Reactivity to confidence ratings in older individuals performing the latin square task

Kit S. Double 1,2 · Damian P. Birney 2

Received: 22 August 2017 / Accepted: 26 November 2018 / Published online: 13 December 2018 © The Author(s) 2018

Abstract
Confidence ratings (CR) are often used to evaluate the metacognitive processes that occur during reasoning and problem solving. Typically CR are elicited with the assumption that they do not affect participants’ underlying cognitive processes. However, recent evidence suggests that eliciting CR can cause changes in cognitive performance. What is not yet clear, are the metacognitive pathways by which CR affect overall performance in older individuals. In order to better understand the mechanisms driving reactivity to CR, we evaluated the impact of eliciting CR in an older sample (N = 89) on two aspects of the metacognitive framework - monitoring and control. Participants first rated their prospective confidence before performing the Latin Square Task either with or without confidence ratings. Participants subsequently self-appraised their performance. We found evidence that eliciting CR leads to poorer metacognitive monitoring. In addition, we found that participants with high initial prospective self-confidence who perform CR adopt a more immediate performance-orientated control strategy, which improves short-term performance but has no effect on overall performance in a timed Latin Square Task.

Keywords Reactivity · Confidence ratings · Metacognition · Decision-making · Problem solving

Metacognition is an important component of reasoning and problem-solving (Ackerman and Thompson 2015). Metacognitive processes, such as confidence and error detection, are important components in both simple perceptual choices and more complex decision-making environments (Yeung and Summerfield 2012). Most often metacognitive processes are measured ‘online’, using confidence ratings elicited after each trial (Fleming and Lau 2014). In general, there is a strong correlation between confidence ratings and cognitive performance (Stankov 2013; Yeung and Summerfield 2014), however, there are also a number of observable systematic biases (Pulford and Colman 1997; Stankov and Crawford 1997). While confidence ratings have been widely adopted into both experimental and individual

Kit S. Double
kit.double@education.ox.ac.uk

1 Department of Education, University of Oxford, Oxford, UK
2 School of Psychology, University of Sydney, Sydney, Australia
differences methodologies, little attention has been given to whether eliciting confidence ratings in an online fashion affects performance – often referred to as reactivity (Fox and Charness 2010; Harris and Lahey 1982; Leow and Morgan-Short 2004).

Reactivity represents a challenge to the study of metacognition because many of the methodologies for measuring metacognition have proven to be reactive, at least under some conditions. For example, some have found think-aloud protocols (Fox and Charness 2010), judgements of learning (Mitchum et al. 2016; Soderstrom et al. 2015; Witherby and Tauber 2017), and confidence ratings (Birney et al. 2017; Double and Birney 2017a) to be reactive, whereas others have found the same measures not to be reactive (Fox et al. 2011; Kelemen and Weaver III 1997; Tauber and Rhodes 2012). Indeed there is little consensus over both the direction and magnitude of reactivity effects, and what factors determine these outcomes (For meta-analytic reviews, see Double et al. 2018; Fox et al. 2011). Recent work has suggested important factors that determine whether a rating is reactive. For example, related word-pairs appear to be reactive while unrelated word-pairs may not be reactive (Double et al. 2018; Janes et al. 2018). Similarly, think-aloud protocols seem to only be reactive when information beyond simple description of one’s thoughts are elicited (Fox et al. 2011).

The current study aims to extend recent findings by examining the effect of confidence ratings on both cognitive and metacognitive processes to better understand the full impact of eliciting confidence ratings from older individuals. Our investigations are situated in the Nelson and Narrens’ framework of metacognition (Nelson 1996; Nelson and Narens 1994) which we adopt to conceptualise how metacognitive consequences may give rise to the differential reactivity effects observed in the literature. Before describing this framework, we reflect on some recent investigations of reactivity.

Reactivity to confidence ratings

Recent evidence suggests that, like other metacognitive measures, eliciting confidence ratings from participants affects their cognitive performance. In a sample of mid-level business managers, Birney et al. (2017) found that eliciting confidence ratings resulted in poorer performance on Raven’s Advance Progressive Matrices (APM) compared to controls, but this effect was complexly moderated by a number of personality factors. In contrast, Double and Birney (2017a) found in a sample of university students, that APM performance was better for participants who performed the task with confidence ratings than without. Importantly, a second experiment in a sample of older adults, showed that the beneficial effect of confidence ratings was moderated by participants’ prior self-reported reasoning confidence. Even after baseline cognitive ability was controlled for, participants who were high in reasoning self-confidence performed better if asked to provide confidence ratings, whereas participants low in reasoning self-confidence performed worse. Double and Birney (2017a, in press) speculated that the self-evaluation involved in confidence ratings is affirming for participants who believe they are performing well, but threatening for participants who believe they are performing poorly.

Reactivity and age

Older individuals may be particularly susceptible to reactivity (Double and Birney 2017a, b; Fox and Charness 2010). For example, Fox and Charness (2010) found that older participants, but not younger participants, performed better on Raven’s Progressive Matrices if they were
asked to think-aloud as they solved the problems. Although it is not yet clear why older people may be particularly affected by confidence ratings, elsewhere it has been speculated that older individuals may benefit from such metacognitive prompts because it allows them to externalise their monitoring and regulatory processes or because such self-reporting prompts monitoring and control that would not otherwise occur (Fox and Charness 2010). This is particularly important because a range of changes in metacognitive abilities have been shown to occur as we age (McDaniel et al. 2008). While there are certainly age-related differences in cognitive control, currently it is unclear whether these should be interpreted as decline or alternatively whether older individuals utilise different strategies to achieve similar outcomes (Hertzog 2016). Similarly, the evidence for age-related changes in monitoring is equivocal, with some aspects of monitoring appearing to decline with age (e.g. overconfidence at test), while other aspects appear to be spared (e.g. monitoring of encoding) (Castel et al. 2015).

Nelson and Narrens metacognitive framework

Nelson and Narrens (Nelson 1996; Nelson and Narens 1994) describe a framework that provides grounds for hypothesizing about the metacognitive effects of eliciting confidence ratings. Their framework distinguishes between an object-level where cognitive activity occurs, and a meta-level that governs the object-level processes (Veenman et al. 2006). These two levels are linked together by an upward monitoring processes and downward control processes, see Fig. 1. These two metacognitive processes are communicative in that they inform each other in order to direct behaviour and cognition towards a goal (Nelson 1996). Importantly the communicated information is imperfect and may be influenced by other aspects of the self, such as motivation and affect (Efklides 2011). The object-level process is typically operationalized as cognitive performance on a criterion activity (Nelson 1996). In addition, we operationalize metacognitive monitoring by assessing retrospective appraisals of cognitive performance and metacognitive control by examining response time data. Below we consider specific hypotheses about how reactivity might play out in this metacognitive framework.

Monitoring effects

Item-level confidence ratings require participants to repeatedly self-assess and quantify aspects of their performance. Metacognitive monitoring is likely to be impacted by this repeated self-assessment. While it is intuitive to assume that more frequent self-assessment will lead to more accurate metacognitive monitoring, this is not necessarily the case. Koriat et al. (2008) distinguishes between information-based and experience-based processes in metacognition, with the former referring to metacognitive inferences based on participants’ beliefs about their own ability and preconceived notions about their competence, and the latter referring to cues based on subjective feelings that occur during a cognitive experience. Metacognitive monitoring should be most accurate when cues that are diagnostic of performance are utilised. Therefore, if confidence ratings direct attention to diagnostic cues, they may indeed result in more accurate metacognitive monitoring (assessed using retrospective appraisals). However, if confidence ratings are largely based on non-diagnostic cues then they may lead to less accurate retrospective appraisals. In particular, confidence ratings may activate pre-existing beliefs about one’s ability (information-based cues) which may accentuate pre-existing over/under
confidence. This information-based cue account would align well with the previous finding that self-confidence moderates reactivity in terms of the effect on cognitive performance (Double and Birney 2017a, in press).

Changes in metacognitive monitoring may act as a pathway for reactivity. Metacognitive monitoring deficits are associated with poorer performance in a large number of domains (see the Dunning-Kruger effect; Dunning 2011; Dunning et al. 2003). Self-regulation theories of self-focused attention (Carver and Scheier 2000, 2001) argue metacognitive reflection is necessary for an individual to regulate their cognitive resources relative to task demand and performance, and to gain insight into their own knowledge (Zimmerman 1998). According to these theories, self-focused attention leads to improved performance monitoring because it provides greater clarity into momentary performance and learning and allows for better regulation of behaviour and cognitive resources (Carver and Scheier 2001). Bannert and Reimann (2012) argue that metacognitive prompts, such as confidence ratings, activate monitoring and self-regulatory processes during learning, and because such performance monitoring processes are not carried out by all learners spontaneously, prompting will be beneficial to learning. However, Bannert and Reimann’s research stems from online learning environments where regulation is operationalized as decisions made by learners in highly interactive environments (e.g. navigating around a webpage). Such findings may not generalise to tasks where regulation is operationalized in simpler performance strategies (e.g. study time allocation).

**Control effects**

We hypothesize that the repeated requirement to self-assess one’s task performance impacts metacognitive control by causing participants to adopt a performance-focused control strategy. That is, by asking participants to rate the likelihood that they answered an item correctly,
confidence ratings direct attention to performance, which is likely to prompt participants to consciously or unconsciously regulate cognitive behaviour within the task to prioritise immediate performance rather than other outcomes (e.g. mastering the task). This may be particularly effective at improving short-term performance in older individuals, because they often use metacognitive strategies that are focused on content mastery rather than immediate performance (Justice and Dornan 2001) and age is negatively correlated with performance orientation (Button et al. 1996).

Consistent with this hypothesis, Mitchum et al. (2016) observed that when asked to perform judgements of learning participants would adopt a more conservative study strategy. They found when participants were asked to study related and unrelated word-pairs, those that did so while providing judgements of learning, spent longer studying the easier related word-pairs and less time studying the difficult unrelated word-pairs, ultimately impairing their performance on unrelated word-pairs at recall (related word-pair performance was unaffected by performing judgements of learning). In support of this notion, Mitchum et al. (2016) found that reactivity was only observed if related and unrelated word-pairs were presented in the same study list, unrelated word-pairs presented alone did not display reactivity to judgments of learning. They argue that when easy and difficult items were both to be recalled, participants modified their study decisions in order to adopt a more performance-orientated mindset rather than trying to master the difficult items.

Timed procedure

The current study utilised a timed reasoning task to examine the impact of confidence ratings on both object-level cognitive performance and the metacognitive control and monitoring processes. A timed task was utilised because the decision making in a timed task relies on metacognitive control to a greater extent than untimed tasks (for a comprehensive theory, see Ackerman 2014). Within a timed procedure, a participant must balance persisting with the current item against progressing to the next within the set time limit. Thus there are two obvious strategies that exist as a trade-off, a short-term strategy where one spends more time on the current item but may get through fewer items, or a more long-term strategy where one spends less time on each item but answers more.

Research questions and hypotheses

In the current study, we examine two related questions – how does providing confidence ratings influence participants performance and metacognitive processes. We investigate whether providing confidence rating during a timed problem-solving task affects performance monitoring, and whether participants react to the confidence ratings by making more performance-orientated decisions during the task.

Based on Double and Birney (2017a, b) previous findings we hypothesize that the effect of confidence ratings will be determined by participants’ prospective self-confidence, such that the performance of high self-confidence participants will be improved, while the performance of low self-confidence participants will be impaired. This was done with the expectation that if confidence ratings prompted high self-confidence participants to adopt a performance-orientated strategy, they would be more concerned with immediate performance and would therefore have superior mean performance but not necessarily better overall performance. Secondly, we hypothesize that performing confidence ratings will influence the performance monitoring of participants by directing their attention to different cues than they would
naturally attend to, although it is not self-evident whether this will lead to improvements or impairments in performance monitoring based on the extant literature.

Method

Participants

A power analysis based on a moderate incremental prediction of a single interaction term ($\Delta R^2 = .08$), suggested an approximate sample of 93. A community sample of 89 participants (82.02% female) was recruited using an advertisement placed in a newsletter of the Australian Broadcasting Corporation as part of a research partnership with the University of Sydney ($M_{age} = 64.18$, SD = 9.07; range = 27–84). Participants performed a reasoning task (Latin square task) either with or without confidence ratings. 46 participants were randomly assigned to the control group (No-CR) and 43 participants were assigned to the confidence ratings group (CR). Participants received no remuneration for participating in the study. After giving informed consent, participants completed the study online using their own computer and administered using Inquisit (Inquisit 2016). All study materials were programmed to present in a standardised fashion.

Materials and procedure

Latin square task

The Latin Square Task (LST; Birney et al. 2006) involves presenting participants with 36 reasoning problems that they must solve. As depicted in Fig. 2, each item consisted of a 4 x 4 matrix filled with a number of symbols as well as a missing symbol. Each row and column could only contain one instance of each symbol (the defining characteristic of a Latin squares and Sudoku type problems) and participants were tasked with determining the symbol that completed the matrix. The LST task was selected due the appropriateness of the task’s difficulty for a community sample, and the fact that the task has been shown to be a reliable predictor of fluid intelligence (Birney et al. 2012). After each item, participants in the CR group had to rate their confidence on a 6-point scale ranging from (0% to 100%) that asked “how confident are you in your previous answer?”, whereas a control condition was presented with a blank screen for 3000 ms in order to control for any effect of a delay. Participants were given 15 min to complete the task with a timer being displayed on screen at all times. The timer was paused when participants were performing confidence ratings. Response times were calculated as the time from stimulus onset to response. Items were presented in randomised blocked design of cognitive complexity (i.e. participants performed an easy item, then medium difficulty item, then difficult item and then repeated the sequence).

Prospective self-confidence/retrospective appraisal

All participants provided a prospective estimate of their performance and a retrospective appraisal of their performance. The prospective estimate was obtained after being presented

---

1 The high percentage of females in the current study is largely representative of typical volunteer samples of older individuals (Dickinson et al. 2012).
with two example questions and answers. Participants were asked to predict their performance on the test as a percentage; this was used as a measure of participants’ initial confidence, referred to henceforth as prospective confidence. The retrospective appraisal was performed immediately after completion of the task and required participants to estimate their overall performance on the task as a percentage.

Scoring

Two measures of performance on the LST were of interest: (1) the overall number of items correct in the time limit (15 min) calculated as a percentage (items correct/total items*100), which we refer to as Overall Accuracy and (2) Mean Accuracy, the number of items correctly answered of those actually attempted calculated as a percentage (items correct/items attempted*100).

Calibration

To assess the calibration between actual performance and participants’ retrospective appraisals we calculated a miscalibration score for each participant using the following formula described in Schraw (2009):

\[ \text{miscalibration}_i = (a_i - p_i)^2 \]

where \( p \) is equal to the actual performance of the \( i \)th participant and \( a \) is equal to the appraisal provided by the \( i \)th participant. Both \( p \) and \( a \) were calculated as a percentage.

Results

The data was analysed using R version 3.2.3 (R Core Team 2017) and plots were produced using ‘ggplot2’ (Wickham 2009). Table 1 presents descriptive statistics of study variables as well as bivariate correlations. Each criterion variable was analysed using the same linear regression model with experimental group, prospective confidence and their interaction as the predictor variables. Prospective confidence was mean centred and experimental group was
dummy coded. Homoscedasticity and the normality of residuals were checked for each regression by plotting standardized residuals vs. fitted values and a Q-Q plot respectively. No substantial violations were noted for any of the regression models.

Performance

Overall Accuracy across groups was 73.41% (SD = 18.12). A regression model ($R^2 = .17$, $F(3, 85) = 5.97$, $p < .001$) indicated that there was no significant difference between experimental groups in terms of Overall Accuracy, $b = -2.33$, $t = -0.65$, $p = .515$. Prospective confidence was a significant positive predictor of overall performance, $b = 38.95$, $t = 2.90$, $p = .005$. Furthermore, the group × prospective confidence interaction was not significant, $b = 1.46$, $t = .08$, $p = .940$.

Average Mean Accuracy across groups was 81.61% (SD = 13.20). The regression model ($R^2 = .35$, $F(3, 85) = 15.56$, $p < .001$) indicated that there was no main effect of experimental group on Mean Accuracy, $b = -2.10$, $t = -0.09$, $p = .362$. Prospective confidence was a significant positive predictor of Mean Accuracy, $b = 24.13$, $t = 2.79$, $p = .007$. Notably, the group × prospective confidence interaction was also significant, $b = 29.82$, $t = 2.41$, $p = .018$. As shown in Fig. 3, participants high in prospective confidence performed better when confidence ratings were elicited, whereas participants low in prospective confidence performed worse when asked to provide confidence ratings. As a follow-up analysis we probed the moderation using the Johnson-Neyman technique, which provides a region of significance in which group differences fall outside of a 95% confidence band (Bauer and Curran 2005). The Johnson-Neyman technique indicated that the group difference between the CR and No-CR groups was significant when prospective confidence was less than .11 standard deviations below the mean (prospective confidence ≤ 62.09%), but not significant with higher values of prospective confidence, see Fig. 4. The results indicated that there was a significant group difference between the CR and No-CR group for participants with prospective confidence below 62.09, such that participants in the CR group performed worse than controls within this region. These findings largely replicate the results of Double and Birney (2017a) with respect to the interaction between confidence and reactivity.

To explain the differential effect of confidence ratings on Overall vs. Mean Accuracy we analysed response time difference. Using the same model as above, a regression using mean response time as the criterion variable was utilised ($R^2 = .06$, $F(3, 85) = 5.97$, $p < .001$).
There was no significant main effect of group on response time, \( b = -0.03, t = -0.013, p = .999 \). Prospective confidence was a significant negative predictor of response time, \( b = -19.20, t = -2.00, p = .049 \), and the group × perspective confidence interaction was significant, \( b = 32.44, t = 2.36, p = .021 \). As shown in Fig. 5, high confidence participants spent longer on each individual item if asked to provide confidence ratings, whereas low confidence participants answered more quickly if asked to provide confidence ratings (compared to No-CR controls). Again, the Johnson-Neyman technique was used to probe the moderation, see Fig. 6. The findings suggested that CR participants responded significantly slower than controls for values of prospective confidence above .30 standard deviations above the mean (prospective confidence \( >69.57\% \)), and quicker for values .29 standard deviations below the mean of prospective confidence (prospective confidence \( <58.71\% \)). This suggests that the effect of confidence ratings on participants’ use of time on the task depended on their prospective confidence.

**Performance monitoring**

To assess performance monitoring we used miscalibration (see Fig. 7 for calibration plots) as the criterion variable using the same model as above. The regression model \((R^2 = .09, F(3,85) = 2.74, p = .048)\) indicated that averaging across prospective confidence, the CR group \((M = 642.36, SD = 781.98)\) was significantly more miscalibrated than the No-CR group \((M = \ldots\).
351.53, SD = 452.00); \( b = 279.90, t = 2.10, p = .039 \). That is to say, the CR group made significantly less accurate retrospective appraisals of their performance. There was no main effect of prospective confidence, \( b = 87.88, t = .175, p = .862 \), and the group \( \times \) prospective confidence was not significant, \( b = -1041.42, t = -1.45, p = .152 \).

To explore this effect further, we examined the predictors of participants’ retrospective appraisals. Retrospective appraisal was entered as the criterion variable, while prospective confidence and Overall Accuracy performance were entered as predictor variables along with the interaction between each of these effects and group. This allowed us to model the extent to which prospective confidence and actual performance differentially predicted participants’ retrospective appraisals (i.e. how did participants weigh their prospective confidence and actual performance when making their appraisals).

The model indicated that neither group, \( b = -.53, t = -0.17, p = .868 \), nor prospective confidence was a significant predictor of participants’ retrospective appraisals, \( b = 25.67, t = 1.96, p = .054 \). Overall Accuracy was a significant positive predictor of participants’ retrospective appraisals, \( b = .47, t = 3.33, p = .001 \). Crucially, the group \( \times \) prospective confidence interaction effect was significant, \( b = 61.02, t = 3.27, p = .002 \). Simple slopes analysis indicated that prospective confidence significantly positively predicted retrospective appraisals only in the CR group, \( b = 86.69, t = 6.54, p < .001 \), but not the No-CR group, \( b = 25.67, t = 1.96, p = .054 \), see Fig. 8. This result suggests that participants in the CR group tended to formulate their retrospective appraisals based on their prospective confidence to a greater extent than the No-CR group.
Confidence ratings often provide important information about metacognitive processes. Our findings, however, add to the growing body of evidence that challenges the assumption that confidence ratings can be elicited as an unobtrusive method of assessing metacognition (Birney et al. 2017; Double and Birney 2017a). Furthermore, like Double and Birney (2017a, 2017b) we found that confidence may be important in determining how older participants react to performing confidence ratings, both in terms of whether or not their performance is facilitated, and the metacognitive control strategies they appear to adopt as a result of performing ratings. In addition to evaluating the effect of confidence ratings on cognitive performance, we also evaluated their impact on meta-level processes. In terms of metacognitive monitoring, the current results suggest that eliciting confidence ratings resulted in less accurate retrospective self-appraisals of performance and supports the notion that confidence ratings draw on information-based cues rather than experience-based cues. In terms of metacognitive control, our findings suggest that performing confidence ratings prompts participants to adopt a different control strategy than they ordinarily would, whereby immediate rather than long-term performance is prioritised.

Reactivity

The present results replicate earlier findings that confidence ratings are reactive and the direction of this effect depends on participants’ self-confidence (Birney et al. 2017; Double
**Fig. 6** Conditional effect of group on mean response time as a function of prospective confidence. The shaded region indicates the values of prospective confidence within which the effect of CR is significant using a 95% confidence interval. Positive conditional effect values indicate slower response time in the CR group, whereas negative values indicate quicker response time in the CR group.

**Fig. 7** Calibration plots for the CR group. Error bars represent ±1 standard error of the mean.
and Birney 2017a). Although reactivity has always been a concern for research using some methods of measuring metacognition such as think-aloud protocols, it is only recently that reactivity to seemingly unobtrusive measures of assessing metacognition has received significant research. For example, recent empirical work by Mitchum et al. (2016) and Soderstrom et al. (2015) has challenged the notion that judgements of learning can be elicited without affecting participants performance. In addition, reactivity to confidence ratings appears to depend on participants’ self-confidence – performance of participants low in confidence appears to be impaired when confidence ratings are elicited. It has been well established that confidence ratings are highly related to actual cognitive performance (Stankov 2000, 2013; Stankov et al. 2014). This correlation may, however, be somewhat inflated by the asymmetrical effect of eliciting confidence ratings, if high self-confidence participants do not react to confidence ratings, whereas low self-confidence participants react negatively then the effect of pre-existing confidence on performance will be exaggerated when confidence rating are elicited and so too will the correlation between confidence and performance. Although further research is needed to confirm this hypothesis, it is worth mentioning that the typical correlations between online (contemporaneous) confidence ratings and performance are far greater than the correlations between self-report, trait-like measures of confidence and performance (Stankov et al. 2012, 2014).

The present reactivity effects also suggest that participants appear to differ in the way they make decisions about progressing through a task when confidence ratings are elicited compared to when they are not. For confidence ratings to be reactive, they must presumably directly attention to information that participants would not otherwise attend to (Ericsson and Simon 1993), which leads them to modify their metacognitive control during the task. This suggests

---

**Fig. 8** Retrospective appraisal as a function of experimental group and prospective confidence
that either the judgements elicited by confidence ratings do not necessarily occur spontaneously when confidence ratings are not provided or that the metacognitive process is, at least in part, an implicit one. This may need to be further considered by theories of cognitive decision-making that argue that individuals use their subjective confidence to decide when to progress to the next item in a test (Ackerman 2014), which may not ordinarily happen in the same way as it does when confidence ratings are elicited.

**Metacognitive monitoring**

Confidence ratings require participants to evaluate their performance. It is therefore intuitive to expect that they would facilitate performance monitoring, which would in turn benefit performance. This belief is informed by the often-held assumption that increasing the frequency of self-monitoring is equivalent to increasing the quality of self-monitoring. However, we find no evidence that performing confidence ratings resulted in more accurate metacognitive monitoring, on the contrary participants who performed confidence ratings in the current study were significantly less calibrated in their retrospective appraisals of their performance. The results tend to indicate that participants in the CR group based their retrospective appraisals on their prospective self-confidence to a greater extent than the No-CR group. This suggests that confidence ratings may direct attention toward self-relevant beliefs and concepts such as prospective confidence, and away from monitoring performance on the task at hand. This finding is in keeping with the idea that confidence ratings rely on information-based cues (Dunning et al. 2003; Koriat et al. 2004). While further replication is needed, the fact that the retrospective appraisals of the CR group deviated less from their prospective confidence and were less aligned with how they performed, suggests that their prospective confidence was more prominent in their metacognitive evaluations as a result of performing the confidence ratings. The implication is that performing confidence ratings may direct attention away from experience-based cues (i.e., task-focus) and onto participants’ information-based cues, particularly existing (prospective) confidence related beliefs (i.e., self-focus). Of course, if one’s information-based cues are diagnostic of actual performance then confidence ratings may result in better metacognitive monitoring (e.g. in a deceptively difficult task). As such, eliciting confidence ratings may be a poor intervention for improving the often observed performance monitoring deficits in older individuals (e.g. Palmer et al. 2014; Soderstrom et al. 2012).

There are also implications for self-regulated learning theory and the role of metacognition in decision-making. Attention is central to many theories of cognitive control and particularly theories about the role of metacognition in learning (Carver and Scheier 2001; Efklides 2011; Nelson 1996). Implicit in these theories is the notion that self-monitoring provides an accurate awareness of our cognitive process. If instead metacognitive prompts such as confidence ratings, lead to a biased view of our internal states then this poses a significant challenge for traditional theories of metacognition, which presume that an efficient and accurate attentional system is used to coordinate cognitive behaviour. This finding additionally poses an interesting question for future research, if introspection does not lead to more accurate performance monitoring, then what person and task characteristics moderate the attention-accuracy relationship?

**Metacognitive control**

Our results suggest that the requirement to provide confidence ratings affected the control-strategy that older participants used. Rather than simply prompting participants to adopt a more performance orientated mindset, we propose that, in keeping with the metacognitive
monitoring findings, this was a direct result of confidence ratings directing attention to information-based cues, namely confidence related beliefs (e.g. “I’m really good at problem solving”). For high-confidence participants, when their self-confidence related beliefs were activated they persisted longer with reasoning problems, believing that they would eventually answer them correctly. On the other hand, when attention was directed to confidence-related beliefs of low-confidence participants, they were less likely to persist with each item, because they lacked confidence that they would eventually answer it accurately.

We thus propose a two-stage model of reactivity to confidence ratings. The first stage premises that participants attend to confidence related beliefs when they perform confidence ratings i.e. they rely on information-based cues. In the second stage, participants make strategic control decisions (e.g. speed-accuracy trade-off) within a task based on those activated beliefs – decisions that will eventually affect their performance on the task. As a result, performing confidence ratings is likely to strengthen and reinforce beliefs (which may or may not be accurate) and encourage individuals to act in accordance with these beliefs (Silvia and Duval 2001).

Limitations

Educators have often been encouraged to introduce metacognitive prompts into classroom environments in order to enhance learning (e.g. Aurah 2014; Bannert and Mengelkamp 2008; Bannert and Reimann 2012; Chen 2007; Garrison 1997). Using confidence ratings as an intervention may be problematic given that these prompts may disproportionately assist high-confidence individuals, and may in fact reinforce inaccurate confidence-related beliefs learners have about their ability. However, the extent to which these effects can be attributed to the specific wording of confidence ratings is not yet clear. Confidence ratings tend to explicitly refer to confidence, and in doing so may be particularly prone to reinforcing individuals’ existing beliefs about their abilities compared with other types of metacognitive prompts (see Double and Birney in press). It may be prudent to investigate use of prompts that require learners to reflect on their cognitive processes (e.g. rating how deeply you thought about a problem) and their cognitive performance without actually providing an evaluation of that performance.

It is also worth mentioning that the current study did not consider the effect of guessing on confidence ratings and appraisals. In the LST, participants have a 25% chance of guessing the correct answer and this may have factored into the manner in which participants used the confidence ratings. For example, in Fig. 7 it is clear that participants average accuracy when they were selecting 0% on the confidence rating was at approximately chance level (25%), and thus they could be considered calibrated if you regard the 0% anchor as guessing. It is unclear how the specific nature of the rating scale and individual differences in how it is used affect reactivity and retrospective appraisals.

Finally, it is not yet clear to what extent these findings generalise to the younger samples that are typically observed in psychological studies (although such samples are themselves far from representative). Further research is necessary to better understand the extent to which population characteristics (e.g. age, intelligence, education etc.) determine the extent to which reactivity effects and the proposed mechanism generalise to other populations. In addition, there a host of other variables that have been shown to affect the accuracy of confidence ratings e.g. task difficulty (Pulford and Colman 1997), awareness of ignorance (Stankov and Lee 2008), alternative answers (Jackson 2016), and gender (Lundeberg et al. 1994) that may also affect reactivity. The current findings represent only a single experiment and substantially more work is needed to establish the extent to which the effects described here are robust, replicable and generalise to other tasks and ratings.
Conclusion

The current results have provided insight into the effect of confidence ratings on both reasoning performance and the accuracy of participants’ metacognitive monitoring in a sample of older participants. The current results provide further evidence that confidence ratings are reactive, and the direction and magnitude of reactivity is determined by pre-existing self-confidence.

Funding This study was funded by the Australian Research Council (grant number DP140101147).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ackerman, R. A. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*(3), 1349–1368.

Ackerman, R. A., & Thompson, V. (2015). Meta-Reasoning: What can we learn from meta-memory. In A. Feeney & V. Thompson (Eds.), *Reasoning as Memory* (pp. 164–178). Hove, UK: Psychology Press.

Aurah, C. M. (2014). The influence of self-efficacy beliefs and metacognitive prompting on genetics problem solving ability among high school students in Kenya. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 74*(9-a(E)), No pagination specified.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning, 3*(1), 39–58.

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science, 40*(1), 193–211.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40*(3), 373–400.

Birney, D. P., Beckmann, J., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: Complexity and learning trajectories in Raven’s progressive matrices depend on self-regulatory processes and conative dispositions. *Intelligence, 61*, 63–77.

Birney, D. P., Bowman, D. B., Beckmann, J. F., & Seah, Y. Z. (2012). Assessment of processing capacity: Reasoning in latin square tasks in a population of managers. *European Journal of Psychological Assessment, 28*(3), 216–226.

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning the development of the Latin Square task. *Educational and Psychological Measurement, 66*(1), 146–171.

Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes, 67*(1), 26–48.

Carver, C. S., & Scheier, M. F. (2000). On the structure of behavioral self-regulation. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 41–84). San Diego, CA: Academic Press.
Carver, C. S., & Scheier, M. F. (2001). *On the self-regulation of behavior*. Cambridge, UK: Cambridge University Press.

Castel, A., Middlebrooks, C., & McGillivray, S. (2015). Monitoring memory in old age: Impaired, spared, and aware. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of Metamemory* (pp. 463–483). NY: Oxford University Press.

Chen, C.-H. K. (2007). Prompting students' knowledge integration and ill-structured problem solving in a web-based learning environment. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 67(10-A), 3709.

Dickinson, E. R., Adelson, J. L., & Owen, J. (2012). Gender balance, representativeness, and statistical power in sexuality research using undergraduate student samples. *Archives of Sexual Behavior*, 41(2), 325–327. https://doi.org/10.1007/s10508-011-9887-1.

Double, K. S., & Birney, D. P. (2017a). Are you sure about that? Eliciting confidence ratings may influence performance on Raven's progressive matrices. *Thinking & Reasoning*, 23(2), 190–206.

Double, K. S., & Birney, D. P. (2017b). The interplay between self-evaluation, goal orientation, and self-efficacy on performance and Learning Paper presented at the Proceedings of the 39th Annual Conference of the Cognitive Science Society, London, England.

Double, K. S., & Birney, D. P. (in press). Do confidence ratings prime confidence? *Psychonomic Bulletin & Review*.

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, 26(6), 741–750.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, 44, 247.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12(3), 83–87.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychology*, 46(1), 6–25.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, mass. MIT Press.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8(July).

Fox, M. C., & Charness, N. (2010). How to gain eleven IQ points in ten minutes: Thinking aloud improves Raven's matrices performance in older adults. *Aging, Neuropsychology, and Cognition*, 17(2), 191–204.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316–344.

Garrison, S. J. (1997). Influence of metacognitive prompting on learning within computer mediated problem sets. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 57(8-A), 3390.

Harris, F. C., & Lahey, B. B. (1982). Subject reactivity in direct observational assessment: A review and critical analysis. *Clinical Psychology Review*, 2(4), 523–538.

Hertzog, C. (2016). Aging and metacognitive control. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of Metamemory*. NY: Oxford University Press.

Inquisit. (2016). Inquisit 5 (Version 5.04). Retrieved from http://www.millisecond.com

Jackson, S. A. (2016). Greater response cardinality indirectly reduces confidence. *Journal of Cognitive Psychology*, 28(4), 496–504.

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review*, 25, 2356–2364. https://doi.org/10.3758/s13423-018-1463-4.

Justice, E. M., & Dorman, T. M. (2001). Metacognitive differences between traditional-age and nontraditional-age college students. *Adult Education Quarterly*, 51(3), 236–249. https://doi.org/10.1177/074171360105100305.

Kelemen, W. L., & Weaver, C. A., III. (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(6), 1394–1409.

Koriat, A., Bjork, R. A., Sheffler, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133(4), 643–656.

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. *A handbook of memory and metamemory*, 117–136.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, 26(01), 35–57.

Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86(1), 114–121.
McDaniel, M. A., Einstein, G. O., & Jacoby, L. L. (2008). New considerations in aging and memory: The glass may be half full. In F. I. M. Craik & T. Salthouse (Eds.), The handbook of aging and cognition (3 ed., pp. 255–310). Hove, England: Psychology Press.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200–219.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*(2), 102–116.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA, US: The MIT Press.

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition, 28*, 151–160. https://doi.org/10.1016/j.concog.2014.06.007.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences, 23*(1), 125–133.

R Core Team. (2017). R: A language and environment for statistical computing (version 3.2.1) [computer software]. Vienna, Austria R Foundation for Statistical Computing. Retrieved from www.R-project.org/

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45.

Silvia, P. J., & Duval, T. S. (2001). Objective self-awareness theory: Recent progress and enduring problems. *Personality and Social Psychology Review, 5*(3), 230–241.

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553.

Soderstrom, N. C., McCabe, D. P., & Rhodes, M. G. (2012). Older adults predict more recollective experiences than younger adults. *Psychology and Aging, 27*(4), 1082–1088.

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence, 28*(2), 121–143.

Stankov, L. (2013). Noncognitive predictors of intelligence and academic achievement: An important role of confidence. *Personality and Individual Differences, 55*(7), 727–732.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*(2), 93–109.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*(4), 961–976.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747–758.

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology, 34*(1), 9–28.

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *The Quarterly Journal of Experimental Psychology, 65*(7), 1376–1396.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Witwerby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition, The Influence of Judgments of Learning on Long-Term Learning and Short-Term Performance*.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Phil. Trans. R. Soc. B, 367*(1594), 1310–1321.

Yeung, N., & Summerfield, C. (2014). Shared mechanisms for confidence judgements and error detection in human decision making. In *The cognitive neuroscience of metacognition* (pp. 147–167). New York, NY: Springer-Verlag Publishing; US.

Zimmerman, B. J. (1998). Academic studing and the development of personal skill: A self-regulatory perspective. *Educational Psychologist, 33*(2–3), 73–86.