A Two-Step Signal Detection Model of Belief Bias

Rachel G. Stephens (r.stephens@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW 2052 Australia

John C. Dunn (john.dunn@uwa.edu.au)
School of Psychological Science, University of Western Australia
Perth, WA 6009 Australia

Brett K. Hayes (b.hayes@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW 2052 Australia

Abstract

When asked to assess the deductive validity of an argument, people are influenced by their prior knowledge of the content. Recently, two competing explanations for this belief bias effect have been proposed, each based on signal detection theory. Under a response bias explanation, people set more lenient decision criteria for believable than for unbelievable arguments. Alternatively, believable and unbelievable arguments may differ in subjective argument strength for both valid and invalid items. Two experiments tested these accounts by asking participants to assess the validity of categorical syllogisms and rate their confidence. Conclusion-believability was manipulated either within- or between-groups. A two-step signal detection model was applied to examine the effects on the relative location of the decision threshold and the distributions of argument strength. Equivalent belief bias effects were found when believability was manipulated within- and between-groups, supporting the view that the belief bias effect is due to response bias.

Keywords: belief bias; deductive reasoning; signal detection theory; response bias

Introduction

An important phenomenon for theories of reasoning is that people show a *belief bias* when asked to assess the logical validity of arguments. The tendency to accept or reject a conclusion as valid is not based purely on logical structure but is also swayed by its compatibility with prior knowledge (e.g., Evans, Newstead, & Byrne, 1993; Markovits & Nantel, 1989; Shynkaruk & Thompson, 2006). Table 1 shows typical stimuli – categorical syllogisms – in which the validity of the argument is crossed with the believability of the conclusion. In the validity discrimination task, participants are asked to judge whether the conclusion below the line necessarily follows from the premises above the line. Key findings based on arguments like these are that people are more likely to endorse valid than invalid arguments, but they are also more likely to endorse arguments with believable than with unbelievable conclusions. In many cases these factors also interact; for example, the difference between the acceptance rates of valid and invalid arguments is often greater for unbelievable than for believable arguments (e.g., Dube, Rotello, & Heit, 2010; Evans, Barston, & Pollard, 1983; Newstead, Pollard, Evans, & Allen, 1992; Roberts & Sykes, 2003).

Table 1: Sample syllogisms.

	Believable	Unbelievable
Valid	No beers are krabbers. Some krabbers are drinks.	No drinks are krabbers. Some krabbers are beers.
	Some drinks are not beers.	Some beers are not drinks.
Invalid	No drinks are krabbers. Some krabbers are beers.	No beers are krabbers. Some krabbers are drinks.
	Some drinks are not beers.	Some beers are not drinks.

Such effects are often seen as evidence that believability affects the quality of deductive reasoning – people's ability to distinguish valid from invalid arguments (see Dube et al., 2010 for a review). Theoretical accounts such as the *selective scrutiny* model (Evans et al., 1983), *misinterpreted necessity* model (e.g., Markovits & Nantel, 1989; Newstead et al., 1992) or the *mental models* approach (e.g., Oakhill, Johnson-Laird, & Garnham, 1989) propose explanations in which believability affects how validity is evaluated.

However, deciding whether an argument is valid also involves *response bias* – the willingness to endorse the argument, regardless of one's ability to discriminate valid and invalid forms. Controversially, recent work has used confidence ratings and signal detection theory to show that belief bias only reflects changes in response bias. That is, people are more willing to respond "valid" for believable arguments (Dube et al., 2010; Trippas et al., 2014). In this view, believability does not change one's subjective evaluation of argument validity.

In reaction to this response bias account, it has been suggested that data patterns consistent with changes in response bias can also be explained by believability affecting the subjective strength of both valid and invalid arguments (Klauer & Kellen, 2011; Singmann & Kellen, 2014). Under this alternative *argument strength* account, if an argument has a believable conclusion (whether valid or

invalid) then it will be viewed as more logically valid and thus garner more endorsements.

In adjudicating between these accounts, a key consideration is that believability is usually manipulated within a single experimental session. The argument strength account is consistent with evidence that response bias is unlikely to change from trial to trial (e.g., Stretch & Wixted, 1998). However, to our knowledge it is currently unknown how believability affects performance if instead it is manipulated *between* different groups of participants, where response bias is free to differ. As we explain below, we hypothesized that if the response bias account is correct then the same belief bias effects on model parameters should appear when believability is manipulated within groups and between groups (i.e., equivalent ordinal effects on response bias and no effects on discriminability).

Given the important implications for theories of reasoning, we aimed to extend the investigation of response bias in deductive reasoning. We took three key steps. First, we sought to replicate the within-group findings of Dube et al. (2010) – including confidence ratings – that they used to support the response bias account. Second, we applied an extended signal detection model that was specifically tailored to the two-step task in which participants first make a binary valid/invalid decision, then rate their confidence. Our goal was to confirm whether such a model would still suggest that believability does not affect accuracy, but that the response bias or argument strength accounts are required. Third, to avoid the issue of whether response bias can change trial-by-trial, in a second experiment we manipulated believability between groups. Our goal was to examine whether the key effects generalized to this design, which would support the response bias account.

To this end, in the following sections, we outline (a) how signal detection theory can be applied to deductive reasoning, (b) the novel two-step signal detection model, and (c) two experiments that manipulate believability within or between groups, to which we apply the model.

Signal Detection Theory and Belief Bias

Signal detection theory (SDT) is a useful framework to examine belief bias because it allows us to separate changes in discriminability (i.e., differentiating valid and invalid arguments) versus response bias (i.e., the "decision stage"; cf. Dube et al., 2010; Rotello & Heit, 2009). In this framework, arguments fall along a continuum of subjective argument strength, with distinct Gaussian distributions for valid and invalid arguments, as shown in Figure 1. The distance between the means of these distributions reflects how well people can distinguish valid and invalid arguments. People also set a response threshold along the continuum, endorsing any argument that exceeds it in strength (i.e., the tallest "Invalid"/"Valid" threshold in the figure). Thus the *hit rate* (endorsement rate for valid arguments) is given by the area under the valid distribution to the right of the threshold, and the *false alarm rate* (endorsement rate for invalid arguments) is given by the area under the invalid distribution to the right of the threshold. Two important ways that performance can change is by the threshold shifting (i.e., changes in response bias), and/or the valid distribution shifting relative to the invalid distribution (i.e., changes in discriminability or sensitivity).

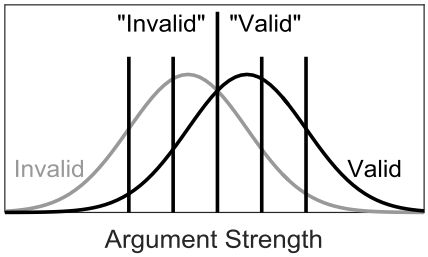

Figure 1: Standard signal detection model.

Adding confidence judgments to the validity discrimination task allows for a more fine-grained analysis of changes in signal detection parameters. It is assumed that people set a response threshold for *n*-1 response options on the confidence scale – five are shown in Figure 1 for a six-point confidence scale. Performance can then be examined using receiver operating characteristic (ROC) curves, which plot hit rates against false alarm rates at different confidence levels (see examples in Figure 2). Evidence for a difference in the discriminability of valid and invalid arguments would be suggested by points from two conditions falling on different curves. Better discrimination is suggested by ROC curves that fall further from the diagonal, towards the upper left – hit rates are higher relative to false alarm rates. In contrast, conventional evidence for a difference in response bias is suggested by points from two conditions falling on different positions along the same curve. A more lenient threshold is suggested by points sitting further towards the right, corresponding to both higher hit rates and higher false alarm rates. Signal detection models can be fit to ROC curves to test for changes in argument discrimination or response bias, which would be supported by reductions in fit due to constraining either the relative location of the valid distribution or the criteria, respectively.

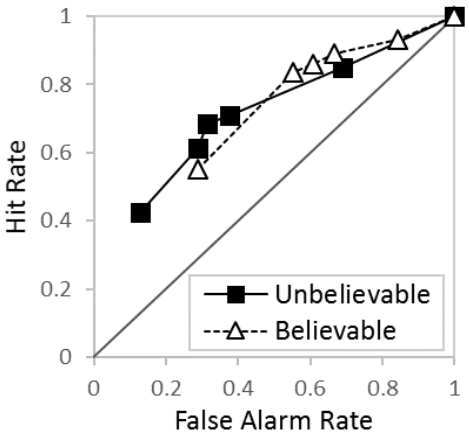

Figure 2: ROC curves from Dube et al. (2010).

An important but controversial result was reported by Dube et al. (2010), who compared and fit ROC curves for believable and unbelievable syllogisms like those in Table 1. Their ROC model fitting showed that argument believability affected response bias but did not affect discriminability. Participants were simply more willing to endorse believable arguments (see Figure 2). This *response bias* account of belief bias is illustrated in the top panel of Figure 3. Here there are two distributions – one for invalid and one for valid arguments – but two sets of decision thresholds – a more lenient set for believable arguments, and a more conservative set for unbelievable arguments (only three criteria per set are shown, to avoid clutter). A similar account has been proposed for belief bias in causal conditional arguments such as modus ponens (Trippas et al., 2014).

However, this is not the only way to interpret overlapping ROC curves. The response bias interpretation has been contested because an alternative *argument strength* account is possible, as illustrated in the bottom panel of Figure 3 (Klauer & Kellen, 2011; Singmann & Kellen, 2014). This approach assumes a single fixed set of decision thresholds, but four different distributions – distinct invalid and valid distributions for both unbelievable and believable arguments. Discriminability is assumed to be the same for believable and unbelievable arguments, but the believable-valid AND believable-invalid distributions are shifted to the right (i.e., they are stronger on average).

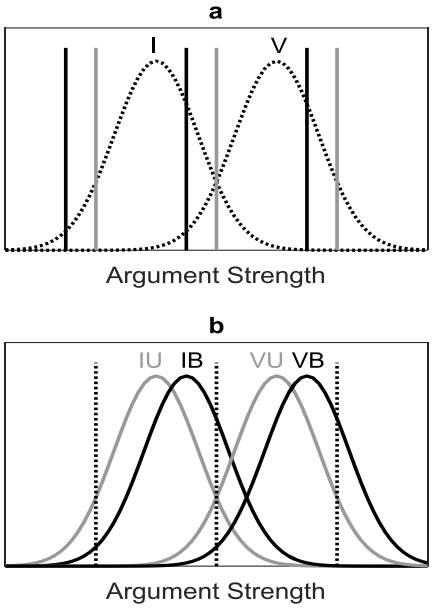

Figure 3: (a) The response bias account. There are fixed invalid (I) and valid (V) distributions. Criteria are shifted to the left for believable arguments (black lines) relative to unbelievable arguments (grey lines). (b) The argument strength account. There are fixed criteria. Invalid-believable (IB) and valid-believable (VB) distributions are shifted to the right, relative to the invalid-unbelievable (IU) and valid-unbelievable (VU) distributions.

Resolving this debate has been difficult because in many of the key studies (e.g., Dube et al., 2010; Trippas et al., 2014), believability has been manipulated within a block of arguments. The response bias account assumes that people will shift their criteria on a trial-by-trial basis, depending on whether an argument is believable or unbelievable. However, this assumption is controversial. In the recognition memory literature, although trial-by-trial shifts in criteria are possible, it appears that often this does not occur (Rotello & Macmillan, 2007; Starns & Olchowski, 2015; Stretch & Wixted, 1998). One way to address this issue is to manipulate believability *between* participants. Uncontroversially, different groups are then free to set different response criteria.

In order to resolve whether belief bias is driven by changes in response bias or argument strength, we carried out two experiments and tested a new signal detection model of reasoning. Experiment 1 confirmed that we could replicate the ROC shifts found by Dube et al. (2010), with believability manipulated within-participants. In Experiment 2, we investigated whether the same effects appeared when believability was manipulated between groups. If the response bias account is correct, then the same distributions of response strength for valid and invalid arguments should apply to those seeing only believable or unbelievable arguments (because there are only two distributions), but the groups will differ in response criteria. Therefore, we would see different hit rates and false alarm rates for the believable and unbelievable argument groups, replicating the Dube et al. (2010) ROC shifts and differences in the response criterion parameter based on model fitting.

Alternatively, if the argument strength account is correct, then the pair of invalid and valid distributions would be in different locations for believable and unbelievable groups. However, each group would be free to set criteria relative to the locations of their invalid and valid distributions – each group has no reason to adopt criteria that are in different locations relative to their distributions. Therefore, we would see the same hit and false alarm rates for both groups, with no ROC shifts nor differences in the criterion parameter.

Accurately testing the competing accounts of belief bias requires model fitting with a model that properly captures the task. Therefore, we extended the signal detection model developed by Dube et al (2010), to treat the valid/invalid decision and confidence judgments as two separate steps. As outlined below, this kind of model is more appropriate for the two-step task than a traditional signal detection model (Moran, Teodorescu, & Usher, 2015). We first present the model. We then report experiments using within- and between-participant manipulations of conclusion believability and fit the model to these data.

Two-Step Signal Detection Model

In the two-step validity discrimination task that we use, participants make a "valid"/"invalid" decision, and then rate their confidence. Despite the sequential nature of these judgments, in the standard procedure for generating

empirical ROC curves, data from the response categories are recoded to form a single scale with judgments that range from high-confidence "valid" to low-confidence "valid", then low-confidence "invalid" to high-confidence "invalid" (e.g., Dube et al., 2010; Trippas et al., 2014). Typically, these ROC curves are then fit using the standard single-step SDT model that we outlined above, with a criterion parameter separating each adjacent pair of recoded confidence levels. However, visual inspection of these empirical ROC curves suggests that they differ from the smooth concave curve typically found – they instead exhibit a "hinge" or "elbow" where valid and invalid response categories join, as apparent in Figure 2, particularly for the unbelievable-ROC. In order to successfully model this feature, the standard SDT model was extended to incorporate changes in evidence accumulation and variability in the period between the initial validity judgment and the subsequent confidence judgment.

The two-step SDT model is similar to a standard SDT model with the exception that confidence judgments are based on a noisy version of the evidence value on which the validity judgment was made. Let $x \sim N(\mu, \sigma)$ be the strength of given argument. Let *c* be a decision criterion such that if $x > c$, respond "valid", else respond "invalid". We propose that a confidence judgment is based on *x**, a noisy memory trace of argument strength, *x*. That is, $x^* = x + x'$, for $x' \sim N(\nu, \eta)$. If $\nu > 0$ then additional argument strength is accumulated in the interval between the two decisions (cf., Moran et al., 2015). Suppose, there are *k* confidence categories labeled, in sequence, from most confident to least confident. Then, associated with these category labels is set of points on the strength continuum, $U = \{u_0, u_1, \ldots, u_k\}$, such that, $u_0 < u_1 < \ldots < u_k$, $u_0 = -\infty$, $u_k = \infty$, and a set of points, $V = \{v_0, v_1, \ldots, v_k\}$, such that, $v_0 > v_1 > \ldots > v_k$, $v_0 = \infty$, $v_k = -\infty$. Then, if the response is "invalid" and $u_i < x^* \leq u_{i+1}$ or, if the response is "valid" and $v_i \leq x^* < v_{i-1}$ then respond with the *i*th category label.

The hypotheses of interest were primarily tested by comparing the fits of nested versions of this model using the likelihood ratio test. Although the response bias and argument strength accounts are formally identical for a traditional signal detection model, this is not strictly true for the two-step model. Therefore, both accounts can be tested when believability is manipulated within-participants.

Experiments

In two experiments, participants evaluated the validity of categorical syllogisms, which included logically valid and invalid arguments with believable or unbelievable conclusions in a 2x2 design. Experiments 1 and 2 manipulated believability within- and between-groups, respectively.

Method

Participants. One-hundred-and-seventeen students (30 males) at the University of New South Wales, Sydney, participated for course credit. Mean age was 18.8 years (SD = 2.3). Participants were randomly allocated to Experiment 1 (N = 38) or one of the groups in Experiment 2 (believable N = 40, unbelievable N = 39).

Stimuli. In Experiment 1, participants evaluated 64 arguments across two blocks of 32 trials, with 16 believable and 16 unbelievable arguments per block – half of which were valid in each case. In Experiment 2, participants evaluated either 32 believable or 32 unbelievable arguments (half valid).

Example stimuli are shown in Table 1. The arguments were based on those of Experiment 2 by Dube et al. (2010), and were constructed using their 16 syllogistic problem frames (e.g., *All X are Y; Some Z are not Y; Therefore some Z are not X*). Half were valid and half were invalid. Each problem frame had the conclusion structure, *Some Z are not X* (or *Some X are not Z*), and was assigned content involving a category-exemplar relationship (e.g., drinks-beers, dogs-poodles, plants-weeds).

Conclusion believability was manipulated by simply reversing the order of the category and exemplar (e.g., *Some drinks are not beers* vs. *Some beers are not drinks*). We verified the believability of the conclusion statements in a separate study by 34 people drawn from a similar population to the main experiments. Based on ratings on a 5-point scale (1 = unbelievable, 3 = neutral, 5 = believable), the 32 statement pairs with the most extreme average ratings were selected from a set of 38 pairs (Believable: M = 4.95, SD = 0.09; Unbelievable: M = 1.59, SD = 0.35). To minimize the effects of premise believability, the premises included a nonsense term (e.g., *krabbers*, *junids*).

The semantic content was split into four subsets of eight category-exemplar pairs, so the content could be assigned to all four believability-by-validity conditions, counterbalanced across participants. Experiment 1 participants (believable and unbelievable within-participants) saw the category-exemplar content once per block and the 16 problem frames twice per block (once as believable and once as unbelievable versions), forming the 64 arguments over two blocks. Content assignment was controlled for this group so that in the second block, each participant saw the same content in the same problem structures as in their first block, but with conclusion believability reversed. At the start of the second block, these participants were warned that there would be similar content but the specific arguments would be different. Experiment 2 participants (believable-only and unbelievable-only groups) saw each category-exemplar content once and the 16 problem frames twice, forming the 32 arguments.

Before beginning the experiment, all participants received two valid and two invalid practice problems with abstract content (e.g., "All M are P…") and different structures that were not included in the main task.

Procedure. Participants were shown the set of arguments in random order, presented one-by-one on a computer, with a line separating the conclusion from the premises. The instructions asked participants to assume that the premises were true and assess whether the conclusion logically followed from them. *Valid* arguments were defined as those for which the sentence below the line was *necessarily true*, given that the information above the line was true (and *invalid = not necessarily true*). Participants were told that the arguments would contain a nonsense word. A trial counter was presented at the top left corner of the screen. Participants clicked on either the "Valid" or "Invalid" button presented underneath a given argument, then rated their confidence on a scale that appeared, ranging from 50 (Guessing) to 100 (Certain) in increments of ten.

Results

Both experiments replicated previously observed argument endorsement patterns and belief bias effects (see Table 2; e.g., Dube et al., 2010; Evans et al., 1983; Newstead et al., 1992). Analysis of variance (ANOVA) revealed that participants endorsed (i.e., responded "valid") valid arguments more often than invalid arguments: Experiment 1, $F(1, 37) = 64.28$, $p < .001$, $\eta^2 = .35$; Experiment 2, $F(1, 77) = 127.24$, $p < .001$, $\eta^2 = .40$. Participants endorsed believable arguments more often than unbelievable arguments: Experiment 1, $F(1, 37) = 38.59$, $p < .001$, $\eta^2 = .12$; Experiment 2, $F(1, 77) = 19.92$, $p < .001$, $\eta^2 = .13$. Notably, as shown in the Table, there was a larger difference between the acceptance rates of valid and invalid arguments for unbelievable than for believable arguments: Experiment 1, $F(1, 37) = 5.50$, $p = .02$, $\eta^2 = .01$; Experiment 2, $F(1, 77) = 9.81$, $p = .002$, $\eta^2 = .05$.

Table 2: Performance in Experiments 1 and 2. Hit rate is p("Valid"|Valid); False alarm rate is p("Valid"|Invalid).

Experiment	Condition	Hit rate	False alarm rate
1	Believeable	0.83	0.56
	Unbelievable	0.71	0.37
2	Believeable	0.82	0.58
	Unbelievable	0.75	0.34

The ROC curves for each experiment are presented in Figure 4 (unfilled points). Both show effects that are consistent with shifts in response criteria and comparable to Dube et al. (2010; cf. Figure 2), although we used more confidence response options. In each experiment, the points for believable and unbelievable arguments fall on similar curves, though the believable points are shifted further to the top-right corner than the unbelievable points.

We first fit an *unconstrained* two-step signal detection model to each experiment. As shown by the filled points in Figure 4, the predicted ROC points correspond reasonably well with the empirical results for both experiments, though there are some small departures for Experiment 1: Experiment 1, $G^2(12) = 22.54$, $p = .03$; Experiment 2, $G^2(12) = 15.83$, $p = .20$.

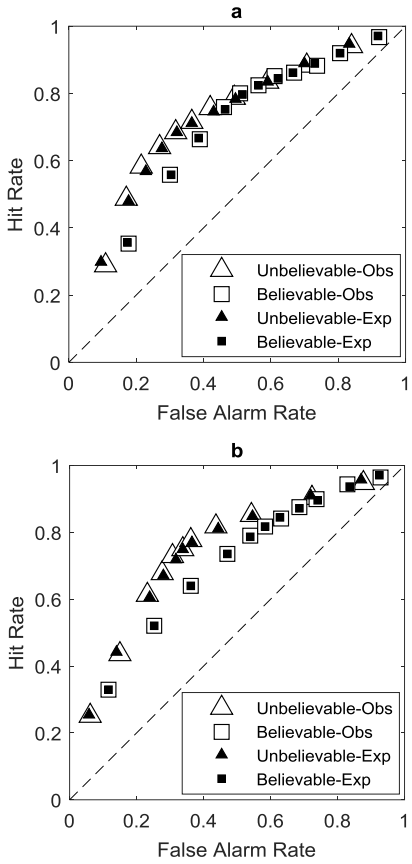

Figure 4: Observed ROC curves (Obs) and expected scores from the unconstrained model (Exp), for Experiments 1 and 2 (panels a and b, respectively).

We compared this unconstrained model against two nested models: a *constant discriminability* model and a *constant criterion* model in which (respectively) discriminability or the "valid"/"invalid" decision criterion for the initial binary judgment was constrained across believable and unbelievable conditions. For both experiments, the fit of the constant discriminability model did not significantly differ from that of the unconstrained model: Experiment 1, $G^2(1) = 0.23$, $p = .63$; Experiment 2, $G^2(1) = 0.001$, $p = .97$. This shows that, in line with Dube et al. (2010), discriminability did not differ between believability conditions.

The constant criterion model led to a reduction in fit compared to the unconstrained model: Experiment 1, $G^2(1) = 47.89$, $p < .001$; Experiment 2, $G^2(1) = 77.75$, $p < .001$. This indicates that, in line with the response bias account, the "valid"/"invalid" decision threshold differed between believability conditions. Importantly, this was true both when believability was manipulated within-groups (Experiment 1) and between-groups (Experiment 2).

When a (non-nested) variant of the two-step model was applied to Experiment 1 that allowed the believable distributions to shift (i.e., the argument strength account), we found that it also provided a satisfactory fit to the data: $G^2(20) = 30.18$, $p = 0.07$. In other words, an argument

strength account of belief bias could also explain the Experiment 1 data. Such a model cannot sensibly be applied to Experiment 2. Nevertheless, as we argued above, the response bias account can more readily explain belief bias effects that occur between-groups.

Discussion

We investigated whether belief bias effects in deductive reasoning could be explained as a response bias effect. Experiment 1 replicated the belief bias effects of Dube et al. (2010), with conclusion believability manipulated within-block. We applied a new two-step signal detection model to better suit the two-step task, and confirmed that belief bias effects are consistent with a shift in response bias, rather than discriminability. Experiment 2 extended the same results to an equivalent task with believability manipulated between-groups.

Under the response bias account (Dube et al., 2010; Trippas et al., 2014), this pattern is explained by a shift in decision threshold, such that there is a more lenient criterion for believable conclusions. Under the argument strength account (Klauer & Kellen, 2011; Singmann & Kellen, 2014), the belief bias effect reflects higher mean strength for believable-valid and believable-invalid arguments than for unbelievable-valid and unbelievable-invalid arguments.

It could be argued that participants in Experiment 1 were unlikely to change their criteria trial-to-trial for different levels of believability, favoring the argument strength account. However, this account would have difficulty with Experiment 2, where participants saw only believable or only unbelievable arguments. There, the two groups had no reason to position their criteria in different locations relative to their distributions. Thus if belief bias primarily reflects a change in argument strength, the belief bias effects should have disappeared. The fact that they did not suggests that the most plausible explanation of belief bias in the current data sets is a change in response bias.

Therefore, addressing the debate between response bias and argument strength accounts of belief bias, we agree that believable conclusions are most likely to affect the decision stage, lowering the decision threshold rather than appearing more logically valid. Just as people may require stronger evidence to endorse that an unusual event occurred (Starns & Olchowski, 2015), it seems that people also require stronger evidence to endorse a syllogism with an unbelievable conclusion. As Dube et al. (2010) concluded, this is problematic for theories of reasoning that propose that believability affects the process of evaluating validity (e.g., Evans et al., 1983; Markovits & Nantel, 1989; Newstead et al., 1992; Oakhill et al., 1989). Future work should address whether the same findings generalize to other reasoning problems such as causal conditionals.

References

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 941–947.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*(3), 831-863.

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., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: Reply to Dube, Rotello, and Heit (2010). *Psychological Review*, *118*(1), 164-173.

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

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

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

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

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

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

Rotello, C. M., & Macmillan, N. A. (2007). Response bias in recognition memory. *The Psychology of Learning and Motivation*, *48*, 61-94.

Singmann, H., & Kellen, D. (2014). Concerns with the SDT approach to causal conditional reasoning: A comment on Trippas, Handley, Verde, Roser, McNair, and Evans (2014). *Frontiers in Psychology*, *5*, 402.

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

Starns, J. J., & Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. *Memory & Cognition*, *43*, 49-59.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379-1396.

Trippas, D., Verde, M. F., Handley, S. J., Roser, M. E., McNair, N. A., & Evans, J. St. B. T. (2014). Modeling causal conditional reasoning data using SDT: Caveats and new insights. *Frontiers in Psychology*, *5*, 217.