Acta Psychologica 180 (2017) 1–7

Contents lists available at ScienceDirect

Acta Psychologica

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

Exploring the determinants of confidence in the bat-and-ball problem

Aba Szollosi[a,b,*], Bence Bago[c], Barnabas Szaszi[b,d], Balazs Aczel[b]

[a] School of Psychology, University of New South Wales, Sydney, Australia
[b] Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[c] Paris Descartes University, Paris, France
[d] Doctoral School of Education, ELTE Eötvös Loránd University, Budapest, Hungary

ARTICLE INFO

Keywords:
CRT
Confidence
Error-monitoring
Attribute substitution
Verification

ABSTRACT

People often fail to solve deceptively simple mathematical problems, a tendency popularly demonstrated by the bat-and-ball problem. The most prominent explanation of this finding is that, to spare cognitive effort, people substitute the difficult task with an easier one, without being aware of the substitution. Despite this latter assumption, recent studies have found decreased levels of post-decision confidence ratings when people gave the answer of an easier calculation, suggesting that people are sensitive to their errors. In the current study, we investigated a mechanism that might be responsible for such a decrease in people's confidence ratings when they make errors: their attempts to make certain that their answer is correct (verification) and the perceived level of task difficulty (verifiability). We found that these two factors predicted people's confidence, suggesting that people's self-assessment of the perceived task difficulty and of their attempt to verify their response might determine their confidence. Implication for current models of post-decision confidence on reasoning problems is discussed.

Certain mathematical tasks, although not particularly difficult, can lure people into giving seemingly correct, yet inaccurate answers. Such behaviour is frequently demonstrated by the Cognitive Reflection Test (CRT; Frederick, 2005). Administering the three tasks of this test is commonly used to measure people's disposition or ability to reflect on their intuitive but incorrect answers (e.g., Toplak, West, & Stanovich, 2011; Travers, Rolison, & Feeney, 2016). For example, in the bat-and-ball problem of the test, participants are asked to solve the following mathematical task: "*A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost*?". "10 cents" is a regularly received, but incorrect answer to this problem, whereas the correct answer is "5 cents".

The most prevalent explanation for this finding is provided by the notion of *attribute substitution* (Kahneman & Frederick, 2005), which proposes that when people are faced with a difficult problem, they tend to substitute it with an easier one in order to spare cognitive resources. For example, the substitution for the bat-and-ball problem would be: "*A bat and a ball cost $1.10. The bat costs $1.00. How much does the ball cost*?". As here the term "*more than the ball*" is missing, it is not surprising that most people give the "10 cents" answer, as this is the correct answer for this easier, substituted task.

An important aspect of the attribute substitution theory is that it proposes that people's errors on this task reflect a lazy or "lax" monitoring (De Neys & Glumicic, 2008; Kahneman & Frederick, 2002). To test this assumption, De Neys, Rossi, and Houdé (2013) asked participants to solve both incongruent and congruent versions of the bat-and-ball problem. They used modified versions of the original bat-and-ball task for the incongruent version, and easier (substitution) tasks for the congruent version. Contradicting the predictions of attribute substitution theory, they found that participants were less confident when they gave the incorrect answer on the incongruent task (10 cents) than when they gave the correct answer on the congruent task (10 cents). This is unexpected since if participants substitute the more difficult task without monitoring the mistake, their confidence should not differ between the two tasks.

A proposed explanation for these finding comes from the conflict detection studies in decision-making (for a summary, see De Neys, 2012; for a discussion see e.g., Aczel, Szollosi, & Bago, 2016; Pennycook, Fugelsang, & Koehler, 2012; Singmann, Klauer, & Kellen, 2014). According to this explanation, people's reasoning errors cannot be just the result of insufficient or lax monitoring, but it also stems from the inability to inhibit the modal answer. Indeed, people have been shown to be sensitive to their errors, reflected on a variety of measures ranging from confidence ratings (De Neys, Cromheeke, & Osman, 2011), and response time measures (De Neys & Glumicic, 2008), to neuro-imaging data (De Neys, Vartanian, & Goel, 2008). In the case of

* Corresponding author at: School of Psychology, University of New South Wales, Sydney, NSW 2052, Australia.
E-mail address: aba.szollosi@gmail.com (A. Szollosi).

http://dx.doi.org/10.1016/j.actpsy.2017.08.003
Received 6 February 2017; Received in revised form 23 May 2017; Accepted 6 August 2017
Available online 10 August 2017
0001-6918/

the bat-and-ball problem, people's sensitivity was indicated by, for example, their decreased confidence ratings when they gave the erroneous "10 cents" answer compared to when they gave the accurate "5 cents" answer (De Neys et al., 2013; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Johnson, Tubau, & De Neys, 2016).

Although the empirical data clearly suggest that people are not as blind to their errors on reasoning tasks as the lax-monitoring view would suggest, it is still unclear what the specific mechanism is that alert people about their errors. Therefore, in the current study, we explored a potential mechanism, people's evaluation of their own answer verification, that can drive the decrease of people's confidence when they make errors. Specifically, we surveyed participants about their perceived task-difficulty and about their attempt to verify their answer. We expected that people are more likely to think that the task was more difficult, and/or they think they were not able to sufficiently verify their answer, when they make errors on a more difficult (in this case, incongruent) task. We also expected that they are less confident in their given answer in these cases.

Conceptually, we defined *verification* as people's self-reported attempt to make certain that the given answer is correct. For example, in one of the earliest models of problem-solving, the "test-operate-test-exit" model (Miller, Galanter, & Pribram, 1960), verification is defined as testing an achieved state against the goal state in order to determine whether further mental operations are needed. On the bat-and-ball problem one can verify their initial answer by substituting the data from the task into a formula (e.g., into the correct formula: $x + [x + \$1.00] = \$1.10$). We expected that people who give the correct answer in the congruent task will report that they verified their answers more than those who give the incorrect answer in the incongruent condition. We also expected that those who report that they verified their answer will be more confident in their answers compared to those who did not.

Along with verification, we also measured the monitoring of the *difficulty* by asking how difficult they found the task to verify. We surveyed this measure since perceived task difficulty may have a different effect on confidence than the actual act of verification itself. We hypothesised that people will regard the incongruent problems as more difficult to verify than the congruent problems and that the harder they find the task to verify, the lower they would rate their confidence. For example, when solving a simple task, such as "$1 + 1 = x$", we expect that people will be more confident in their answer compared to when they solve a more difficult one, such as "$(x - 1)^2 = (3 - 4x)^3$".

1. Experiment 1

1.1. Methods

1.1.1. Participants

The sample consisted of 233 university students (184 women; $M = 21.56$ years, $SD = 4.2$ years), all native speakers of Hungarian, who received course credit in exchange for participating. We omitted the data of five participants for not responding to all the questions they were asked. Data collection was approved by the institutional ethics committee and informed consent was obtained from all participants.

Table 1
Frequency of reporting verification in Experiment 2.

Condition - accuracy	Verification
Incongruent – incorrect ($n = 76$)	48.7% (37)
Incongruent – correct ($n = 27$)	66.7% (18)
Congruent – correct ($n = 111$)	59.5% (66)

Note. The number of participants are in parentheses.

1.1.2. Materials and procedure

In an online survey, alongside other heuristics and biases tasks, we asked participants to complete either a congruent or an incongruent version of the two modified CRT tasks, constructed by De Neys et al. (2013). In a between-subjects design, each participant answered only one question after an unrelated task. First, the participants saw the description of the task. After they indicated that they understood the description, the question appeared. To make sure that participants did not use any external help to solve the question, a 15 s time-pressure was applied. After giving an answer, we asked the participants to rate how confident they were that the answer they gave was the correct answer (0 – completely unconfident; 100 – completely confident). We then asked them to indicate whether they verified their answer (Yes/No)[1] and then how verifiable they thought the question was (1 – very easy to verify; 5 – very hard to verify). Each question appeared on a different page, without the possibility of returning to previous pages. The tasks were translated to Hungarian; English version of the tasks and all questions are available in the Appendix A.

1.2. Results

In all analyses, we excluded respondents ($n = 10$) who did not give the heuristic or the normative answer (5, 10, 45 or 90). Response accuracies were in line with previous observations: correct answers were collected from 26.2% of the participants in the incongruent group and 96.5% of the congruent group.

1.2.1. Verification

First, we expected that more people would report that they did not verify their answer among incorrect than among correct responders. Therefore, we compared the correct responders on the congruent and on the incongruent task to the incorrect responders on the incongruent task. The proportion of self-reported verification (Table 1) was higher for correct responders in both the congruent (66.7%) and incongruent (59.5%) groups than for the incorrect responders in the incongruent group (48.68%). However, chi-square tests revealed that this difference was not significant in either the congruent correct, incongruent incorrect comparison, $\chi^2(1, N = 187) = 1.70$, $p = 0.192$, or in the incongruent correct, incongruent incorrect comparison, $\chi^2(1, N = 103) = 1.92$, $p = 0.166$.

For the analyses of confidence, we dropped correct responders in the incongruent and incorrect responders in the congruent condition ($n = 31$) in all further tests, following De Neys et al. (2013) procedure, since we aimed to investigate the purported substitution process. As Fig. 1 displays, confidence was at ceiling in the congruent condition both for those who reported verification and for those who reported no verification (both $Mdns = 100$). In the incongruent condition, however, the confidence of participants who reported that verification they did not verify their answer were less confident in their answer ($Mdn = 70$) than those who reported that they did ($Mdn = 100$).

To statistically test this pattern, we aimed to predict confidence as a function of congruency and verification. As apparent in Fig. 1, the confidence ratings were heavily skewed, violating the assumptions of linear regression. To handle this issue, we dummy coded the ratings to "full confidence" for confidence ratings that were equal to 100, and "uncertain" for all confidence ratings that were smaller than 100. Table 2 reveals that the proportion of the dummy-coded confidence ratings were similar to the original confidence ratings.

For a statistical test, we used a generalised mixed-effect model with a binomial link function in R (R Core Team, 2017). We predicted confidence from congruency, verification, and their interaction.

[1] Instead of the technical word "verification", in our questions we used the plain Hungarian word for "checking", which refers to the attempt to make certain whether one's answer is correct. For the ease of understanding in text, we use the word verification.