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

$$\begin{array}{l} \text{No addictive things are inexpensive.} \\ \text{Some cigarettes are inexpensive.} \\ \hline {}^{*}\text{Therefore, some addictive things are not cigarettes.} \end{array} \quad (1)$$

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