A meta-analysis and systematic review of reactivity to judgements of learning

Kit S. Double, Damian P. Birney & Sarah A. Walker

To cite this article: Kit S. Double, Damian P. Birney & Sarah A. Walker (2018) A meta-analysis and systematic review of reactivity to judgements of learning, Memory, 26:6, 741-750, DOI: 10.1080/09658211.2017.1404111

To link to this article: https://doi.org/10.1080/09658211.2017.1404111

Published online: 21 Nov 2017.

Article views: 391

View Crossmark data

Citing articles: 1 View citing articles
A meta-analysis and systematic review of reactivity to judgements of learning

Kit S. Double, Damian P. Birney and Sarah A. Walker

School of Psychology, University of Sydney, Sydney, Australia

ABSTRACT

Judgements of learning (JoL) are often used in memory research as a means for assessing an individual's metacognitive beliefs about their learning. JoL have been shown to reliably predict performance as well as learning behaviours and decisions. Participants may, however, modify their behaviour in response to performing JoL. There has, however, been little consensus as to the reliability and direction of the effect. We report on a meta-analyses that assesses the evidence that memory performance is reactive to JoL. The results indicate that overall providing JoL does not have a significant effect on memory performance ($g = 0.054$, 95% CI $-0.027$ to $0.135$). However, sub-groups analysis showed that this effect depends on the nature of the stimuli to be recalled, with moderate positive reactivity observed for related word pairs ($g = 0.323$, 95% CI $0.083$ to $0.563$) and word lists ($g = 0.384$, 95% CI $0.146$ to $0.622$) but no reactivity when pairs were unrelated or a mixture of related and unrelated pairs. These results indicate that researchers should be aware that eliciting JoL may well influence participants' underlying encoding processes, especially when using related word pairs or word lists.

Judgements of learning (JoL) are metacognitive ratings whereby participants indicate the likelihood that they will recall learnt material on a future test (Rhodes, 2016). JoL are typically elicited after each item within a series of items that are to be remembered. For example, when participants are memorising a series of word pairs, JoL are often elicited after the presentation of each word pair. This methodology may make JoL susceptible to what is commonly referred to as reactivity (Rhodes & Tauber, 2011). Reactivity occurs when an individual either intentionally or unintentionally alters their performance or behaviour in response to being measured (Harris & Lahey, 1982). Reactivity can occur when participants are observed by an experimenter (e.g., McCarney et al., 2007), or when a participant self-reports on their own behaviour (e.g., Fox & Charness, 2010).

Due to the potential for reactivity, researchers have criticised the use of self-report measures collected concurrently with behaviour (Leow & Morgan-Short, 2004). So called “online” self-report measures, however, remain commonplace in cognitive psychology (Fox, Ericsson, & Best, 2011). Online self-report measures have been widely used in studies of memory (Ackerman, 2014; Koriat, 2000, 2012; Koriat & Ma’ayan, 2005), reasoning (Pallier et al., 2002; Stankov, 2000; Stankov & Lee, 2008), and decision-making (Balakrishnan & Ratcliff, 1996; Jackson & Kleitman, 2014). Despite widespread use of online self-report measures, empirical studies directly examining reactivity are relatively rare (Fox et al., 2011). Furthermore, reactivity may vary depending on a number of factors including participant characteristics, methodological differences, and features of the self-report measure itself (Ericsson & Simon, 1993). The current systematic review and meta-analysis focuses on the effect of online self-report measures on cognitive performance, specifically reactivity to JoL on memory performance.

A number of researchers have argued that the additional data gathered from online self-reports measures justifies the risk of reactivity (Ericsson & Simon, 1993). Online self-report measures have typically been used to provide insight into an individual’s metacognitive processing (Koriat & Helstrup, 2007). The accuracy of these self-report measures can often provide a behavioural measure of the effectiveness of an individual’s metacognitive monitoring (Thompson, 1999; Yan, 2015). For example, a large body of research has demonstrated that JoL made after a delay are more accurate than immediate JoL (e.g., Narens, Nelson, & Scheck, 2008; Nelson & Dunlosky, 1991), which may help guide educational practice (Dunning, Heath, & Suls, 2004). Furthermore, online self-report measures have been shown to predict performance, not only on the task at hand, but also on real-world outcomes such as academic achievement (Stankov, 2013).

In other areas, reactivity to self-report measures has been used as the basis off an intervention. For instance, in a series of experiments, Kazdin (1974) demonstrated that the direction of reactivity depends on the valence of the behaviour. Participants will tend to reduce negative

CONTACT Kit S. Double kit.double@sydney.edu.au School of Psychology, University of Sydney, A18, NSW 2006, Australia

© 2017 Informa UK Limited, trading as Taylor & Francis Group
behaviours, while at the same time increasing positive behaviours. In contrast, eliciting self-report measures does not appear to affect neutral behaviours. As such, reactive self-monitoring has been used as the basis for reducing a wide variety of negative behaviours including compulsive nail biting (Adesso, Vargas, & Siddall, 1979), smoking (McCarthy, Minami, Yeh, & Bold, 2015), and over-eating (Latner & Wilson, 2002). Additionally, self-monitoring has been used to increase positive behaviours such as staying on-task (Carr & Punzo, 1993) and exercise maintenance (Izawa et al., 2005).

JoL are one of the most frequently used concurrent self-report measures. JoL have been used to gain a range of insights into metacognitive evaluations and the study decisions individuals make when learning material. For example, JoL are often used to evaluate how various experimental manipulations can affect a participant’s subjective confidence in their learning (e.g., Susser, Jin, & Mulligan, 2016; Zechmeister & Shaughnessy, 1980). Traditionally, reactivity is rarely assessed and it is particularly rare that evaluating reactivity was the primary aim of a study (Fox et al., 2011). In a previous review of JoL reactivity, Rhodes and Tauber (2011) compared immediate JoL and delayed JoL in terms of the effect they have on memory performance. Their review found that, compared to immediate JoL, delayed JoL have a small positive effect on memory performance. The present systematic review and meta-analysis aims to extend on this work by summarising the findings of studies that compare immediate JoL to no judgement control conditions in order to assess whether immediate JoL, like delayed JoL, affect memory performance relative to “uncontaminated” controls.

Reactivity occurs when a self-report measure elicits information that would not otherwise be consciously attended to (Ericsson & Simon, 1993). Reactivity effects are problematic from a measurement perspective because they suggest that a self-report measure is unintentionally directing attention to information related to the judgement. This may over-inflate the importance of such information for an individual and modify their judgement and/or their learning. For example, if a participant is asked to provide a JoL, they may be more likely than otherwise to base study time decisions on whatever information is utilised when making the JoL. As a result, eliciting JoL may over-emphasise the importance of judgement relevant information and inflate correlations between such judgements and behavioural measures such as study time and performance.

Assessing reactivity effects is important not only for determining if a self-report measure is unintentionally affecting the performance of a participant but can also indicate whether such judgements occur without prompting. If providing JoL affects the performance of participants then this suggests that such metacognitive judgements are not necessarily made spontaneously in the same way as they are when JoL are elicited. This would imply that under conditions where JoL are not elicited, the equivalent metacognitive judgements are not made, or at least not made in the same way. Reactivity research provides some insight into basic learning processes, and could be utilised to better understand the contexts where such judgements occur spontaneously (and where they do not). Together, they facilitate a better understanding of the role of metacognition in various settings and the conditions that give rise to metacognitive monitoring.

Although there are relatively few studies looking at reactivity to JoL, there are a number of reasons why a meta-analytic approach is appropriate. Firstly, as reactivity is rarely the primary aim of a study, such effects are often under-emphasised when reporting results, and thus it is important for these findings to be more closely evaluated. Secondly, reactivity effects are not necessarily large and may well not reach significance in individual studies with small sample sizes (Fox et al., 2011). A meta-analytic approach may therefore observe trends that may have otherwise been overlooked. Finally, specific characteristics of the stimuli used within an experiment (e.g., the relatedness of the word pairs used) may affect the magnitude and direction of reactivity effects (Mitchum, Kelley, & Fox, 2016), therefore it is preferable to average across a variety of stimuli to assesses overarching trends in the data, as well as utilise meta-analytic procedures to determine if specific stimuli characteristics determine reactivity effects in a systematic way.

This review is timely given that a number of recent studies have started to investigate reactivity to JoL directly. Mitchum et al. (2016) reported a series of five experiments that examined the contextual situations under which reactivity to JoL occurs. They found that reactivity in a word-pair learning tasks depended on the relatedness of the word pair and the presence of cues that indicate the relative difficulty of a word pair (i.e., whether unrelated and related pairs were presented as part of the same list). Specifically, recall of unrelated word pairs was disrupted by the provision of JoL, whereas recall of related pairs was not reactive to JoL. Alternatively and in apparent contradiction, using a similar paradigm Soderstrom, Clark, Halamish, and Bjork (2015) found that recall to related pairs was facilitated by JoL whereas unrelated pairs were not reactive. Such studies indicate that pair relatedness may be important at determining reactivity effects, a hypothesis which we will test across studies in the present meta-analysis.

In addition to indicating the possible moderating role of word pair relatedness, the contradictory nature of these findings is representative of the mixed findings regarding reactivity to JoL in the extant literature. Overall, a number of authors have found positive reactivity (Dougherty, Scheck, Nelson, & Narens, 2005; Soderstrom et al., 2015; Zechmeister & Shaughnessy, 1980), whereas other have found no reactivity (Benjamin, Bjork, & Schwartz, 1998; Kelemen & Weaver, 1997; Tauber & Rhodes, 2012), and still others have found negative reactivity (Mitchum et al., 2016). Hence, a meta-analysis will be useful to summarise this contradictory literature.
Method

Study eligibility

Study types
In order to be eligible for inclusion, a study had to be a published peer-reviewed article that compared performance on a cognitive task performed with a JoL vs. with no judgement control condition. Both between-subject designs and within-subject designs were eligible, provided that participants were randomly allocated to condition/order. Only English language papers were considered.

Participant types
Only studies using adult neuro-typical and non-clinical populations were eligible for inclusion.

Outcome types
Studies needed to assess performance on a recognised cognitive task, including memory tasks, reasoning tasks, intelligence tasks, and decision-making tasks.

Judgement types
Studies were eligible if they assessed performance with vs. without JoL, regardless of the specifics of the scale or the wording of the question.

Study selection
Given that reactivity is rarely assessed directly, a deliberately broad set of search terms was selected. A search of PsychINFO and Science Direct was conducted in August 2016. Search terms for JoL ("judgements of learning" OR "judgments of Learning" OR "JOL" OR "delayed judgments of learning" OR "delayed judgements of learning" OR "confidence ratings" OR "confidence judgements” OR "metacognitive judgements” OR "metacognitive ratings”) were combined with cognitive ability search terms ("reactivity" OR "cognitive Performance" OR “cognitive ability” OR "Memory” OR "problem solving” OR "decision making" OR "Intelligence”) (see Note 1). Reference lists were also scanned for eligible studies.

Assessing eligibility
Two reviewers independently screened the titles and abstracts of the search results for eligibility. When eligibility was unclear full text versions of the articles were assessed. In the instance of disagreement between the two reviewers a third reviewer was consulted to assess eligibility and resolve any disagreements. If a study was eligible but did not provide sufficient data regarding the comparison, then the original authors were contacted requesting the necessary data.

Meta-analytic approach
A fixed effects meta-analysis was carried out as they are considered more conservative and less susceptible to bias (Greenland, 1994; Poole & Greenland, 1999). The primary outcome was standardised mean difference (SMD) between JoL and No-JoL conditions, calculated using Hedge’s g. A positive SMD indicates comparatively better cognitive performance in the JoL group (i.e., positive reactivity). Between-subjects and within-subjects designs were combined by first converting to a common effect size metric (Morris & DeShon, 2002). Stimuli type was entered as a sub-group variable with four categories (related, unrelated, mixed, word list) found. All analyses were performed using comprehensive meta-analysis (CMA). Figures were created using the “metafor” package (Viechtbauer, 2010) in R version 3.2.4 (R Core Team, 2015).

Results

Description of studies
The study selection processes and the associated outcomes are presented in Figure 1. Eight studies comprising 17 experiments were included in the final meta-analysis. Four experiments were within-subjects designs and the remaining 13 were between-subjects designs. Two experiments manipulated pair relatedness within subject so the effects were entered separately (increasing the total number of effect sizes to 19). One author provided additional data upon request. A summary of the included studies is presented in Table 1.

Included studies consisted of 1321 participants (mean JoL group size = 45.54 (SD = 25.86); mean control group size = 44.77 (SD = 25.59); mean within-subjects study size = 36.75 (SD = 13.60)). For studies in which multiple JoL effects were reported (e.g., immediate and delayed JoL) the immediate JoL vs. control condition effect was used due to the fact that delayed JoL have previously been shown to have a greater effect on memory than immediate JoL and thus examining the immediate JoL condition should provide a minimal baseline for possible reactivity effects (Rhodes & Tauber, 2011). All included experiments involved either word-pair learning (n = 14) or word-list learning (where participants were presented a list of single words to learn; n = 3).

Study design
In the 14 word-pair learning experiments, participants were presented with a series of word pairs to learn after which they completed a recall test wherein they were presented with one member of the pair and had to recall the other. In four of the experiments, the JoL group was allowed to restudy the word after making the JoL. In 4 of the word-pair learning experiments participants were able to decide how long to study the word pair, whereas 10 of
the word-pair experiments had a fixed study duration with a mean study duration of 6.91 s (SD = 2.02 s). In one word-pair learning experiment presentation time was experimentally manipulated (3 vs. 12 s). The three word-list experiments presented participants with a list of words to learn after which participants performed either a recognition test \((n = 1)\) where they had to discriminate between learnt words and distractors or a free-recall test \((n = 2)\) where they had to recall as many words from the list as possible.

**Stimuli**

The mean number of word pairs to be recalled was 70.35 pairs \((SD = 75.54)\). The mean number of words in the word-list experiments was 106.8 words \((SD = 142.52)\). Three experiments included only related pairs, six included only unrelated pairs, and five included both related and unrelated pairs. In addition, Mitchum et al. (2016; Experiment 1) manipulated the direction of the associative relatedness (forward and backward) and Mitchum et al. (2016; Experiment 2) manipulated the emotional valence of the words, although neither manipulation was found to affect reactivity. Furthermore, Tauber and Rhodes (2012) manipulated the abstractness of the words in their list, but found no effect of abstractness on reactivity. Mitchum et al. (2016; Experiments 1 and 2) and Soderstrom et al. (2015) performed within-subjects comparisons of reactivity using related and unrelated pairs presented as part of the same series. Mitchum et al. found recall of unrelated pairs was subject to negative reactivity, but recall of related pairs showed no reactivity, while Soderstrom et al. found related pairs exhibited positive reactivity whereas there was no reactivity in unrelated pairs. As reactivity was demonstrably different between related and unrelated pairs we opted not to average across the stimuli types, instead including the unrelated and related word pairs separately in the meta-analysis (see Note 2). However, as Soderstrom et al. (2015) only reported that \(p > .05\) for the unrelated pairs effect, only the related pairs effect could be included in the meta-analysis.

**Reactivity**

Three experiments found that performance was better under the JoL condition compared to the no-judgement condition. Nine experiments found no significant difference between the JoL group and controls. Five experiments found that the effect depended on cue relatedness (described above). Finally, no experiments found negative reactivity.

**Meta-analysis**

The meta-analyses indicated that overall there was no significant effect of JoL on memory performance \((g = 0.054, 95\% CI \approx -0.027 to 0.135, p = .190, see Figure 2)\). Heterogeneity between the included studies was assessed using a Chi-square test, which assesses whether observed differences in results are likely due to chance alone (Schwarzer, Carpenter, & Rücker, 2015). This Chi-square test indicated there was significant heterogeneity among the studies \((Q = 53.59 p < .001)\), as such a number of sub-groups analyses were performed to attempt to explain the heterogeneity in observed effect (see below). Funnel plots comparing the relationship between observed effects and standard error were inspected for asymmetry, where symmetry would indicate possible risk of publication bias in the included studies (Schwarzer et al., 2015). A regression test for funnel plot asymmetry (Viechtbauer, 2010) indicated that the meta-analyses showed significant signs of asymmetry and the Egger’s test, a test of symmetry, was not significant \((p = .124)\), suggesting that the risk of publication bias in the included studies is relatively low, see Figure 3.
### Table 1. A summary of study characteristics of all included studies.

<table>
<thead>
<tr>
<th>Author Year</th>
<th>Exp.</th>
<th>Word type</th>
<th>Stimuli</th>
<th>Other man.</th>
<th>No. wds</th>
<th>Study time</th>
<th>Set study time</th>
<th>Re-study</th>
<th>Test type</th>
<th>Test delay</th>
<th>Design</th>
<th>Reactivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitchum et al. 2016</td>
<td>1*</td>
<td>RWP/UWP</td>
<td>Word pair</td>
<td>Pair direction</td>
<td>60</td>
<td>Participant defined</td>
<td>No</td>
<td>Recall</td>
<td>Immediate</td>
<td>BS</td>
<td>Negative/No</td>
<td></td>
</tr>
<tr>
<td>Mitchum et al. 2016</td>
<td>2*</td>
<td>RWP/UWP</td>
<td>Word pair</td>
<td>Valence</td>
<td>44</td>
<td>Participant defined</td>
<td>No</td>
<td>Recall</td>
<td>Immediate</td>
<td>BS</td>
<td>Negative/No</td>
<td></td>
</tr>
<tr>
<td>Mitchum et al. 2016</td>
<td>4a</td>
<td>UWP</td>
<td>Word pair</td>
<td>48</td>
<td>Participant defined</td>
<td>No</td>
<td>Recall</td>
<td>Immediate</td>
<td>BS</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitchum et al. 2016</td>
<td>4b</td>
<td>UWP</td>
<td>Word pair</td>
<td>48</td>
<td>Participant defined</td>
<td>No</td>
<td>Recall</td>
<td>Immediate</td>
<td>BS</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitchum et al. 2016</td>
<td>5</td>
<td>RWP/UWP</td>
<td>Word pair</td>
<td>48</td>
<td>5 s</td>
<td>Fixed</td>
<td>No</td>
<td>Recall</td>
<td>Immediate</td>
<td>BS</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Soderstrom et al. 2015</td>
<td>1a</td>
<td>RWP</td>
<td>Word pair</td>
<td>60</td>
<td>8 s</td>
<td>Fixed</td>
<td>Yes</td>
<td>Recall</td>
<td>3 min</td>
<td>BS</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Soderstrom et al. 2015</td>
<td>1b</td>
<td>RWP</td>
<td>Word pair</td>
<td>60</td>
<td>8 s</td>
<td>Fixed</td>
<td>Yes</td>
<td>Recall</td>
<td>3 min</td>
<td>BS</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Soderstrom et al. 2015</td>
<td>2</td>
<td>RWP</td>
<td>Word pair</td>
<td>50</td>
<td>8 s</td>
<td>Fixed</td>
<td>Yes</td>
<td>Recall</td>
<td>3 min</td>
<td>BS</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Dougherty et al. 2005</td>
<td>2</td>
<td>UWP</td>
<td>Word pair</td>
<td>Duration/dummy</td>
<td>52</td>
<td>3/12 s</td>
<td>Fixed</td>
<td>Yes</td>
<td>Recall</td>
<td>Half</td>
<td>Immediate/half 2 days</td>
<td>BS</td>
</tr>
<tr>
<td>Kelemen and Weaver 1997</td>
<td>1</td>
<td>RWP/UWP</td>
<td>Word pair</td>
<td>Abstract vs. concrete</td>
<td>60</td>
<td>6s</td>
<td>Fixed</td>
<td>No</td>
<td>Recall</td>
<td>10-20 min</td>
<td>WS</td>
<td>Positive</td>
</tr>
<tr>
<td>Kelemen and Weaver 1997</td>
<td>3</td>
<td>RWP/UWP</td>
<td>Word pair</td>
<td>60</td>
<td>8 s</td>
<td>Fixed</td>
<td>No</td>
<td>Recall</td>
<td>10-20 min</td>
<td>WS</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Zechmeister and Shaughnessy 1980</td>
<td>1</td>
<td>UWL</td>
<td>Word list</td>
<td>24</td>
<td>6s</td>
<td>Fixed</td>
<td>No</td>
<td>Recall</td>
<td>1 min</td>
<td>WS</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Yang et al. 2015</td>
<td>1</td>
<td>UWL</td>
<td>Word list</td>
<td>360</td>
<td>2 s</td>
<td>Fixed</td>
<td>no</td>
<td>Recognition</td>
<td>24 h</td>
<td>WS</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Tauber, Dunlosky, and Rawson 2015</td>
<td>1</td>
<td>UWP</td>
<td>Word pair</td>
<td>Duration/dummy</td>
<td>60</td>
<td>8 s</td>
<td>Fixed</td>
<td>no</td>
<td>Recall</td>
<td>Half</td>
<td>Immediate/half 2 days</td>
<td>BS</td>
</tr>
<tr>
<td>Tauber et al. 2015</td>
<td>2</td>
<td>UWP</td>
<td>Word pair</td>
<td>66</td>
<td>8 s</td>
<td>Fixed</td>
<td>no</td>
<td>Recall</td>
<td>2 days</td>
<td>BS</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Tauber et al. 2015</td>
<td>3</td>
<td>UWP</td>
<td>Word pair</td>
<td>66</td>
<td>8 s</td>
<td>Fixed</td>
<td>no</td>
<td>Recall</td>
<td>2 days</td>
<td>BS</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

Notes: UWP = unrelated word pair; RWP = related word pair; UWL = unrelated word list; s = seconds; no. wds = number of words/pairs; other man. = other manipulations. Experiments marked with an asterisk were entered using separate effects for related and unrelated pairs. For the Soderstrom et al. (2015) paper only the related words pairs could be included in the meta-analysis because there was insufficient effect data reported for related word pairs.

---

**Sub-groups analysis**

Firstly, a sub-groups analysis examined whether there was a difference in between and within-subject designs. The sub-groups analysis indicated that there was no significant difference between the effect sizes as a function of between-subjects vs. within-subjects design ($p > .05$). A second sub-groups analysis compared effects as a function of stimuli type. Four groups (related, unrelated, mixed, word list) were formed. The sub-groups analysis indicated significant differences in reactivity as a function of stimuli type, see Figure 4. Moderate positive reactivity was observed for related word pairs ($g = 0.323, 95\% CI 0.083 to 0.563, p = .008$) and word lists ($g = 0.384, 95\% CI 0.146 to 0.622, p = .002$) but no reactivity was found when pairs were unrelated ($g = -0.014, 95\% CI -0.156 to 0.128, p = .128$) or in studies with a mix of related and unrelated pairs ($g = -0.049, 95\% CI -0.169 to 0.071, p = .424$).

**Discussion**

Researchers are often interested in the metacognitive judgements that an individual makes when learning new material. Although these metacognitive processes are likely to determine a number of important learning behaviours (e.g., the allocation of study time), the current meta-analytic results suggest that JoL, the primary way in which we currently assess such processes, may unintentionally influence the underlying behaviour in some circumstances. This finding has important implications for research practice and challenges the presumption that JoL can be used as an unobtrusive measure of metacognitive processes.

The current results imply that participants either intentionally or unintentionally modify their behaviour in response to JoL in some situations. This has a number of implications for researchers interested in measuring metacognitive processes during learning. Although the potential for reactivity may not be particularly problematic in some experimental contexts, it is worth noting that having participants perform JoL may improve their monitoring processes and decrease the likelihood that participants will engage in ineffective learning practices, which may be of primary interest to a researcher. Given such measures are not uncommon in the memory research literature, reactivity should be considered as a potential threat to the interpretation of findings. In order to avoid reactivity effects, researchers may wish to use alternative measures of metacognitive processes that are less reactive, such as think-aloud protocols (for a meta-analytic review, see Fox et al., 2011). Think-aloud protocols may be less reactive because they do not elicit extra information that an individual would not otherwise have attend to, instead participants are encouraged to verbally report their internal thoughts only as they are (for a review, see Ericsson & Simon, 1993).

The results of the sub-group analysis provide a clear indication that the type of stimuli drives the magnitude...
and direction of reactivity. While, related word pairs and word lists showed a moderate positive reactivity effect; non-significant effects were found in all other subgroups. This raises the question of why (positive) reactivity only occurs in related word pairs and word lists? A possible explanation comes from recent work examining reactivity to confidence ratings, where it has been shown that positive reactivity only occurs for high-confidence participants, whereas negative or no-reactivity is observed in low-confidence participants (Double & Birney, 2017). Participants generally have higher confidence in their ability to recall related word pairs.
compared to unrelated word pairs and it may be that performing a JoL draws participant’s attention to their subjective confidence, which is beneficial when confidence is high but detrimental when confidence is low. Double and Birney (2017) put forward the cognisant confidence hypothesis as a way of explaining the mechanisms underlying heterogeneity in reactivity effects. The cognisant confidence hypothesis proposes that, rather than providing a direct enhancement of metacognitive monitoring, self-report measures draw attention to activated self-related beliefs (e.g., confidence and goals), which may be beneficial to performance if such beliefs are affirming but in contrast may impair performance if those beliefs are negative or threatening. More research is needed to establish if confidence similarly moderates JoL reactivity, particularly as such findings may not generalise from confidence ratings as JoL differ from confidence ratings in that they are a prospective rather that retrospective judgements which may change the nature of the relationship between confidence and reactivity. Speaking more broadly, additional empirical research is needed to examine the mechanisms underlying the effect of pair relatedness on reactivity, particularly as the effect may be complex and depend on numerous factors such as participant goals and precise experimental characteristics (Mitchum et al., 2016). In particularly it is currently unclear why word-list experiments show reactivity effects, although this effect should be interpreted cautiously as only three word-list experiments were included in the meta-analysis. Nevertheless, the current findings provide a clear indication of when reactivity is likely to occur, which is both useful for future investigations of reactivity and for establishing when researchers should be most concerned about unintentional reactivity.

Although some researchers have suggested that JoL enhance performance by increasing the quality of monitoring processes, the specifics of such a cognitive effect (e.g., the precise cognitive mechanisms that improve when JoL are performed) are relatively unknown. JoL require a participant to retrospectively evaluate the effectiveness of their learning. At least in the studies reported here, this is done repeatedly during testing and thus a prospective evaluation is also likely to be primed. That is, the continued task demand to evaluate one’s own performance may lead to enhanced learning due to incidental improvements in the effectiveness of the learning processes or as a strategic response by the individual to perform better in the face of frequent self-assessment. Future research is needed to

\begin{figure}
\centering
\includegraphics[width=\textwidth]{funnel_plot.png}
\caption{A funnel plot of overall effects as a function of standard error.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{stimuli_effect_sizes.png}
\caption{Effect sizes as a function of stimuli type.}
\end{figure}
disentangle and specify the mechanisms by which reactivity occurs and the contextual factors that give rise to reactivity effects.

The results also have interesting practical implications because they suggest that there is sometimes a benefit to performing behaviours that enhance monitoring. As such, JoL may be a useful tool for enhancing learning particularly for low-performing students who do not naturally engage in performance monitoring effectively (for a review see Dunning, Johnson, Ehrlinger, & Kruger, 2003). Further experimental work is necessary to investigate the role of JoL in enhancing monitoring in those that display a monitoring deficit. Compared with previous metacognitive interventions, JoL are noteworthy in that they are reasonably domain general. While there are a host of prompts that have specifically been devised for use in a particular domain, for instance reading (Pressley & Afflerbach, 1995) and mathematics (Kramarski & Mevarech, 2003), JoL can be implemented into most learning and training environments with little modification across content domains, making them an easy tool to integrate into classroom and learning environments. For example, some researchers have used computerised programs to prompt and enhance the metacognitive processes with success (Azevedo, 2005).

It is also pertinent to note that the above effects may have important individual difference qualifiers. If JoL enhance the performance monitoring processes of participants, then the direction and magnitude of any reactivity effects will depend on the difference between the quality of an individual’s natural unprompted monitoring processes and their prompted monitoring process. Further, consideration is needed to determine the task and person characteristics that influence the direction and magnitude of reactivity.

Reactivity to online self-report measures is a general problem in psychology and it is worth noting that reactivity to other common self-report measures has received far less attention and may be equally problematic. For example, confidence ratings have often been incorporated into fluid intelligence tasks (Crawford & Stankov, 1996; Stankov, 2000; Stankov & Lee, 2008) with little consideration of whether judgements influence performance. Although the above results can only speak to reactivity to JoL, it seems pertinent that reactivity to a broad range of self-report measures receives greater attention in cognitive psychology.

It is worth acknowledging that the current meta-analysis may have been limited by the nature of reactivity research. Evaluating reactivity effects is rarely the primary aim of a study; often such comparisons are only carried out as a pre-cautionary step (Fox et al., 2011). As a result, while every effort to find and include relevant studies has been made, it is possible that relevant effects have been evaluated but not directly mentioned in paper abstracts and therefore overlooked in the current analysis.

The current meta-analysis has provided an indication that JoL are positively reactive when related word pairs are used. These findings suggest that researchers, when designing experimental paradigms that rely on online JoL, should consider the potential for reactivity. A further implication of these results is the potential to use JoL in some situations as a mechanism to enhance learning. The next challenge is to determine the mechanisms that influence the magnitude and direction of reactivity effects, in particular the mechanisms driving the link between pair relatedness and reactivity, as well as the extent to which the above findings generalise beyond JoL.

Notes

1. Originally the review was open to looking at reactivity to confidence ratings as well, because they are retrospective judgments and would have provided a useful comparison to JoL, hence the more expansive list of keywords and outcome types. Unfortunately, no such studies examining confidence ratings were found thus the focus of the review was subsequently narrowed.

2. Including two outcomes from the same set of participants in the meta-analysis is sometimes problematic due to the fact that the participants are not independent. This procedure was, however, considered preferable to averaging out the effects due the marked differences between related and unrelated pairs. In order to prevent inflation of the weighting of the two studies in question, half (rounded to the nearest person) of their respective samples was used when calculating weighted effect sizes.

3. For the meta-analysis, data was averaged across the 3- and 12-s conditions as recommended by Borenstein, Hedges, Higgins, and Rothstein (2009).

Disclosure statement

No potential conflict of interest was reported by the authors. The authors would like to thank Amit Lampit for helpful discussion regarding the analyses.

Funding

This work was supported by Australian Research Council [grant number DP140101147].

References

References marked with an asterisk indicate studies included in the meta-analysis.


*Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don’t. *Bulletin of the Psychonomic Society, 15*(1), 41–44.