did not vary significantly with instructional conditions.

2.2.2. ROCs and modeling

Traditional analyses led to following conclusions: There were significant logic and belief effects in both conditions, and a significant interaction effect in the standard instruction condition only. To understand the processes that underlie the belief bias effect, we plotted the ROC data, shown as the points in Fig. 4. These ROCs were generated by plotting the most conservative, highest-confidence (rating of 1 only), hit and false alarm rates as the left-most point, and then relaxing the minimum confidence one level at a time (ratings of 1 or 2, ratings of 1, 2, or 3, etc.) to determine the remaining operating points. Note first that the form of each ROC is curved, not linear, and that the ROCs fall at the same or similar height in the space for believable and unbelievable arguments. This pattern suggests equal sensitivity to validity for those two problem types, because ROCs that fall higher in the space reflect higher accuracy levels. Next, note that the operating points for the believable problems are consistently shifted to the upper-right compared to the operating points for the unbelievable problems. This shift is a clear indication of a response bias shift attributable to conclusion believability, which yields an increase in both the hit and false alarm rates at each confidence level.

To assess these visual observations quantitatively, we fitted an SDT model to the data using Dube et al.'s (2010) approach. The model, like those shown in Fig. 2, has parameters that describe the sensitivity of reasoning to logic as well as parameters that capture response biases. We fitted the model to the data from each instruction condition using a maximum likelihood criterion, allowing all parameters to vary freely between conditions, or with the restrictions that either the logic or the bias parameters did not differ for believable and unbelievable problems within a condition. These constrained versions of the model are nested within the unconstrained model, and so their fits may be compared with the difference in G^2 statistics, $\Delta G^2 = G^2_{\text{restricted}} - G^2_{\text{full}}$, with degrees of freedom equal to $df_{\text{restricted}} - df_{\text{full}}$.

The modeling results are shown in Table 3.[1] For both types of instructions, the full model captured the data well, and the quality of the fit was not impaired by the assumption of that sensitivity to logic was the same for believable and unbelievable problems. In contrast, the model fit was significantly reduced by the assumption that response biases were the same for believable and unbelievable problems. For these reasons, the overall best account of these data is provided by the SDT model with the equal-logic constraint across believability status; the closeness of the model's predictions to the empirical data can be seen in Fig. 4.

Framed in terms of the more traditional conclusions in the belief bias literature, the SDT analyses led to following conclusions: Participants made more "valid" responses to valid than invalid arguments (a logic effect in traditional language), and they made more "valid" responses to believable than unbelievable conclusions (a belief effect). However, the two variables did not interact under either instruction set: Participants' ability to discriminate valid from invalid problems does not depend on conclusion believability. These conclusions are consistent with those reached by Dube et al. (2010), who suggested that, contrary to the traditional analyses, the effect of belief bias is solely on response tendencies.[2]

The implications of Experiment 1 are that the belief bias effect is best described as a response bias effect, but traditional analyses can erroneously imply that reasoning accuracy is better for problems with unbelievable rather than believable conclusions. Across conditions, it is evident that our instructional manipulation influenced reasoning accuracy: Overall accuracy is slightly *lower* with the augmented instructions. This observation was supported by simultaneous fits to the ROCs in each condition, either allowing all parameters to vary freely ($\Delta G^2_{df=12} = 23.74$) or equating accuracy across instructions ($\Delta G^2_{df=4} = 14.61$, $p < .01$). Although this is an interesting finding, we would hesitate to interpret it without first assessing its generality. Response bias was marginally affected by the augmented instruction condition, as shown by an SDT fit in which response bias was constrained to be equal across instruction conditions ($\Delta G^2_{df=10} = 17.40$, $p = .066$). Again, though, regardless of overall accuracy level, the effect of conclusion believability was merely one of response bias.

[1] Full details, including parameter values, are available from the authors.

[2] An alternative interpretation of these data is that the criterion is fixed but the distributions of both valid and invalid unbelievable problems is shifted rightward compared to the distributions for believable problems. Such a shift is mathematically indistinguishable from our interpretation, as both yield the same data. However, we view that alternative as implausible in light of the results reported by Dube et al. (2010, Exp. 1) and described in the introduction, in which abstract syllogisms were shown to subjects under two bias conditions; the data were virtually identical those from a belief bias study on the same syllogistic structures (but with believable/unbelievable content).

Table 2
'Valid' response rates for each stimulus type and condition in each experiment, and accuracy measures averaged over participants' scores.

Exp.	Instruction condition	*H* = P('valid'|Valid)		*F* = P('valid'|Invalid)		*H*–*F*		*d*′	
		Bel.	Unbel.	Bel.	Unbel.	Bel.	Unbel.	Bel.	Unbel.
1	Standard	0.82	0.71	0.63	0.34	0.19	0.37	0.56	1.06
	Augmented	0.80	0.63	0.62	0.44	0.18	0.19	0.48	0.50
2	Standard	0.87	0.63	0.68	0.41	0.19	0.22	0.60	0.57
	Conservative	0.82	0.63	0.56	0.36	0.26	0.36	0.76	0.79

These results resemble our simulations, in which an underlying difference in bias was picked up by traditional analyses as a logic effect as well as an interaction between logic and belief. Moreover, by using the confidence ratings to plot ROC curves, we can see that the traditional analyses are inappropriate, because they assume linear ROCs that are not present in our data. In contrast, the assumptions of SDT model-based analyses, namely curved ROCs that are consistent with underlying Gaussian distributions of argument strength, are supported by the data.

3. Experiment 2

Having shown in Experiment 1 that traditional analyses and SDT model-based lead to different conclusions regarding belief bias and the effects of instructions, we next investigated an alternative instructional manipulation, in which participants were told to be more conservative, namely that when they are guessing, they should respond "not valid" rather than "valid." In a recognition memory paradigm, Rotello, Macmillan, Hicks, and Hautus (2006) showed that an analogous instructional manipulation (i.e., say "new" if you aren't sure that that test item was studied) shifted participants' response criteria without affecting memory accuracy. However, for a reasoning task, it is possible that instructions that emphasize care when responding "valid" could affect sensitivity to logic as well as response biases. Again, our aim was to compare conclusions drawn from traditional analyses and SDT model-based analyses. Moreover, by repeating the standard instruction condition from Experiment 1, we were able to observe whether traditional measures lead to consistent or inconsistent conclusions across the two experiments.

3.1. Method

The method was the same as Experiment 1 except for the following. Eighty-eight undergraduates participated; half were randomly assigned to the standard instructional condition and the remainder were tested in the conservative instructional condition.

The instructions were as follows (standard instructions are shown; additional material for the conservative instructions only is indicated in bold).

> This experiment is designed to find how people solve logical problems. Your task is to decide whether each conclusion follows logically from the information given in that problem. The premises—the information given—appear above the line and the conclusion appears below the line.
>
> You must assume that all the information you are given is true; this is very important. If, and only if, you judge that a given conclusion logically follows from the information given you should answer ''Valid.'' If you think that the given conclusion does not necessarily follow from the information given you should answer ''Not Valid.'' Also, you will be asked how confident you are in this judgment.
>
> **Do not answer ''Valid'' unless you are very sure that the conclusion follows logically. If you are not sure or you have to guess, it is better to answer ''Not Valid.''**
>
> Please take your time and be certain that you have the logically correct answer.
>
> **REMEMBER TO DO YOUR BEST. BUT IF YOU HAVE TO GUESS, YOU SHOULD ANSWER ''Not Valid.''**

3.2. Results and discussion

The proportions of "valid" responses given to each stimulus type in each condition are shown in Table 2.

3.2.1. Traditional analysis

As in Experiment 1, the response rate data in Table 2 were subjected to a 2 (validity status: valid, invalid) × 2 (belief status: believable, unbelievable) × 2 (instruction condition: standard, conservative) ANOVA. Subjects gave more "valid" responses to valid problems ($F(1,86) = 87.77$, MSe = 0.06, $p < .001$), showing a significant logic effect. They also gave more "valid" responses to problems with believable conclusions ($F(1,86) = 72.22$, MSe = 0.07, $p < .001$), showing a reliable belief effect. In this study, there was no significant interaction of belief and validity ($F(1,86) = 0.45$).

The interaction component of the standard belief bias effect has been the most inconsistent of the three components of the belief bias effect, a phenomenon that some researchers have explained in terms of dual-process theories of deduction (Evans, 2006; Evans, 2007; Shynkaruk & Thompson, 2006). For example, Evans and Curtis-Holmes (2005) found the interaction was smaller when a response deadline was imposed, which they argued limited the contribution of System 2/analytic reasoning processes. Perhaps relatedly, Shynkaruk and Thompson obtained a

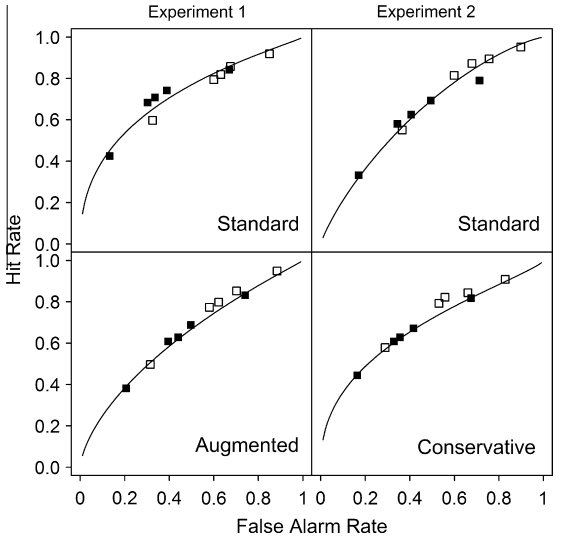

Fig. 4. ROCs for both Experiments. Filled symbols = unbelievable problems; open symbols = believable. Solid curve = equal-accuracy SDT fit.

reliable interaction only for subjects with high reasoning ability, an observation they also interpreted from a dual-process perspective. In contrast, Dube et al. (2010) argued that the variability of the interaction effect was likely due to differences in the level of overall accuracy, the form of the ROC, and the magnitude of the belief effect (see Fig. 1). Regardless of the precise source of that variability, our point is that the absence of the interaction effect in this experiment is not particularly unusual for traditional measures.

There was a marginal effect of instruction ($F(1,86) = 3.59$, MSe = 0.07, $p = .062$), with, appropriately, fewer positive responses being given in the conservative condition. No interactions with instruction condition were significant (all $ps > .2$).

In sum, traditional analyses would lead to somewhat different conclusions for Experiment 2 than Experiment 1, namely significant main effects of logic and belief, no interaction between the two, and no significant effect of instructions. It is most notable to compare the standard instruction conditions for the two experiments; traditional analyses lead to the conclusion that there was a fundamental difference in results, with an interaction between logic and belief in Experiment 1 but not Experiment 2.

3.2.2. ROCs and modeling

The ROCs for Experiment 2 are shown in Fig. 4. As in Experiment 1, these ROCs are curved and asymmetric, and therefore violate the linearity assumption implicit in the traditional analysis. For this reason, we used the same modeling strategy as in Experiment 1; the resulting fit statistics are shown in Table 3. The modeling results largely replicate those from Experiment 1: The SDT model fit the data well in each condition, and comparative model fitting indicates that the assumption of equal response bias for believable and unbelievable arguments can be strongly rejected. However, in contrast to Experiment 2, in the standard condition there appears to be a small but statistically significant accuracy difference between believable and unbelievable arguments, in favor of unbelievable arguments. It is rather difficult to spot this difference in Fig. 4 itself, that is, the two curves look nearly equally close to the top-left corner of ROC space. To put this into perspective another way, the effect size, w, ranges from 0 to 1 and increases with the misfit of the model (Cohen, 1988). Effect size is estimated to be 0.691 for the assumption of equal bias in the control condition, and less than half that size, 0.274, for the assumption of equal accuracy. Indeed, for each condition in the two experiments, w is estimated to be at least 2.5 times larger under the assumption of equal bias than the assumption of equal accuracy.

As expected, the conservative instruction condition yielded more conservative response bias than the standard instructions. This observation was supported by simultaneous fits of the SDT model to both ROCs in each condition, either allowing all parameters free to vary ($G^2_{df=12} = 31.23$) or equating response bias across instructions. The response bias restriction dramatically reduced the quality of the model fit: $\Delta G^2_{df=10} = 39.99$, $p < .001$. As anticipated, we also observed a small difference in reasoning accuracy across instructions: Subjects' ability to discriminate valid from invalid problems was slightly greater in the conservative instruction condition. Constraining the accuracy

Table 3
Fit statistics for the SDT model to the data from Experiments 1 and 2.

Experiment	Instructions	Constraint	G^2	df	ΔG^2
1	Standard	None	16.87	6	
		Equal accuracy	19.62	8	2.79
		Equal bias	168.41	11	151.54***
	Augmented	None	6.87	6	
		Equal accuracy	10.08	8	3.21
		Equal bias	73.95	11	67.08***
2	Standard	None	18.96	6	
		Equal accuracy	27.22	8	8.26*
		Equal bias	159.03	11	140.08***
	Conservative	None	12.27	6	
		Equal accuracy	13.84	8	1.57
		Equal bias	110.78	11	98.51***

* $p < .05$,
*** $p < .001$.

parameters to be equal across instructions reduced the quality of the model fit: $\Delta G^2_{df=4} = 9.82$, $p = .044$. We note that augmented instructions led to a slight decrease in accuracy in Experiment 1, however those were different instructions, and we would emphasize that these accuracy differences in each experiment were slight, and reflected changes in overall accuracy.

Again, the main finding is that traditional analyses and SDT model-based analyses reached different conclusions. Most notably, traditional analyses lead to different conclusions for the standard conditions for the two experiments, suggesting that the interaction between logic and belief that has been crucial for theoretical development in reasoning research appears in Experiment 1 but not Experiment 2. (For an illustration of how points from a single ROC curve can be taken as supporting either an interaction or no interaction when traditional measures are used, refer back to the Introduction and Fig. 1) In contrast, SDT analyses indicated no such interaction for Experiment 1 and a small but statistically significant interaction for Experiment 2. In addition, the main effect of the instructional manipulation on response bias in Experiment 2 was picked up by the SDT analysis, but did not quite reach the level of statistical significance for the traditional analysis using the belief index. SDT analyses supported the conclusion that the belief bias effect is manifested primarily or exclusively as a response bias effect (as in Dube et al., 2010; Dube et al., 2011), although as noted, there was a very small but statistically significant accuracy difference in the standard condition. Finally, by using the confidence ratings to plot ROC curves, we can see that the traditional analyses are inappropriate, because they assume linear ROCs.

4. General discussion

Our initial simulations suggested that when argument strength has an underlying distribution that is Gaussian in form, there is great potential for traditional analyses such as the logic index, belief index, and interaction index to draw incorrect conclusions. For example, when conclusion believability and instructions are known to affect only response bias, our simulations showed that there is still a good chance that the traditional logic effect and interaction effect measures will nonetheless show significant effects instructions. The problems for traditional measures get worse as sample sizes get larger. These simulations expand upon our previous work (Dube et al., 2010; Dube et al., 2011; see also Rotello et al., 2008) which focused only on the interaction between logic and belief, and did not address the sample size issue.

In our two new experiments on belief bias in syllogistic reasoning, we found that the underlying form of the data violated the assumptions of traditional analyses, and instead was consistent with the assumptions of SDT model based-analyses. We also found that these two forms of analysis led to different conclusions from the experiments. The SDT analyses showed strong differences in response bias between believable and unbelievable arguments, and between instructional conditions, for both experiments. Traditional analyses showed an interaction between logic and belief for Experiment 1, but not for Experiment 2, and found that in Experiment 1, the interaction was eliminated by instructions.

4.1. Other studies of belief bias

Few experiments on belief bias have collected confidence ratings that allow researchers to conduct SDT-model based analyses. Fortunately, a recent study by Eliades et al. (2012, Experiment 1) also collected confidence ratings (although these were not used by the authors for SDT analyses). Their materials varied in content, being either neutral (as in most belief bias experiments), emotional but not sexual, or related to sex abuse. Subjects were asked to make their validity assessments using confidence ratings, allowing us to generate the empirical ROCs and to estimate accuracy measures using SDT. We compared these SDT fits to traditional difference score analyses.

4.1.1. Traditional analysis

Traditional analyses of the hit and false alarm rates (shown in Table 4) revealed that only the emotional stimuli supported the standard triumvirate of effects (validity, $F(1,71) = 355.18$, MSe = .116, $p < .001$; belief, $F(1,71) = 3.69$, MSe = .060, $p < .06$; interaction, $F(1,71) = 4.73$, MSe = .026, $p < .05$). The sex abuse related materials also yielded standard validity ($F(1,72) = 345.68$, MS = .108, $p < .001$) and belief effects ($F(1,72) = 15.78$, MSe = .059, $p < .001$), and a significant interaction ($F(1,72) = 4.31$, MSe = .088, $p < .05$). However, the interaction effect was atypical in that believable stimuli yielded more accurate reasoning than unbelievable, the opposite of what has usually been observed and explained by theories of reasoning. Finally, the neutral stimuli yielded both validity ($F(1,72) = 914.16$, MSe = .059, $p < .001$) and belief effects ($F(1,72) = 4.01$, MSe = .247, $p < .05$) but no interaction ($F(1,72) = 0.03$, $p > .85$), suggesting equal reasoning accuracy for believable and unbelievable problems. Because the items for each content set were not equated on believability, and because premises themselves contained information that might be more or less believable (e.g., "No Scottish residents are good people," "Some women are lawyers"), we will not attempt to interpret the different patterns of effects across materials, including the reversal of the standard interaction effect (though we note that Points A and C in Fig. 1 suggest that such results are not unexpected from an SDT perspective). Instead, we focus on the comparison with SDT-based interpretations of the same data.

4.1.2. ROCs and modeling

The ROCs from this experiment, shown in Fig. 5, are strongly curved in all cases. Therefore, these data are inconsistent with the traditional difference score measures, which imply linear ROCs. To emphasize the discrepancy, we have superimposed the linear ROCs implied by the use of *H–F* in each panel of Fig. 5; clearly the empirical data do not conform to the expectations of that measure. Instead, they appear consistent with our SDT model.

As in Experiments 1 and 2, we fit the SDT model to the data from each stimulus set (neutral, emotional, sex abuse

Table 4
Traditional response rate summaries for data from Eliades et al. (2012, Exp. 1).

Materials	*H* = *P*('valid'|Valid)		*F* = *P*('valid'|Invalid)		*H*–*F*	
	Bel.	Unbel.	Bel.	Unbel.	Bel.	Unbel.
Neutral	0.92	0.86	0.05	0.00	0.87	0.86
Emotional	0.82	0.80	0.10	0.00	0.72	0.80
Sex abuse related	0.86	0.67	0.07	0.03	0.79	0.64

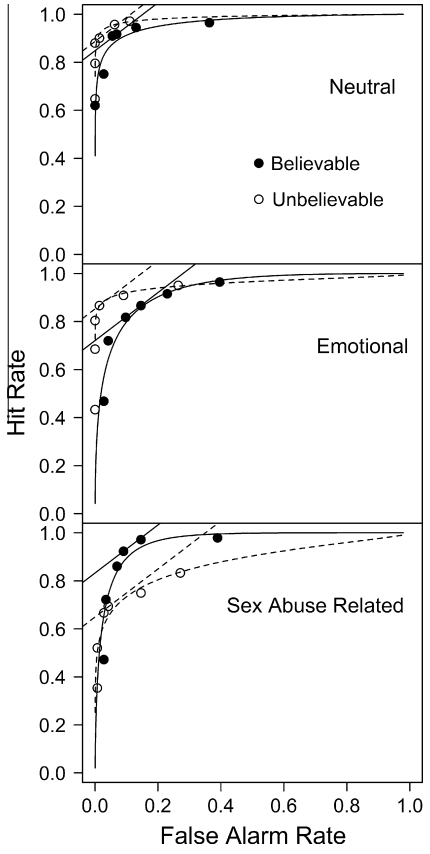

Fig. 5. Observed ROCs from Eliades et al. (2012, Exp. 1) with *H*–*F*-implied ROCs (lines) and SDT-fitted ROCs (curves) superimposed.

related) in three ways, allowing both accuracy and response bias parameters to vary with believability status, or constraining either the accuracy or bias parameters to be the same for both believable and unbelievable stimuli. For the sex abuse related and emotional materials, the SDT analysis concurred with the traditional analysis: Both accuracy and response bias parameters varied with conclusion believability; constraining either set of parameters significantly reduced the fit (sex abuse: $\Delta G^2_{\text{accuracy},df=2} = 21.30$, $p < .001$, $\Delta G^2_{\text{belief},df=6} = 12.26$, $p < .06$; emotional: $\Delta G^2_{\text{accuracy},df=2} = 19.81$, $p < .001$, $\Delta G^2_{\text{belief},df=6} = 28.96$, $p < .001$). For the neutral stimuli, however, the traditional and model-based approaches led to different conclusions. Whereas the difference-score approach implied that reasoning accuracy did not differ with believability, the SDT analysis implied that both accuracy and response bias parameters were affected by conclusion believability; constraining either to be equal significantly reduced the fit (neutral: $\Delta G^2_{\text{accuracy},df=2} = 6.31$, $p < .05$, $\Delta G^2_{\text{belief},df=6} = 37.92$, $p < .001$). The best-fitting ROCs are shown as the curves superimposed on the data in Fig. 5.

In sum, the data from Eliades et al. (2012) reinforce the conclusions we have drawn from our own data, namely that ROCs from belief bias experiments are curved, making traditional analyses inappropriate, and that SDT model-based analyses lead to different conclusions than traditional analyses, here most notably in the neutral condition. Again, although the results do suggest an interaction between logic and believability, we hesitate to interpret this finding because, unlike many studies, the materials were not standardized for overall believability and premise plausibility, nor were they matched across stimulus content.

Another recent study, by Trippas et al. (2013), generally reinforces these points. This study also collected confidence ratings for experiments on the belief bias effect in syllogistic reasoning. Unlike Eliades et al., Trippas et al. did plot ROCs and perform SDT analyses. Most notably, the ROCs in Trippas et al. were curved, invalidating traditional analyses. Although it was not the main point of the study, Trippas et al. did report some traditional analyses, which sometimes agreed with SDT analyses and sometimes disagreed. For example, in their Experiment 1, both traditional and SDT analyses indicated that there was an interaction between logic and belief for more complex syllogisms but not for less complex syllogisms. Experiment 2 introduced two new variables: time limits to respond as well as participants' cognitive ability (measured by an IQ test). Traditional analyses did not reveal interactions of these new variables with the standard belief bias effects. In contrast, SDT analyses pointed to significant interactions, e.g., reasoning accuracy depended on an interaction between timing and believability for high-ability participants but not low-ability participants. In addition to finding bias differences between believable and unbelievable arguments, Trippas et al. did sometimes find interactions between logic and belief when looking at subsets of the data, e.g., for some problems or some participants, suggesting that in limited cases the belief bias effect may be more than a response bias effect.

Trippas et al.'s (2013) results differ from Dube et al. (2010), Dube et al. (2011), as well as the present Experiment 1, although we would note that our own research has not looked at subsets of the data in the same way. We leave open the possibility that model-based analyses may point to more nuanced findings at a detailed level when comparing problems or subsets of participants, while holding to the more general point that these

model-based analyses are necessary because the traditional analyses are inappropriate due to the curved ROCs. Moreover, the results from Trippas et al. may help explain the finding from the present Experiment 2, in which beyond a large bias difference between believable and unbelievable arguments, there was a small but significant accuracy difference, which could have been due to small variations in the nature of belief bias across participants. In general, our claim is not that there can never be accuracy differences between believable and unbelievable arguments, but rather that (1) in most of our own experiments, the belief bias effect was simply a response bias effect, and (2) in work by other researchers using traditional measures, any conclusion that accuracy differed between believable and unbelievable arguments rested on incorrect assumptions about the data and was probably wrong.

4.2. Analyses with d'

As we have noted, most reasoning experiments have not collected confidence ratings, and are prone to misinterpretation due to incorrect analyses. Without confidence ratings, researchers are left with a choice among single-point sensitivity measures such as the traditional difference score measures we have described. We have argued that these traditional measures are problematic as they assume underlying linear ROCs that are not evident empirically. The observed ROCs are curved and asymmetric, consistent with underlying evidence distributions that are unequal in variance and approximately Gaussian in form. Under those conditions, an accuracy measure such as d_a or A_z is strongly preferred (Rotello et al., 2008). However, calculation of d_a and A_z requires information about the slope of the ROC, which is not generally available without confidence ratings. An alternative is to calculate *d'*, a single-point measure of sensitivity that is consistent with an equal-variance signal detection model. In the Appendix, we present and apply d'-based analogs of the logic, belief, and interaction indices. We show in simulations that d' analyses are more conservative in general than traditional measures, that is, they are less likely to erroneously lead to the conclusion that a difference is significant when it is not.

Despite the potential of d' analyses, they are also inappropriate when the underlying evidence distributions are unequal in variance (as in all of the belief bias data we have analyzed). For this reason, applying d' analyses to own two experiments was not particularly promising. In particular, for these experiments, d' analyses led to the same conclusions as traditional measures. Hence, although d' analyses may, in theory, be an improvement over traditional analyses, our own experience does not provide instances making that point. We conclude that full ROC analyses relying on confidence ratings are preferred whenever possible.

4.3. Other model-based analyses

Although we have emphasized the value of SDT analyses of reasoning, by no means are these the only model-based way to analyze reasoning experiments in general or belief bias in syllogistic reasoning in particular. For example, Klauer et al. (2000, p. 855) also raised concerns about the interaction index used in traditional analyses, noting that differences between valid and invalid arguments for believable versus unbelievable items are not comparable when the absolute level of performance also differs as a function of believability (cf., Cook & Campbell, 1979), and that the variance of a proportion differs with its magnitude. We share this concern about comparability of baselines, but that complication does not apply directly to main effects, such as the logic index and belief index, which we have argued are also problematic. Klauer et al. advocated a form of discrete-state modeling known as multinomial processing tree (MPT) modeling. To be as clear as possible, it is not the purpose of this article to compare MPT and SDT models of reasoning (see, e.g., Dube et al., 2010; Dube et al., 2011; Klauer & Kellen, 2011, for some recent comparisons, and Pazzaglia, Dube, & Rotello, 2013, for a more general analysis). Indeed, the development of MPT models is broadly consistent with our own approach, which is to develop theoretically-motivated measures of reasoning making assumptions that are supported by the data.

Superficially, the most basic form of MPTs bears some resemblance to traditional analyses, in terms of predicting linear ROCs under many conditions. Hence, such models would face initial difficulties when applied to curved ROCs from belief bias experiments; in particular, they require elaboration to handle confidence-rating data. However, we note that MPTs are not equivalent to traditional analyses. We have applied a variant of the basic Klauer et al. (2000) MPT model to our own Experiments 1 and 2; as expected, the basic MPT model did not capture the curvature of the ROCs. The best fits required that both reasoning accuracy and response bias change with the believability of the conclusion, suggesting an interpretation of the data that is more consistent with the traditional difference score analysis, which pointed to an interaction between logic and belief for Experiment 1 but not Experiment 2.[3]

4.4. Other reasoning problems

Our focus so far has been on reasoning with categorical syllogisms, yet we have asserted that in principle, the concerns about incorrect analyses apply to other types of reasoning problems as well, such as conditional reasoning. This assertion rests on the assumption that ROCs would be curved rather than linear for these other reasoning problems. Here, we reanalyze previously collected data and show that the finding of curved ROCs is widespread outside of categorical syllogisms. Fortunately for these purposes, several published and unpublished studies

[3] In addition, we acknowledge that some versions of MPT models have the potential to account for curved ROCs. For example, Malmberg (2002) showed that by incorporating free parameters for how internal states are mapped onto ratings, MPT models can account for curved ROCs derived from confidence ratings. However, we would emphasize that Malmberg did not address the focus of our own work, traditional measures based on raw scores or differences. As described in Section 1, traditional measures, in widespread use within reasoning research, necessarily assume linear ROCs, and to our knowledge there is no published claim that traditional measures are consistent with curved ROCs.

collected confidence ratings that allow SDT-based analysis. Fig. 6 displays the curved ROCs for 6 such studies.

Fig. 6A shows a reanalysis of Rips (2001), in which participants made deductive validity judgments for four types of arguments, modus ponens (P), conjunctive syllogism (N), conjunction elimination (A), and disjunctive syllogism (O). The participants also made confidence ratings on a 10-point scale, although these were not used for any analysis. Here, we show ROC curves for this study, with the position of the modus ponens curve (closest to the top-left) suggesting that modus ponens was the easiest of the four problem types. (For this purpose, results are pooled over the causally consistent and causally inconsistent materials, and inductive plausibility judgments as opposed to deductive validity judgments are omitted.) The figure shows that the ROCs are clearly curved (and asymmetric) for all four problem types including modus tollens, which has been frequently used in studies of conditional reasoning. Hence, the functional form of these ROCs would put any traditional analysis of this study at risk.

In Fig. 6B, we show the results of an unpublished replication of Rips (2001) that we conducted. This replication had 52 participants making deduction judgments and used the same materials and procedure as the original study. The findings were similar to the original study. In Fig. 6C, we show the results of Heit and Rotello (2005, Experiment 1), a variant of the Rips study in which the materials were modified to strip out information related to causal knowledge (e.g., "Jill rolls in the mud" became "Jill does D"). Here, for the first time, we present separate ROCs for deduction judgments on all four argument types. Again, there are curved ROCs for conditional arguments and the other arguments, making traditional analyses inappropriate.

Markovits and Handley (2005, Study 1) collected probability ratings on a 1–7 scale for conditional arguments, namely modus ponens (P) and modus tollens (T). That study did not include ROC analysis, however in Fig. 6D we show that the ROC curves for both modus ponens and modus tollens are curved and asymmetric (with the usual result from conditional reasoning research that modus tollens arguments are more difficult). Markovits, Forgues, and Brunet (2010, Study 1) conducted a related experiment in which participants were given modus ponens and modus tollens arguments along with information about exceptions and/or disabling conditions. Fig. 6E shows the results pooled over modus ponens and modus tollens, with the same form of ROCs. Although it is not our purpose here to consider the conclusions of Markovits and colleagues in detail, we note that the curved, asymmetric ROCs do put any traditional analyses at risk. More generally, we note that the form of ROCs for conditional arguments in these studies as well as the original Rips (2001) study and our own replications suggest that curved ROCs would be the norm for conditional reasoning.

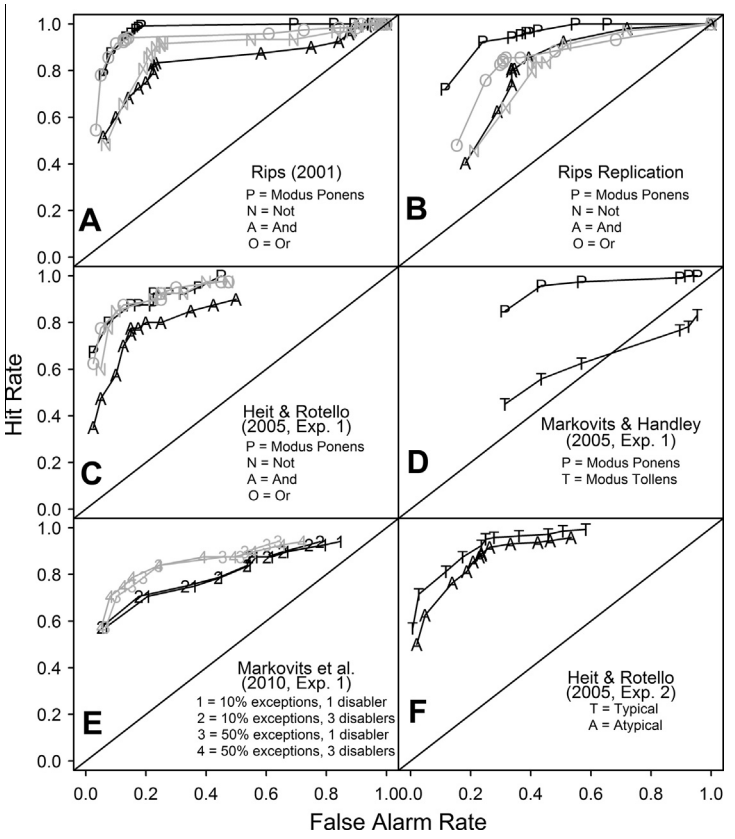

Fig. 6. ROCs generated from confidence ratings in six studies of deductive reasoning. See text for details.

Finally, we note that in our own previous work (Heit & Rotello, 2010; Heit & Rotello, 2012; Rotello & Heit, 2009), we have conducted several other experiments that included deduction judgments. In general, these experiments employed categorical arguments like *Mammals have property X, therefore horses have property X* (valid) and *Horses have property X, therefore mammals have property X* (invalid). These experiments provide additional evidence for the existence of curved ROCs in deductive reasoning, and the inappropriateness of traditional analyses. To give one example, in Fig. 6F we provide a reanalysis of Heit and Rotello (2005, Experiment 2). This experiment also varied typicality, for example substituting the atypical exemplar *dolphins* for the typical exemplar *horses* in the above arguments. Loosely speaking, these were belief bias studies in the sense that there are stronger beliefs associating horses with mammals compared to dolphins with mammals (cf., Sloman, 1998). Here, for the first time we provide ROCs separately for arguments with typical and atypical categories, in this case showing that beliefs about typicality of horses relative to mammals lead to more accurate reasoning about typical items.

5. Conclusion

Our argument can be summarized as follows: Experimental analyses are model-dependent; different models may lead to different conclusions; and analyses based on signal detection theory are justified based on the form of the ROC data whereas traditional analyses are not. Therefore, traditional difference-score analyses of reasoning performance in general, and the belief bias effect in syllogistic reasoning in particular, are flawed and likely to lead to erroneous conclusions.

The belief bias effect has been a central phenomenon in reasoning research for three decades. Major theoretical developments such as mental model theory (e.g., Oakhill & Johnson-Laird, 1985) and dual-process theory (e.g., Evans & Curtis-Holmes, 2005) have grown up alongside work on the belief bias effect using traditional measures, with inconsistent results across experiments sometimes being taken as further support for a theory. For example, Evans et al. (1994) took the pattern of results across their own experiments as well as Newstead et al. (1992) as supporting a modified form of mental model theory in which unbelievable conclusions lead to an extra step of verification. More recently, Thompson and Evans (2012) took the historic record of belief bias results as supporting a default-interventionist version of dual-process theory in which believable conclusions are generally accepted but unbelievable conclusions trigger a deeper form of reasoning. In both cases, the theoretical conclusions are aimed at explaining accuracy differences between believable and unbelievable conclusions that cannot be safely inferred based on previous research.

As reviewed in Section 1, belief bias is not only of longstanding theoretical interest for basic research on reasoning, but has also been studied in the context of neuroscience, emotion and cognition, individual differences, and informal argumentation. Our analyses suggest that those studies should be revisited. Moreover, our findings that ROCs are curved rather than linear for conditional reasoning suggest that decades of theoretical work based on traditional analyses of conditional reasoning experiments are also at risk. To give one recent and interesting example of a belief bias study, Stollstorff et al. (2013) concluded that when reasoning about emotional materials, people with a specific genetic variant linked to serotonin transport had relatively low accuracy on arguments with unbelievable conclusions. This apparent reduction of accuracy was explained in terms of a lack of inhibitory control for the group with that particular genetic variant. Essentially, different people with different genetic variants carried out different stages of processing. However, the finding of an accuracy difference between genetic groups, based on traditional analyses, may itself be incorrect, putting at risk the theoretical conclusions regarding inhibition and reasoning.

The present research extends Dube et al. (2010), Dube et al. (2011) in important ways. First, we have addressed not only the interaction index but two other measures commonly used in belief bias studies, the logic index and belief index. Second, we have provided simulations showing how all three traditional measures are likely to lead to incorrect conclusions. Third, we have compared traditional measures and SDT-based analyses in two new experiments on the theoretically important topic of whether the belief bias effect can be eliminated by instructions. Fourth, we have provided additional simulations, addressing other potential substitutes for traditional measures, such as d'-based analyses. Fifth, we have shown that curved, asymmetrical form of ROCs for syllogistic reasoning also appears in other forms of deduction, including conditional reasoning.

Indeed, our conclusions apply to any experimental situation in which difference score measures of response rates are used without clear evidence of linear ROCs for that task and domain (see Rotello et al., 2008; Fig. 1; and Jaeger, 2008, for related arguments). A recent survey of individual-subject ROCs in memory and perception tasks concluded that they are consistently and overwhelmingly curved (Dube & Rotello, 2012), like the group-level ROCs we have reported on a variety of reasoning tasks (Dube et al., 2010; Dube et al., 2011; Heit & Rotello, 2005; Heit & Rotello, 2008; Heit & Rotello, 2010; Heit & Rotello, 2012; Heit et al., 2012; Rotello & Heit, 2009). Moreover, Dube and Rotello reviewed many results from the memory, perception, and reasoning literatures suggesting that the curvature of ROCs does not depend on the method of elicitation (namely the confidence rating approach that has been our focus here).

The merits of SDT-based measures have been recognized more widely in some areas of cognition research than others. For example, many memory researchers have been sensitive to limitations of raw-score measures, instead turning to d', ROC analyses, and other model-based analyses (e.g., Cleary, 2005; Mickes, Johnson, & Wixted, 2010; Slotnick, 2010). In addition, SDT and ROC analyses are also used extensively in perception research (e.g., Guido, Lu, Vaughan, Godwin, & Sherman, 1995; Macmillan, Kaplan, & Creelman, 1977; Sewell & Smith, 2012). Likewise, in a review of the effects of social stereotypes

on memory, Stangor and Macmillan (1992) noted the limitations of raw scores, and some of the studies they reviewed used SDT analyses (but not ROCs). A recent line of work on social cognition has focused on the QUAD model (Sherman et al., 2008), which has been applied to numerous findings showing how implicit attitudes affect various judgments about newly observed stimuli, for example in implicit association tests (IAT). Although the QUAD model does have separate parameters for sensitivity and bias, some algebra shows that the correct response rate is a linear function of the error rate, meaning that this model also assumes linear ROCs. To the best of our knowledge, the assumption of linear ROCs in IATs is untested, and may be unwarranted. Turning briefly to studies of animal learning, we note that most of the classic studies of contingency judgment by animals, reviewed by Alloy and Tabachnik (1984), involved simple difference scores between response rates in various conditions, such as a comparison between predictable and unpredictable cues. These analyses assume linear ROCs, whereas Wixted and Squire (2008) have reported that ROCs derived from animal learning studies (involving rats or pigeons) have usually been curved.

With our own two experiments, we have shown that the curved and asymmetric ROCs lead traditional analyses (assuming linear ROCs) and d' analyses (assuming symmetrical ROCs) to incorrect conclusions about the nature of belief bias in syllogistic reasoning. Our own ROC-based analyses indicated that the belief bias effect is predominantly a response bias effect, and that bias shifts which occur as a consequence of instructions may be falsely interpreted as having an influence on reasoning accuracy. We do not rule out the possibility that by applying ROC-based analyses to subsets of data, even more nuanced conclusions could be drawn (cf., Trippas et al., 2013), and indeed in our own Experiment 2 there was evidence for a very small but statistically significant accuracy difference that is consistent with the results of Trippas et al. Instead, we emphasize that the application of SDT models allows such subtleties to be detected and measured accurately.

In most general terms, researchers must validate the assumptions of their analyses against the details of their data (Kinchla, 1994). In the case of traditional difference score measures of reasoning, most researchers have not checked the assumptions of their analyses, that is, they have not collected confidence ratings to assess the shape of ROC curves. We strongly suspect that in most cases, the ROCs would be curved, invalidating the traditional measures. Consequently, a great deal of reasoning research that has used traditional measures is at risk.

Acknowledgements

We thank Marios Eliades and Isabelle Blanchette for providing the raw data from Eliades et al. (2012), Henry Markovits for providing raw data from Markovits and Handley (2005) and Markovits et al. (2010), Lance Rips for providing raw data from Rips (2001), Wendy Contreras, Graham Ellis, and Aljanee Whitaker for their assistance in running experiments, Nicolas Raboy for programming assistance, John Dunn and Dries Trippas for feedback on a previous version of this manuscript, and Chad Dubé for many thoughtful discussions of these issues. This material is based upon work while Evan Heit was serving at the National Science Foundation (US). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix

Here we present and apply d'-based analogs of the logic, belief, and interaction indices. The three effects may be estimated as follows:

$$d'_{Logic} = z(P(\text{`valid'}|Valid)) - z(P(\text{`valid'}|Invalid))$$

$$d'_{Belief} = z(P(\text{`valid'}|Believable)) - z(P(\text{`valid'}|Unbelievable))$$

$$\begin{aligned} d'_{Interaction} = {} & z(P(\text{`valid'}|Valid, Unbel)) \\ & - z(P(\text{`valid'}|Invalid, Unbel)) \\ & - [z(P(\text{`valid'}|Valid, Bel)) \\ & - z(P(\text{`valid'}|Invalid, Bel))] \end{aligned}$$

In the absence of ROC data, is it better to use the traditional difference score approach or to use the d' based measures? Both measures are flawed. Difference scores entail the assumption of a linear ROC that has not been observed. The d' approach entails the assumptions of a curved ROC that is symmetric, implying equal-variance Gaussian evidence distributions. Although the curvature in the empirical ROCs is clearly consistent with the SDT model, the equal-variance assumption is consistently violated, rendering d' confounded with response bias. Simulations reported by Rotello et al. (2008) suggest that both d' and the difference score ($H - F$) measures are likely to lead to erroneous conclusions, with the probability of error being higher for the latter measure.

One way to reduce the probability of misinterpreting the data is to choose a conservative measure, one that is less likely to yield a significant result when its assumptions are violated. We used a bootstrapping strategy to assess the relative probabilities that d' and the traditional difference score measures would yield a significant result. We focused on the interaction effect, as this is considered to be the hallmark of the belief bias effect (e.g., Evans et al., 1983). To increase power, the data from the standard instruction conditions of both experiments were combined, yielding $N = 89$. For each of 2000 bootstrapped samples, 89 participants' data were randomly selected, with replacement. From the sampled data, we computed the interaction index using both the traditional, H–F, and d' based measures; these measures were compared against the null hypothesis value of 0 using a single-sample t-test. The results are shown in Fig. A1. In the figure, it is apparent that the measures are correlated. Correlation does not imply interchangeability of the measures, as is also apparent: The probability that the interaction effect is declared significant is much higher for H–F than for d'. If only one of those measure declares a significant effect, it is 24 times

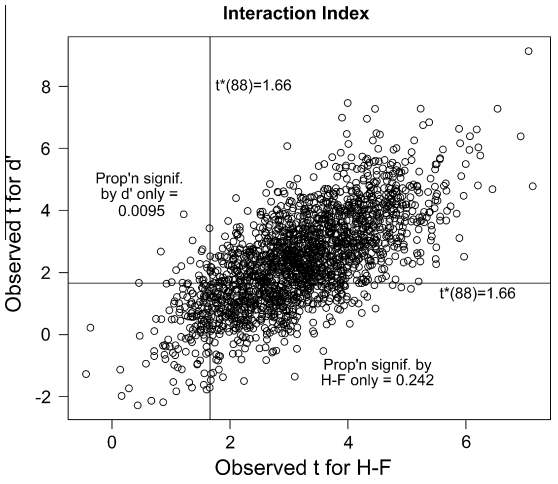

Fig. A1. Results of significance tests for the interaction effect, calculated using tradition measures (*x*-axis) or *d′* (*y*-axis), for 2000 bootstrapped samples from the combined data of the standard instruction conditions of both experiments.

more likely to be the *H–F* interaction. On this basis, we conclude that *d′* measures, though flawed, are superior to traditional difference score measures.

Despite that conclusion, *d′* analyses of our own two experiments were not particularly promising. We used *d′* to estimate the logic, belief, and interaction effects in each experiment, and compared the effects across instructional conditions using independent samples *t*-tests. Calculated *d′* values for our own two experiments are shown in Table 2.

In Experiment 1, the *d′* analysis implied that subjects were better able to discriminate valid from invalid problems in the standard instruction condition (d'_{Logic} = 0.92 v. 0.55, $t(87) = 2.18$, $p < .05$, Cohen's $d = 0.46$), although subjects in both conditions gave more "valid" responses to valid conclusions (standard: $t(44) = 7.70$, Cohen's $d = 1.15$; augmented: $t(43) = 4.51$, Cohen's $d = 0.68$). Subjects in both conditions made more "valid" responses to problems with believable conclusions (standard: $t(44) = 5.83$, Cohen's d = 0.87, augmented: $t(43) = 4.08$, Cohen's $d = 0.61$), and instructions did not influence the belief effect (d'_{Belief} = 0.63 v. 0.52, $t(87) = 0.64$, Cohen's $d = 0.14$). Finally, the *d′*-based interaction index was large and positive in the standard instruction condition (0.52, $t(44) = 4.03$, $p < .001$, Cohen's $d = 0.60$) but near 0 in the augmented instruction condition (0.03, difference $t(87) = 2.91$, $p < .01$, Cohen's $d = 0.04$). In other words, the *d′* analyses led to the same conclusions as the traditional difference score analyses, and differed from the conclusions based on ROCs.

In Experiment 2, *d′* analyses indicated that there were no effects of instruction condition for any measure (maximum $t(86) = 1.37$), but both d'_{Logic} and d'_{Belief} were greater than zero overall (logic: $t(87) = 8.91$, Cohen's $d = 0.95$; belief: $t(87) = 8.21$, Cohen's $d = 0.88$). In contrast, $d'_{\text{Interaction}}$ was not different than zero ($t(87) = 0.06$). Again, the *d′* analyses were consistent with the traditional measures rather than the ROCs.

To conclude, applying *d′* analyses holds some promise over traditional difference scores. The *d′* analyses are a better fit with the underlying data in terms of curved ROCs, but they still make an equal-variance assumption that has not been observed empirically. The *d′* analyses appear to be more conservative that traditional measures, e.g., they are less likely to indicate an interaction between logic and belief. Although applying *d′* analyses to our own data essentially led to the same conclusions as traditional analyses, based on simulations we would predict that *d'* analyses do generally have a better chance of coming to correct conclusions. Better still, of course, is to take the extra step of collecting confidence rating data so that the full ROC analysis may be conducted.

References

Alloy, L., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91*, 112–149.

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

Blanchette, I., & Campbell, M. (2012). Reasoning about highly emotional topics: Syllogistic reasoning in a group of war veterans. *Journal of Cognitive Psychology, 24*, 157–164.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717–726.

Cleary, A. M. (2005). ROCs in recognition with and without identification. *Memory, 13*, 472–483.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 130–151.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117*, 831–863.

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: A reply to Klauer and Kellen (2011). *Psychological Review, 118*, 155–163.

Eliades, M., Mansell, W., Stewart, A. J., & Blanchette, I. (2012). An investigation of belief-bias and logicality in reasoning with emotional contents. *Thinking & Reasoning, 18*, 461–479.

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, judgement 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. S. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning under time pressure. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie), 56*, 77–83.

Evans, J. St. B. T., Legrenzi, P., & Girotto, V. (1999). The influence of linguistic form on reasoning: The case of matching bias. *The Quarterly Journal of Experimental Psychology: Section A, 52*, 185–216.

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, K., Rotello, C. M., Li, X., & Rayner, K. (2009). Scene perception and memory revealed by eye movements and ROC analyses: Does a cultural difference truly exist? *Quarterly Journal of Experimental Psychology, 62*, 276–285 [PMCID: PMC2668147].

Goel, V., & Vartanian, O. (2011). Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cognition & Emotion, 25*, 121–131.

Griggs, R. A. (1989). To "see" or not to "see": That is the selection task. *The Quarterly Journal of Experimental Psychology, 41*, 517–529.

Guido, W., Lu, S.-M., Vaughan, J. W., Godwin, D. W., & Sherman, S. M. (1995). Receiver operating characteristic (ROC) analysis of neurons in the cat's lateral geniculate nucleus during tonic and burst response model. *Visual Neuroscience, 12*, 723–741.

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., & Evans, J. St. B. T. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning, 10*, 175–195.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.). *Psychology of learning and motivation* (Vol. 39, pp. 163–199). San Diego: Academic Press.

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

Heit, E., & Rotello, C. M. (2005). Are there two kinds of reasoning? In *Proceedings of the twenty-five annual meeting of the cognitive science society* (pp. 923–928). Mahwah, NJ: Erlbaum.

Heit, E., & Rotello, C. M. (2008). Modeling two kinds of reasoning. In *Proceedings of the thirtieth annual meeting of the cognitive science society* (pp. 1831–1836). Austin, TX: Cognitive Science Society.

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 805–812.

Heit, E., & Rotello, C. M. (2012). The pervasive effects of argument length on inductive reasoning. *Thinking & Reasoning, 18*, 244–277 (Special issue on Reasoning and Argumentation).

Heit, E., Rotello, C. M., & Hayes, B. K. (2012). Relations between memory and reasoning. In B. Ross (Ed.). *Psychology of Learning and Motivation* (Vol. 52, pp. 57–101). Amsterdam: Elsevier.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101*, 166–171.

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). *Psychological Review, 118*, 164–173.

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

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

Macmillan, N. A., Kaplan, H. L., & Creelman, D. (1977). The psychophysics of categorical perception. *Psychological Review, 84*, 452–471.

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., Forgues, H. L., & Brunet, M.-L. (2010). Conditional reasoning, frequency of counterexamples, and the effect of response modality. *Memory & Cognition, 38*, 485–492.

Markovits, H., & Handley, S. (2005). Is inferential reasoning just probabilistic reasoning in disguise? *Memory & Cognition, 33*, 1315–1323.

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.

Mickes, L., Johnson, E. M., & Wixted, J. T. (2010). Continuous recollection versus unitized familiarity in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 843–863.

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

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

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

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139*, 1173–1203.

Platt, R. D., & Griggs, R. A. (1993). Facilitation in the abstract selection task: The effects of attentional and instructional factors. *The Quarterly Journal of Experimental Psychology, 46*, 591–613.

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

Pollard, P., & Evans, J. St. B. T. (1987). On the relationship between content and context effects in reasoning. *American Journal of Psychology, 100*, 41–60.

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

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science, 12*, 129–134.

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., 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., 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.

Schyns, P. G., & Oliva, A. (1999). Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition, 69*, 243–265.

Sellen, J. L., Oaksford, M., & Gray, N. S. (2005). Schizotypy and conditional reasoning. *Schizophrenia Bulletin, 31*, 105–116.

Sewell, D. K., & Smith, P. L. (2012). Attentional control in visual signal detection: Effect of abrupt-onset and no-onset stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 1043–1068.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115*, 314–335.

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.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35*, 1–33.

Slotnick, S. D. (2010). "Remember" source memory ROCs indicate recollection is a continuous process. *Memory, 18*, 27–39.

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

Stangor, C., & Macmillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111*, 42–61.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven: Yale.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672.

Stollstorff, M., Bean, S. E., Anderson, L. M., Devaney, J. M., & Vaidya, C. J. (2013). Rationality and emotionality: Serotonin transporter genotype influences reasoning bias. *Social Cognitive and Affective Neuroscience, 8*, 404–409.

Stupple, E. J. N., Ball, L. J., Evans, J. St. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology, 23*, 931–941.

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

Thompson, V., & Evans, J. St. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning, 18*, 278–310.

Trippas, D., Handley, S. J., & Verde, M. J. (2013). The SDT model of belief bias: Complexity, time and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1393–1402.

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.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20*, 273–281.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7*, 275–278.

Wixted, J. T., & Squire, L. R. (2008). Constructing receiver operating characteristics (ROCs) with experimental animals: Cautionary notes. *Learning & Memory, 15*, 687–690.