Psychological Topics, 28 (2019), 1, 115-124
Original Scientific Paper UDC: 159.955
doi:https://doi.org/10.31820/pt.28.1.6

Further Tests of the Metacognitive Advantage Model: Counterfactuals, Confidence and Affect

André Mata
Universidade de Lisboa, CICPSI, Faculdade de Psicologia, Lisboa, Portugal

Abstract

This study tested whether people have an accurate sense of how good their reasoning is, as measured by their confidence in their responses, and how good they feel after they give those responses. First, incorrect responders were unjustifiably confident in their responses. However, correct responders were even more confident, and this confidence boost was found to come from their awareness of alternative solutions that are intuitive but incorrect. An affect measure revealed the same pattern: correct responders felt better, and incorrect responders felt worse, after they solved reasoning problems, but this was only the case when post-reasoning affect was measured after participants were instructed to think of alternative solutions. Implications are discussed for the possibility of implicit error monitoring, the role of counterfactual thinking in meta-reasoning, and the use of affective measures in meta-reasoning research.

Keywords: meta-reasoning, confidence, affect, counterfactual thinking, conflict detection

Introduction

Research on meta-reasoning has studied the confidence that people have in their reasoning. Several findings stand out in this literature: 1) incorrect responders are confident (i.e., overconfident; Kruger & Dunning, 1999; Mata, Ferreira, & Sherman, 2013; Pennycook, Ross, Koehler, & Fugelsang, 2017); but 2) correct responders are even more confident, presumably because they are aware of alternative misleading responses that they were able to override (Mata & Almeida, 2014; Mata et al., 2013); and 3) even though incorrect responders are overconfident, they might nevertheless have some awareness of their errors (De Neys, 2012).

This study offers new tests of these hypotheses. First, this study uses two kinds of metacognitive measures: confidence and affect. Whereas confidence has been widely investigated in meta-reasoning research (e.g., De Neys, Cromheeke, &

✉ André Mata, Faculdade de Psicologia, Alameda da Universidade, 1649-013 Lisboa, Portugal. E-mail: *andremata@psicologia.ulisboa.pt*

Osman, 2011; Mata et al., 2013; Thompson, Prowse Turner, & Pennycook, 2011), the use of affective measures is more recent (Klauer & Singmann, 2013; Morsanyi & Handley, 2012; Trippas, Handley, Verde, & Morsanyi, 2016). The purpose of using these measures is to test whether incorrect responders can somehow feel it when they commit reasoning errors. Whereas previous research asked reasoners to express how much they liked certain problems, the present study assessed reasoners' general affect. Specifically, the most popular measure of affect was used: the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988). Can people's momentary affect signal their metacognitive assessments? Is the impact of meta-reasoning so deep as to make people feel better or worse from one moment to the next, not just liking some propositions better than others, but actually feeling overall better or worse? And how do different metacognitive measures – confidence and affect – relate to each other?

The second contribution of this research is to offer a new test of the metacognitive advantage model (Mata et al., 2013) whereby the reason why incorrect responders are confident (i.e., overconfidence), but correct responders are even more confident than incorrect ones (i.e., a confidence boost), is that the former lack awareness of alternative solutions and think that theirs is the only possible response, whereas correct responders are aware of the alternative intuitive solution. If this is true, any procedure that heightens this awareness of alternatives should make correct responders feel better and more confident, whereas incorrect responders should feel worse and more doubtful. This study experimentally manipulated this awareness of alternatives by using a counterfactual instruction: After having completed a set of reasoning problems (the Cognitive Reflection Test; Frederick, 2005), participants were asked to indicate alternative solutions, different from the ones they produced. Participants rated their confidence in their solutions twice, before and after they thought of the alternative solutions. Correct responders were expected to feel more confident, and incorrect responders less so, and this was expected to depend on whether they thought of the relevant counterfactuals. Thus, the difference in confidence before vs. after the counterfactual thinking exercise should relate to the type of counterfactual solution generated: correct responders should be more confident to the extent that they thought of the alternative intuitive solution, whereas if incorrect responders realize that there are better alternative solutions, they should become less confident.

As for the affective measure, participants completed the PANAS both before and after they completed reasoning problems. Critically, the timing of the second PANAS measurement was manipulated across participants: some participants completed it after having solved the problems, but before the counterfactual exercise where they were explicitly instructed to think of alternative solutions; other participants completed it after they solved the problems and after the counterfactual exercise. This manipulation aims to test the implicit nature of error detection (De Neys, 2012). If error detection is efficient and spontaneous, then it should not be

necessary to explicitly prompt participants to consider alternative solutions; incorrect responders should know at some level that their responses are not the only possible ones, or even the most correct ones. If, on the other hand, incorrect responders only show signs of error detection upon being explicitly asked to reconsider their solutions, this would suggest that implicit error detection is not as spontaneous and efficient as previously hypothesized.

Finally, if affect taps onto confidence, that is, if it serves as a metacognitive signal of how sound one's reasoning is, then post-reasoning affect and confidence should be related.

Method

Participants

Ninety-one undergraduates from the University of Heidelberg participated and received course credit for their participation.

Procedure

First, participants responded to the PANAS, which measures positive and negative affect, each with 10 items. Participants were asked to indicate to what extent they felt several emotions (e.g., interested, proud, nervous, upset) at that moment, on a scale from 1 - *very slightly or not at all*, to 5 - *extremely*. The 20 items were presented in random order. The PANAS scores presented in the results section below were calculated by averaging the ratings for positive affect items and negative affect items separately, and then subtracting the latter from the former, such that positive scores indicate positive affect.

Then, participants answered versions of the 3 problems in the Cognitive Reflection Test (modified in their content, so as to minimize familiarity), after which they were asked "How confident are you that your answers to the previous problems are correct, on a scale from 1 - *not at all confident* to 9 - *very confident*?", as well as "How satisfied are you with the answers that you gave to the previous problems, on a scale from 1 - *not at all satisfied* to 9 - *very satisfied*?"

After that, participants received the following instructions: "Now you are going to see the same three problems again. But this time, instead of presenting your solutions, we want you to indicate alternative solutions that other people might have given to these problems. For each problem, write down a different solution than the one you gave." Participants were then showed the problems again, and for each of them they were asked: "What response might other people have given to this problem, other than the one you gave?"

Participants were asked to respond to the PANAS scale a second time, using the same instructions as before. For half the participants, this second PANAS measurement was done only after they had listed alternative solutions that other participants might have presented. For the other half, it was requested before the alternative generation task.

Finally, participants were asked to rate their confidence in their initial responses again: "Think back to the responses that you gave the first time the problems were presented to you. How confident are you that the responses that you gave in the beginning are correct, on a scale from 1 - *not at all confident* to 9 - *very confident*?"; "How satisfied are you with the answers that you gave in the beginning, on a scale from 1 - *not at all satisfied* to 9 - *very satisfied*?"

In short, in one condition, participants 1) completed the first PANAS measurement, 2) solved the problems, 3) expressed their confidence, 4) completed the second PANAS measurement, 5) generated alternative responses, and 6) expressed their confidence again. In the other condition, steps 4 and 5 were completed in the opposite order.

Results

Some participants failed to comply with the counterfactual thinking instructions: they either did not generate any alternative solution, or they simply repeated their solution. Rather than seeing this as a mere distraction or failure to comply, it might instead be seen as a consequence of the metacognitive difficulty that incorrect responders have in considering alternative responses. That is, the reason why they are so confident in their responses is that they cannot entertain the possibility that there are other (more valid) solutions. Indeed, performance (i.e., the number of correct responses; M = 1.74, SD = 1.13) correlates with how many alternative solutions participants were able to generate (M = 2.76, SD = 0.62), r = .21, p = .047. For the remainder of the analysis, only those participants who complied with the counterfactual instruction are considered (15 participants did not generate alternative solutions for one or more of the problems, and were therefore excluded).

Affect

Table 1 shows the mean PANAS scores across time, performance and order conditions.

Table 1

Mean Affect (and SD) Before and After Counterfactual Thinking by Performance and Order Condition

Number of Correct Problems	Order 1 (PANAS 2 Before Counterfactual Thinking)		Order 2 (PANAS 2 After Counterfactual Thinking)	
	PANAS 1	PANAS 2	PANAS 1	PANAS 2
0	2.03 (0.90)	1.98 (0.73)	1.35 (0.45)	0.98 (0.63)
1	1.50 (0.62)	0.99 (0.68)	1.00 (0.62)	0.39 (0.75)
2	1.41 (0.76)	1.50 (0.23)	1.54 (0.66)	1.54 (0.57)
3	1.24 (0.91)	1.21 (0.98)	1.28 (0.88)	1.75 (0.74)

Aggregating across conditions, performance was not related to pre-reasoning affect (PANAS 1), $r = -.09$, $p = .418$, but it correlated at a marginally significant level with post-reasoning affect (PANAS 2), $r = .20$, $p = .081$.

Critically, the difference in pre- vs. post-reasoning affect varied across order conditions. Performance did not correlate with the change in affect (a subtraction score measuring the difference in affect at time 1 vs. 2) when the second PANAS measurement occurred before the counterfactual instruction, $r = .20$, $p = .229$, but it did so when it was presented afterwards, $r = .50$, $p = .002$. Specifically, when post-reasoning affect was measured after the counterfactual exercise, affect decreased for incorrect responders and increased for correct responders.

These results suggest that counterfactual thinking influenced participants' affect. In order to further examine this possibility, the specific counterfactuals that participants generated were analyzed. In all 3 problems, incorrect responders generated several alternative responses, none of them consensual. For instance, for consistently incorrect responders (i.e., those with no correct response), the most frequent choices were never chosen by more than 25% of the responders, and only 2 responses (out of all the counterfactuals generated across problems and participants) corresponded to the correct solutions. Consistently correct responders, on the other hand, were much more consensual in indicating the intuitive solutions as alternative responses: for all three problems, these solutions were chosen in more than 50% of the cases.

A counterfactual score was created considering whether the alternative solutions generated by the participants were the intuitive solutions (scored as -1 for each problem) or the deliberative solutions (+1). First, performance correlates with this score, $r = -.68$, $p < .001$, suggesting that correct responders were aware of the intuitive alternative solution, and/or that incorrect responders were aware of the deliberative alternative solution. However, as stated above, this ability to think of the alternative response was much clearer for correct responders than for incorrect ones. The absolute counterfactual score (i.e., averaging across problems, and making negative and positive scores equivalent) is low for incorrect responders (e.g., for those who responded incorrectly to all problems, $M = 0.14$, $SD = 0.22$), whereas for

correct responders it is high (e.g., for those who responded correctly to all problems, $M = 0.63$, $SD = 0.19$).

Aggregating across conditions, the counterfactual score correlated with the difference in affect (PANAS 1 vs. PANAS 2), $r = -.31$, $p = .006$. But this was only the case when the second PANAS measurement was done after the counterfactual exercise, $r = -.42$, $p = .009$, not when it was done before, $r = -.17$, $p = .290$. Thus, counterfactual thinking changed affect in the condition where that was expected: when it came before the second affect measurement and could therefore influence it. Finally, a mediational analysis (using the bootstrapping procedure suggested by Preacher & Hayes, 2008) tested whether the effect of performance on the difference in pre- vs. post-reasoning affect is accounted by the counterfactual score. However, this analysis did not show a significant indirect effect, 95% CI = (-0.21, 0.16). This null result holds for both order conditions.

Confidence

Results on the two confidence items were aggregated (at time 1, $\alpha = .88$; at time 2, $\alpha = .94$). Table 2 shows the mean confidence levels across time and performance.

Table 2

Mean Confidence (and SD) Before and After Counterfactual Thinking by Performance

Number of Correct Problems	Confidence 1	Confidence 2
0	6.71 (1.99)	6.00 (2.32)
1	4.94 (2.66)	4.47 (2.06)
2	7.26 (1.71)	7.05 (1.67)
3	7.81 (1.06)	7.93 (1.23)

Confidence ratings before the counterfactual instruction show that incorrect responders were overconfident (for instance, those with 0 correct responses expressed confidence levels above the midpoint of the scale, $t(11) = 2.97$, $p = .013$). However, correct responders were even more so, such that the number of correct responses correlates with confidence, $r = .36$, $p = .001$. Confidence ratings after the counterfactual instruction show the same pattern, only even more clearly: $r = .49$, $p < .001$.

The difference in confidence after vs. before between the counterfactual instruction correlated with performance (i.e., the number of correct responses), although this was only marginally significant, $r = .21$, $p = .068$. There was a trend, such that correct responders grew more confident, whereas incorrect responders became less confident (see Table 2).

Confidence should change as a function of whether responders considered alternative responses. Indeed, the counterfactual score described above correlated

with the difference in confidence at time 1 vs. 2, r = -.33, p = .003. Actually, it already correlated with confidence even before the counterfactual exercise, r = -.28, p = .014, suggesting that these alternative solutions were already present in the minds of some responders, even before the explicit instruction to generate them (consistent with Mata & Almeida, 2014; Mata et al., 2013). However, this correlation was even larger after the counterfactual exercise, r = -.50, p < .001; comparing the correlation before vs. after the counterfactual exercise, z = 3.00, p = .003. Moreover, the effect of performance on the difference in pre- vs. post-counterfactual confidence is accounted by the counterfactual score: A mediational analysis using the bootstrapping procedure suggested by Preacher and Hayes (2008) revealed a significant indirect effect, 95% CI = (0.18, 1.30). This suggests that the degree to which participants grew more or less confident in their responses depended on the alternative solutions that they considered.

Finally, the relation between affect and confidence was examined. Pre-reasoning affect (PANAS 1) was not related with confidence either before or after the counterfactual exercise, respectively r = .10, p = .384, r = .04, p = .718. However, post-reasoning affect (PANAS 2) was related to both confidence measurements, respectively r = .40, p < .001, r = .39, p < .001. Moreover, this relation between confidence and post-reasoning affect was more pronounced when post-reasoning affect was measured after versus before the counterfactual exercise. Indeed, confidence (at time 2) predicted affect (at time 2), but only when affect was measured after the counterfactual exercise, r = .55, p < .001, not when it was measured before, r = .19, p = .258. In addition, confidence (at time 2) predicted the change in affect (from time 1 to time 2). This last result was observed both when post-reasoning affect was measured before counterfactual thinking, r = .44, p < .001, and after counterfactual thinking, r = .52, p < .001.

Discussion

This study investigated whether people have a good sense of how sound their reasoning is, as measured by a more traditional metacognitive measure (confidence), as well as an affective measure, which has only recently been used in this area of research. First, and replicating previous findings, incorrect responders were overconfident in their responses (Kruger & Dunning, 1999; Mata et al., 2013; Pennycook et al., 2017). Second, even though the confidence of incorrect responders was high, that of correct responders was even higher. Third, this confidence boost was found to come from correct responders' awareness of alternative solutions, which are intuitive and tempting, but misleading (Mata et al., 2013).

This pattern also showed in participants' affect: Correct responders felt better after solving the problems, whereas incorrect responders felt worse. The use of the PANAS measure shows that metacognitive appraisals can carry over to reasoners'

overall affect. Further supporting this idea that affect serves as a metacognitive cue, confidence and post-reasoning affect were found to be related.

Instruments such as the Need for Cognition Scale (Cacioppo & Petty, 1982; Epstein, Pacini, Denes-Raj, & Heier, 1996) ask people whether they consider themselves good at reasoning, and how much they enjoy reasoning (e.g., "I enjoy solving problems that require hard thinking."). However, those are self-report measures, which suffer from the limitations of introspection (Nisbett & Wilson, 1977). Indeed, such self-report measures of thinking disposition have been found to map poorly onto people's actual abilities to think properly about challenging reasoning problems. In particular, incorrect responders tend to overestimate their thinking abilities (Pennycook et al., 2017; see also Kruger & Dunning, 1999). However, the present results suggest that this enjoyment aspect of deliberative thinking can be real: Participants' affect really did change depending on their reasoning; or rather, their meta-reasoning.

Indeed, affect changed differently for correct responders (for the better) and incorrect responders (for the worse), but this was only the case after participants reflected on possible alternative responses. This study manipulated the timing of measurement of post-reasoning affect: For some participants, it was measured before reflecting on alternative responses, whereas for others it was measured after this counterfactual thinking exercise. Critically, affect only changed (in relation to the pre-reasoning baseline) when it was measured after participants thought of alternative responses. Specifically, it increased for correct responders, whereas it decreased for incorrect responders. This result serves a further demonstration that people's feelings about their responses are at least in part based on the knowledge (or lack thereof) that there are alternative responses – in the case of correct responders, worse responses that they avoided; in the case of incorrect responders, better responses that they missed (though the latter kind of counterfactual thinking was very rare).

The fact that affect only changed when participants were explicitly prompted to consider alternative responses has implications for the debate on implicit error detection and logical intuitions (De Neys, 2012). Indeed, although this study did not use the standard test of implicit conflict detection (comparing metacognitive measures for conflict vs. no-conflict problems; e.g., De Neys et al., 2011), it does offer some answers to the question of how spontaneous error detection is. On the one hand, whereas correct responders were aware of the alternative intuitive solutions, incorrect responders were not aware of the correct deliberative solutions (that is, after all, why they failed to solve the problems), even when they were instigated to think of alternative responses. On the other hand, confidence and affect increased for correct responders and decreased for incorrect ones, even when the latter did not generate the critical counterfactuals. Considering that incorrect responders were not able to generate the counterfactual correct responses, the effects of the counterfactual exercise on their affect and confidence must have been more implicit in nature: they

might have suspected that their responses were not accurate, though they did not know exactly why. Thus, with regard to whether incorrect responders are sensitive to their errors (De Neys, 2012), results are mixed. However, it seems clear at least that this epistemic process of doubting one's incorrect responses is not as spontaneous as previously hypothesized. Indeed, incorrect responders were confident at first; only when were prompted to think of alternative responses did they revise their confidence from time 1 to time 2. This pattern was even clearer for the affective measure: when post-reasoning affect was measured before the counterfactual exercise, their affect was high, at the same baseline level where it was in the beginning of the experiment. Only when post-reasoning affect was measured after the counterfactual exercise, did they lower their confidence. This speaks against the spontaneous nature of the error monitoring process.

At the same time, it is interesting to note that, while incorrect responders were the ones who generated the fewest relevant counterfactuals (i.e., the correct solutions), while correct responders generated more relevant counterfactuals (i.e., the intuitive but incorrect solutions), the counterfactual exercise seems to have changed the confidence of incorrect responders more than that of correct responders. Indeed, an additional analysis comparing the absolute change (i.e., making positive and negative differences comparable) in confidence from time 1 to time 2 shows that the best reasoners (those who responded correctly to all problems) only changed their confidence slightly ($M = 0.17$, $SD = 0.38$), whereas the worst performers (those who responded incorrectly to all problems) revised their confidence to a considerable degree ($M = 1.67$, $SD = 1.56$), $t(11.56) = 3.28$, $p = .007$. This might speak to the spontaneous nature of the counterfactual process of thinking of alternatives for deliberative responders (Mata et al., 2013). That is, the counterfactual exercise might have less of an effect on correct responders to the extent that they already engage in it spontaneously.

In conclusion, these results lend further support for the metacognitive advantage model (Mata & Almeida, 2014; Mata et al., 2013), whereby metacognitive judgments are informed by participants' awareness of alternative (better or worse) responses. This advantage was particularly evident for deliberative reasoners, who benefited the most from thinking of alternative responses.

References

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116-131.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*(1), 28-38.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS ONE, 6*(1), e15954.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology, 71*, 390-405.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25-42.

Klauer, K. C., & Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1265-1273.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Mata, A., & Almeida, T. (2014). Using metacognitive cues to infer others' thinking. *Judgment & Decision Making, 9*(4), 349-359.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology, 105*(3), 353-373.

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good - I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 596-616.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-249.

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning-Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24*(6), 1774-1784.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879-891.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*, 107-140.

Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1448-1457.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

Received: December 17, 2018