MEMORY, 2017
https://doi.org/10.1080/09658211.2017.1406523

When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the "Displaced-JOL effect"

Young Bui[a], Mary A. Pyc[a,b] and Heather Bailey[a,c]

[a]Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA; [b]Dart NeuroScience, San Diego, CA, USA; [c]Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA

ABSTRACT

Judgments of learning (JOL) made after a delay more accurately predict subsequent recall than JOLs made immediately after learning. One explanation is that delayed JOLs involve retrieving information about the target item from secondary memory, whereas immediate JOLs involve retrieval from primary memory. One view of working memory claims that information in primary memory is displaced to secondary memory when attention is shifted to a secondary task. Thus, immediate JOLs might be as accurate as delayed JOLs if an intervening task displaces the target item from primary memory, requiring retrieval from secondary memory, prior to making the JOL. In four experiments, participants saw related word-pairs and made JOLs predicting later recall of the item. In Experiment 1, delayed JOLs were more accurate than JOLs made shortly after learning, regardless of whether a secondary task intervened between learning and JOL. In Experiments 2–4, the secondary task demands increased and JOLs made shortly after learning with an intervening task were just as accurate as delayed JOLs, and both were more accurate than immediate JOLs with no intervening task (Experiment 4). These results are consistent with a retrieval-based account of JOLs, and demonstrate that the "delayed-JOL effect" can be obtained without a long delay.

ARTICLE HISTORY
Received 27 January 2017
Accepted 6 November 2017

KEYWORDS
Judgments of learning; working memory; metacognition; primary memory; secondary memory

Metacognition refers to the ability to understand the state of one's knowledge (Dunlosky & Metcalfe, 2009; Flavell, 1979). Nelson and Narens (1990) proposed one of the earliest models for how we understand our own knowledge, which included two components of metacognition – monitoring and control – that are distinct, yet work in conjunction with one another to guide behaviour, or self-regulated learning (see also, Greene & Azevedo, 2007; Winne & Hadwin, 1998). The monitoring component reflects the degree to which an individual can accurately assess the state of one's memory, and presumably is based on various cues that provide information about how well an item has been learned (e.g., Koriat's cue-utilisation framework; Koriat, 1997). The control component represents metacognition on the object-level, and acts as the manifestation of the monitoring component. That is, it involves decisions about how to study, what to study, and when to study. Critically, these two components work in a bidirectional fashion to influence one another. For example, when a student studies for an exam, the monitoring component helps guide their metacognitive judgment of whether they understand the material well enough to pass the test. The student may take a practice exam, and after doing poorly, use the feedback from that exam to decide that they indeed do not understand the material well enough (monitoring component), and therefore need to study more (the control component; but see Koriat, Ma'ayan, & Nussinson, 2006 for a retroactive account in which a retrieval attempt precedes and influences monitoring).

The fact that these two components continuously guide one another is important, because it implies that we are always making metacognitive judgments about our environment. To the extent that our metacognitive judgments are inaccurate because of various factors (e.g., fluency of retrieval; e.g., Benjamin, Bjork, & Schwartz, 1998), our self-regulated learning decisions (i.e., deciding how much more to study) may be affected. Indeed, research has supported this supposition – when people are more accurate at monitoring the state of their memory, they perform better on later retention tests (e.g., Dunlosky & Rawson, 2012; Rawson, O'Neil, & Dunlosky, 2011; Thiede, 1999; Thiede, Anderson, & Therriault, 2003). The constant bidirectional relationship between monitoring and control further supports the notion that metacognitive processes are crucial in making decisions during self-regulated learning.

Judgments of learning

Judgments of learning (JOLs) reflect subjective judgments about an individual's ability to recall target information at a later time. In most studies examining JOLs, participants are shown simple materials (e.g., word-pairs) and are

CONTACT Heather Bailey ✉ hbailey@ksu.edu 📇 Kansas State University, 414 Bluemont Hall, Manhattan, KS 66506, USA

asked to predict the likelihood of being able to recall the item at a later time. These judgments are often made on a 0–100 scale, with 0 indicating a 0% chance of correctly recalling an item and 100 indicating a 100% chance of correctly recalling an item. Afterwards, participants receive a test on the items, and JOLs are compared to recall performance to assess the accuracy of participants' monitoring abilities. Two of the most common measures derived from JOLs and recall are *calibration* (absolute accuracy; a difference score of mean JOLs and mean recall, indicating how close JOL estimates are to actual recall) and *resolution* (relative accuracy; the correlation of JOLs and recall, used to estimate the degree to which participants give higher JOLs to recalled items and lower JOLs to non-recalled items).

JOLs are assessments of how well an individual has learned a particular piece of information. The extent to which JOLs are accurate is of particular importance because these judgments are used to control decisions during the self-regulated study (Ariel, Dunlosky, & Bailey, 2009; Metcalfe & Finn, 2008). That is, predictions about the likelihood of later recalling a target item are used to control further study (Benjamin et al., 1998; Mazzoni, Cornoldi, & Marchitelli, 1990; Metcalfe & Kornell, 2003; Nelson, Dunlosky, Graf, & Narens, 1994; Nelson & Narens, 1990; Son & Metcalfe, 2000). Indeed, Nelson and Dunlosky (1991) have argued that "the accuracy of JOLs is critical, because if the JOLs are inaccurate, the allocation of subsequent study time will correspondingly be less than optimal" (p. 267).

The Delayed-JOL effect

Considerable research has focused on conditions under which JOLs are more versus less accurate. One such manipulation that strongly influences JOL accuracy is the delay between initial learning and when the JOL is made. In a seminal study, Nelson and Dunlosky (1991) instructed participants to study a series of word-pairs ("ART : girl"). Participants were shown a partially intact word-pair ("ART : ______"), and asked to predict the likelihood that they would recall second word ("girl") on a later test either immediately after the study trial or after a delay. Results showed that JOL accuracy was greater when they were made after a delay compared to immediately, which is known as the *delayed-JOL effect*. Over time, this effect has received substantial attention with regards to its robustness (see Dunlosky & Metcalfe, 2009), including the benefit of delay on metacognitive monitoring in neuropsychological patients (Moulin, Perfect, Akhtar, Williams, & Souchay, 2011) as well as children at various stages of development (Schneider, Visé, Lockl, & Nelson, 2000). However, it should be noted that studies have also identified important boundary conditions (e.g., Dunlosky & Nelson, 1992), as well as instances in which delayed JOLs are inaccurate (for a meta-analysis, see Rhodes & Tauber, 2011).

Nelson and Dunlosky (1991) proposed a retrieval-based account such that when participants make a JOL, they attempt to retrieve the correct response and use the outcome from the retrieval attempt as a basis for making the JOL (for a different account of the delayed-JOL effect, see Sikstrom & Jonsson, 2005). Though it should be noted here that delayed JOLs might not elicit the same type of retrieval attempt as would be elicited on an intentional memory test (Son & Metcalfe, 2005; Tauber, Dunlosky, & Rawson, 2015). If participants are able to retrieve the response, they make a high JOL. If they are not able to retrieve the response, they make a lower one. Furthermore, it is assumed that JOLs can be made by retrieving information from either primary memory or secondary memory,[1] but that performance on a later test is likely based on retrieval from secondary memory. In situations where a JOL is made immediately after study, information about the to-be-remembered item is likely to be in primary memory. However, due to the transient nature of primary memory, information stored there is unlikely to be available later at final test. As a result, JOLs based on information in primary memory reduces the accuracy of JOLs because that information is less diagnostic of later recall. By contrast, according to this retrieval-based account, JOLs made after a delay are based on information retrieved from secondary memory. Since final test performance is also based on retrieval from secondary memory, delayed JOLs are made from a more diagnostic source, which leads to more accurate JOLs (although the diagnosticity of delayed JOLs is limited by certain features of the task, such as list length; Rhodes & Tauber, 2010).

Interestingly, Nelson and Dunlosky's explanation for the delayed-JOL effect suggests that the effect can emerge independently of making JOLs at a delay. More specifically, the delayed-JOL effect is not due to temporal delay per se, but to retrieving information from secondary memory (but see Tauber et al., 2015). For delayed JOLs, retrieval is a byproduct of the temporal delay between learning and the JOL. However, theoretically, it should be the case that JOLs made immediately after learning can be relatively accurate as long as the information used to make the JOL is retrieved from secondary memory. Before highlighting a scenario in which this might occur, we will first shift our focus to a recent model that makes predictions about the relationship between primary and secondary memory in the context of working memory tasks. We then discuss predictions of this model in relation to predicting the accuracy of JOLs.

Dual-component model of working memory

Unsworth and Engle (2007) have put forth a dual-component model of working memory, which posits that individual differences in working memory ability reflect differences in the ability to retrieve information from secondary memory when to-be-remembered information is displaced from primary memory. More specifically, there is a limit on the amount of information (four items; Cowan, 1999) that can be actively maintained in primary memory at any given time. When this limit is exceeded, some of the to-be-remembered information is displaced into secondary memory, from which it is retrieved when needed. Additionally, Unsworth and Engle suggested that to-be-remembered information also can be displaced from primary memory into secondary memory when the attentional resources needed to maintain the information in primary memory are diverted to a secondary task. More recently, it has been proposed that individual differences in working memory ability are also due to the maintenance of information (via attentional control; Unsworth, 2016; Unsworth, Fukuda, Awh, & Vogel, 2014; Unsworth, Spillers, & Brewer, 2009) and the size of the focus of attention (i.e., primary memory; Cowan, 2001).

The extent to which this secondary memory component underlying working memory is similar to the secondary memory component underlying long-term episodic memory has received limited empirical attention. Work by Loaiza, Rhodes, and Anglin (2013; see also Loaiza, Duperreault, Rhodes, & McCabe, 2015) has suggested that long-term semantic representations influence performance on working memory and episodic memory tasks in a similar manner, such that factors underlying long-term episodic memory may also moderate working memory functioning. To the extent that this is the case, experimental manipulations should produce effects on working memory tasks similar to those observed on tests of long-term episodic memory. This was the approach taken by Rose, Myerson, Roediger, and Hale (2010), who examined the effects of level of processing (Craik & Tulving, 1975) on the performance of a working memory task. The results of this study demonstrated that attending to different types of features (visual, phonological, semantic) of words at the time of encoding did not produce effects on working memory tasks like those typically observed on long-term memory tasks. However, subsequent studies have found such level of processing effects of to-be-remembered items on working memory tasks (Loaiza, McCabe, Youngblood, Rose, & Myerson, 2011; Rose, Buchsbaum, & Craik, 2014; Rose & Craik, Experiment 2, 2012). To reconcile this discrepancy, Rose and Craik (2012) suggested that the extent to which performance on working memory and long-term memory tests share common processes (and therefore similar patterns of results) depends on the extent to which working memory tests disrupt active maintenance of information in primary memory. When the degree of disruption is sufficient, performance on working memory tests is forced to rely on retrieval from secondary memory.

This approach of modifying working memory tasks to incorporate manipulations known to produce robust effects on long-term episodic memory tests can provide unique insights into the relationship between primary and secondary memory. Importantly, this approach allows one to test predictions made by the dual-component model of working memory with regards to the relation between primary memory and secondary memory during short-term memory tests. According to Unsworth and Engle's (2007) model, items are displaced from primary memory into secondary memory under two conditions: (1) When the capacity (four) of primary memory is exceeded, and (2) when attentional resources needed to maintain the to-be-remembered information in primary memory are diverted to a secondary task. Under these specific conditions, performance on a short-term memory task is thought to rely on retrieval from secondary memory.

Current study

If the demands of a secondary task (e.g., solving math problems) are sufficient to displace items from primary memory into secondary memory, then the accuracy of JOLs made shortly after study should be affected. More specifically, the addition of a demanding secondary task interleaved between study and JOL should increase the accuracy of immediate JOLs because the additional task will result in participants having to retrieve items from secondary memory, thereby providing them with diagnostic information.

Kelemen and Weaver (1997, Experiments 1–3) explored this possibility. In addition to having participants make JOLs under conditions similar to Nelson and Dunlosky (1991) – immediately after learning and after a longer delay (average of 4 min) – Kelemen and Weaver had participants make JOLs after brief periods (1–30 sec) filled with a short-term memory distractor task. Across those three experiments, results generally indicated that a few seconds of distraction via the secondary task increased relative accuracy compared to immediate JOLs, but not to the same extent as did delayed JOLs. Kelemen and Weaver (1997) concluded that short-term memory distractions increase JOL accuracy compared to immediate JOLs, and that temporal delays increase them even further.

However, this interpretation should be taken with caution, as the conditions used in their study do not disentangle the effects of task demands and temporal delay. That is, although it is possible that the distractor task increased JOL accuracy, it may also be the case that the temporal delay created by the distractor task increased JOL accuracy. Indeed, evidence has suggested that even very brief delays can increase JOL accuracy (e.g., Rhodes & Tauber, 2011). Thus, stronger evidence that short-term distractor tasks increase JOL accuracy requires unconfounding the effects of time and task demand. One way to do this is to keep the temporal delay between learning and JOL constant, and manipulate the task demand during this interval (e.g., Bjork & Allen, 1970; Bui, Maddox, & Balota, 2013; Roediger & Crowder, 1975). This allows for direct examination of JOL accuracy as a function of task demands, and is the approach we take in this current study.

Turning to the goals of this study, we sought to evaluate whether JOLs made after completing a secondary task are (1) more accurate than JOLs made after the same length of time, but not requiring completing a secondary task, and (2) just as accurate as JOLs made after a longer delay. Our prediction is that JOLs will be most accurate when information is retrieved from secondary memory than when it is not. Specifically, delayed JOLs will be more accurate than immediate JOLs that do not have an intervening secondary task. More critically, immediate JOLs made with an intervening secondary task will more accurate than immediate JOLs made without an intervening secondary task, and just as accurate as delayed JOLs. Supporting these goals would be important for two reasons. First, this work would provide additional evidence that delayed JOLs are indeed more accurate because they elicit retrieval from secondary memory. Second, this study may provide evidence consistent with the view that the secondary memory component thought to be important for working memory tasks (Unsworth & Engle, 2007) is the same as the secondary memory component thought to underlie long-term episodic memory tasks (e.g., Atkinson & Shiffrin, 1968).

JOL conditions

Across four experiments, participants studied weakly associated word-pairs, some of which participants were asked to provide JOLs on. Importantly, these JOLs were made in one of three conditions: delayed, displaced, and maintenance. In the delayed condition, JOLs were made after a larger number of intervening trials consisting of various tasks (other intervening items & math problems; the target item is retrieved from secondary memory), and we will refer to these as *delayed JOLs*. In the displaced condition, JOLs were made after a couple of math problems (the target item is displaced from primary memory to secondary memory). We will refer to these as *displaced JOLs*. Critically, this new condition will allow us to evaluate whether secondary task demands – and not just temporal delay – influences JOL accuracy. We chose math problems because they have consistently been used as secondary tasks in working memory span tasks (e.g., Operation Span [OSPAN]; Turner & Engle, 1989; Unsworth, Heitz, Schrock, & Engle, 2005). Further, math problems require enough attentional resources to displace other information from primary memory (e.g., Unsworth & Engle, 2007). Finally, JOLs in the maintenance condition were made on the target item after a couple of intervening word-pairs (the target item presumably is still maintained in primary memory because the task is relatively easy), and we will refer to these as *maintenance JOLs*. At this point, we acknowledge that JOLs made shortly after learning in our study are not the same as those used by Nelson and Dunlosky (1991). In their study, JOLs were made immediately after learning the target, whereas JOLs in our study were made after a very short delay. Given this difference, we will refer to this condition as "maintenance" rather than "immediate"; however, we believe that our maintenance JOLs and Nelson and Dunlosky's immediate JOLs are both made while the target information is still residing in primary memory. We will revisit this assumption in more detail in the General Discussion.

Experiment one

Method

Design and participants

Sixty undergraduate students from Washington University in St. Louis (30 females; *M* age = 19.3 years, *SD* = 1.3) participated in the study to partially fulfil a course requirement, and reported being proficient English speakers. JOL type (delayed vs. displaced vs. maintenance) was manipulated within participants. Sample sizes for all studies were based on power analyses conducted using G*Power with an alpha level of .05, 95% power, and an effect size of 0.65 (based on data from Experiment 3 in Kelemen & Weaver, 1997).

Materials

One hundred weakly related word-pairs were selected from a database of word associates (Nelson, McEvoy, & Schreiber, 1998). Cues had an average word length = 5.09 and an average number of syllables = 1.49, whereas targets had an average word length = 4.98 and average number of syllables = 1.47. These word-pairs were chosen for their weak forward and backward associative strength (*M* = .01; Maddox, Balota, Coane, & Duchek, 2011). Ten word-pairs served as primacy buffer items, and were presented at the beginning of the study phase for all participants. Thirty of the word-pairs were target items for which participants made JOLs, with 10 word-pairs assigned to and counterbalanced across each of the three conditions of interest. The remaining 60 word-pairs served as filler items. In addition to the word-pairs, 20 math problems were selected to serve as the secondary task, and were selected such that they were similar with regards to how long they took to verify (*M* = 3186.7 ms, *SD* = 89.2), reducing the amount of variability in the duration of the secondary task (math problems).

Procedure

Participants were told that they would see a series of word-pairs ("ART : girl"), each for 5 s, and that they would need to recall the second item in the pair when prompted with the first word on a test that took place later during the same session. They were also told that on some trials, they would be given 5 s to verify ("Yes" or "No") via a button click the accuracy of a given math problem ("[4 × 2]−1 = 9?"). The math problems remained onscreen for the entire 5 s, regardless of how long it took participants to verify the answer. Finally, participants were told that on some trials, they would be given a word-pair that was presented earlier during the study phase, but with only the first word intact ("ART : ______"). On these trials, they were given 5 s to indicate their likelihood of recalling the second word ("girl") on a later test by typing in a number between 0-100 in a provided text box, where 0 indicates that there is no chance they will be able to correctly recall the second word later, and 100 indicates that they will for sure be able to correctly recall the second word later. The JOL prompt remained on the screen for the entire 5 s, regardless of how long it took participants to type in their response.

For the maintenance condition, after an initial study trial with a given item ("ART : girl"), participants were presented with a filler word-pair ("TOOL: hand"), which was immediately followed by a JOL prompt for the target maintenance item ("ART : ______"). For target word-pairs assigned to the displaced condition, the presentation of the target item was followed by a math problem, which was immediately followed by a JOL prompt for the target item. Finally, the presentation of target word-pairs assigned to the delayed condition was followed by six trials that loosely consisted of a varying number word-pairs, math problems, and JOLs for target items assigned to one of the other two conditions. After these six trials, participants made a JOL on the target item assigned to the delayed condition. It is important to note that the study phase was structured in such a way that filler items and math problems made it difficult for participants to anticipate when they would be making a JOL for any given item. For example, the presentation of two consecutive word-pairs was not always followed by a JOL (as is the case in the maintenance condition), nor was a math problem always followed by a JOL (as is the case in the displaced condition). By making JOL prompts appear unpredictable to the participants, we decreased the likelihood that participants would differentially engage in the processing of the word-pairs across the different conditions (see Figure 1 for a visual depiction of the different JOL conditions).

After the study and JOL phase, participants were administered a non-verbal distractor task (Consonant–Vowel/Odd-Even Switching Task; e.g., Duchek et al., 2009) for 5 min. After 5 min had elapsed, participants were tested on the word-pairs presented during the study phase. Participants were given the first half of each word-pair ("ART : ______"), and were given up to 10 s to type in the correct target word ("girl") in a blank text box on the computer screen. The order of the word-pairs in the test phase was randomised across all participants.

Results

For all experiments, we focused on the effect of condition on JOL, recall, and the difference score between those two measures (i.e., calibration). In addition, for each participant, we derived a resolution measure by computing a Goodman–Kruskal gamma correlation between JOL and subsequent recall, with one gamma correlation calculated for each of the three conditions of interest (delayed vs. displaced vs. maintenance). The descriptive statistics for each measure as a function of condition are shown in Table 1.[2] All of these measures were analysed using one-way repeated measure analyses of variance (ANOVA), with follow-up tests of significant interactions conducted with Bonferroni corrections.

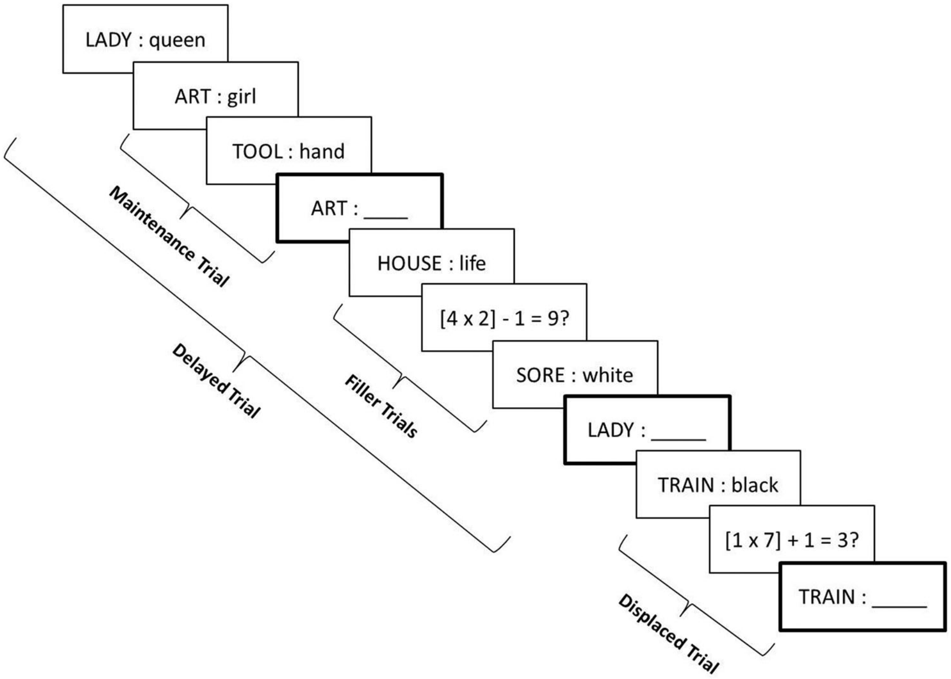

Figure 1. General proceure of learning phase in Experiment 1. Bolded boxes indicate JOL trials.

Table 1. Descriptive statistics for JOL, recall, calibration, and resolution as a function of condition for Experiments 1–3.

		Delayed	Displaced	Maintenance	*F*	*p*	η^2
Experiment One	**JOL:**	.45 (.20)	.51 (.21)	.55 (.20)	3.43	.035	.04
n = 60	**Recall:**	.42 (.25)	.40 (.23)	.41 (.24)	0.10	.906	.00
	Calibration:	.03 (.15)	.11 (.24)	.14 (.21)	4.67	.011	.05
	Resolution:	.78 (.41)	.60 (.41)	.53 (.39)	6.27	.002	.07
Experiment Two	**JOL:**	.51 (.20)	.49 (.21)	.52 (.20)	0.31	.737	.00
n = 60	**Recall:**	.50 (.23)	.47 (.26)	.44 (.22)	0.97	.382	.01
	Calibration:	.02 (.20)	.03 (.21)	.08 (.24)	1.66	.193	.02
	Resolution:	.82 (.27)	.81 (.36)	.65 (.46)	3.91	.022	.04
Experiment Three	**JOL:**	.44 (.22)	.45 (.21)	.46 (.23)	0.10	.904	.00
n = 60	**Recall:**	.36 (.23)	.35 (.24)	.29 (.21)	1.98	.141	.02
	Calibration:	.08 (.21)	.10 (.22)	.17 (.23)	3.04	.049	.04
	Resolution:	.79 (.32)	.79 (.34)	.61 (.49)	3.90	.022	.04

Note: Standard deviations are in parenthesis. JOL scores were converted from percentage scores to proportion to allow for a direct comparison against recall performance (proportion correct). Calibration scores are calculated as the difference between JOL and recall scores, and resolution scores are presented as gamma correlations.

Results from this experiment showed an effect of condition on JOLs, *F*(2, 117) = 3.43, $p < .035$, $\eta^2 = .04$. Follow-up tests indicated that delayed JOLs were lower than maintenance JOLs, *t*(59) = 5.08, $p < .001$, $d = 0.50$, and displaced JOLs, *t*(59) = 2.43, $p = .018$, $d = 0.29$. Finally, displaced JOLs were marginally lower than maintenance JOLs, *t*(59) = 1.79, $p = .078$, $d = 0.20$. There was no difference in recall between the three conditions, *F*(2, 177) = 0.10, $p = .906$, $\eta^2 = .00$. However, the condition did have an effect on calibration, *F*(2, 177) = 4.67, $p = .011$, $\eta^2 = .05$. Follow-up tests indicated that calibration was better in the delayed condition than in the maintenance, *t*(59) = 5.03, $p < .001$, $d = 0.60$; and displaced conditions, *t*(59) = 2.73, $p = .008$, $d = 0.40$. Additionally, calibration for the displaced condition was not different from the maintenance condition, *t*(59) = 1.09, $p = .14$, $d = 0.13$.

As can be seen in Figure 2(A), a calibration curve was used to evaluate the accuracy of JOLs in predicting test performance (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). In this context, we can aggregate items receiving

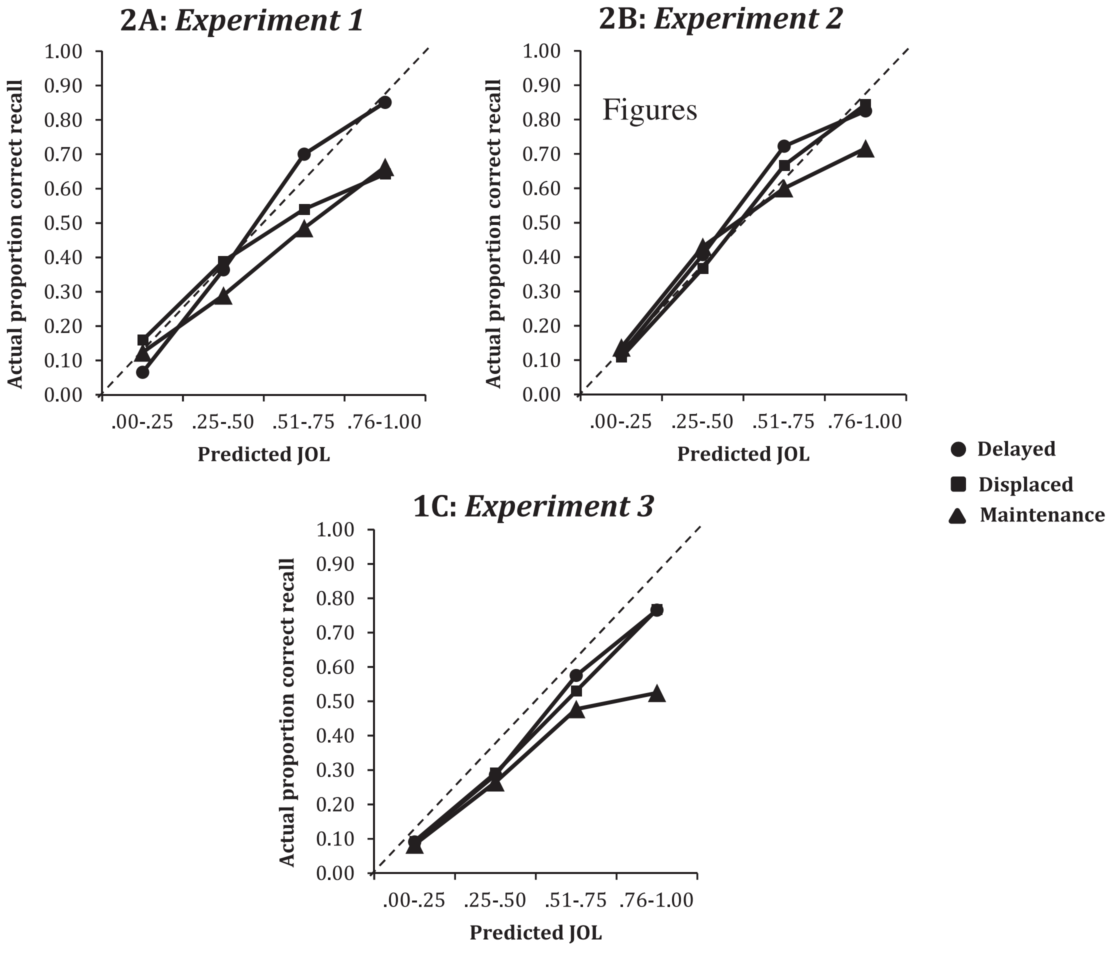

Figure 2. Calibration curves for each experiment as a function of each of the three JOL conditions. The dotted diagonal line represents perfect calibration.

the same JOLs and compute the mean recall for those items, which permits examination of the relationship between predicted and actual proportion recalled. For example, given all of the items on which a participant gave JOL estimates of 50%, what proportion was correctly recalled on the final test? The extent to which actual recall is closer to the 50% estimate represents better calibration. In Figure 2, the dashed diagonal line represents perfect calibration, and absolute accuracy of the delayed JOLs falls closest to the diagonal. Perhaps more importantly, delayed JOLs are closer to the diagonal than either displaced and maintenance JOLs, which do not appear to differ from one another.

In keeping with the analyses conducted by Nelson and Dunlosky (1991), a linear regression equation was calculated with JOLs (x) predicting recall (y). To do so, we aggregated items with the same predicted recall, and then actual recall performance was predicted from the predicted recall performance. This analysis provides an alternative method for assessing the different calibration curves. Given that the diagonal line in the figure represents perfect calibration (and thus represented in a linear equation as: $y = 1.0x + 0.0$), values for the slope that are closer to 1.0 would presumably represent better calibration. Indeed, a dummy-coded interaction term confirms that the slope for the delayed condition ($y = 0.70x + .00$) is significantly greater than that of the displaced ($y = 0.40x + 0.13$) and maintenance conditions ($y = 0.45x + .05$), p's < .001. In addition, the slopes in the displaced and maintenance conditions did not differ from one another, $p = .31$.

Finally, gamma correlations were calculated to examine the extent to which resolution (or relative accuracy) differed between the three conditions. Again, resolution is the degree to which participants give higher JOLs to recalled items and lower JOLs to non-recalled items. It is important to note that it was not always possible to calculate a gamma correlation for all three conditions for each participant. For example, correlations could not be calculated when participants did not use a wide enough range of values for JOLs, or if they recalled all (or none) of the target items in a given condition. In this experiment, these scenarios were rare (9% of the data), but when they did occur, we used an expectation-maximisation procedure (Dempster, Laird, & Rubin, 1977) to estimate the missing values (see note 2). We chose this method for managing missing data because it overcomes the issue of underestimating standard errors that other techniques encounter (e.g., mean substitution, regression substitution), which in turn can artificially make it easier to detect effects. A one-way ANOVA revealed an effect of condition on resolution, $F(2, 177) = 6.27$, $p = .002$, $\eta^2 = .07$. Follow-up tests indicated that resolution in the delayed condition was better than the maintenance condition, $t(59) = 3.72$, $p < .001$, $d = 0.63$, and the displaced condition, $t(59) = 2.69$, $p = .005$, $d = 0.44$. Additionally, resolution in the displaced condition was no different than the maintenance condition, $t(59) = 1.14$, $p = .13$, $d = 0.18$.

Discussion

The results of Experiment 1 suggest that calibration (difference scores and regression slopes) and resolution (gamma correlations) measures were better when JOLs were made after several intervening trials (delayed), compared to more immediate JOLs when an intervening secondary task was (displaced) and was not (maintenance) present. The finding that metacognitive judgments were more accurate in the delayed condition compared to the maintenance condition replicates the results found by Nelson and Dunlosky (1991). However, we were unable to find evidence supporting our prediction that accuracy of metacognitive judgments in the displaced condition would be similar to the delayed condition and greater than the maintenance condition. Instead, metacognitive judgments in the displaced condition were less accurate than the delayed condition, and more similar to the maintenance condition.

One possible explanation is that our secondary task (i.e., one math problem) did not displace to-be-remembered information from primary memory into secondary memory. Unsworth and Engle (2007) claim that a secondary task will only displace information to secondary memory to the extent that it requires attention. They state that "if attention is removed, because new information is intentionally being processed or because attention has been captured by environmental stimuli (e.g., a flashing light), representations are displaced from PM" (Unsworth & Engle, 2007, p. 107). However, their model makes no predictions about the relative attentional demand of the secondary task.

Contrary to Unsworth and Engle's view, Rose and Craik (2012) proposed a more graded view in which the likelihood of to-be-remembered items being retrieved from secondary memory on working memory tasks depends on the extent to which tasks immediately preceding recall disrupt the maintenance of items in primary memory. Accordingly, they present aggregated data from several studies that reveal a positive correlation between time spent making levels-of-processing judgments (response time differences between deep and shallow judgments) and levels-of-processing effects (benefit of deep over shallow levels of processing) on working memory tasks. Additionally, Rose et al. (2014) report results in which levels-of-processing effects on working memory tasks are seen when secondary tasks are more difficult (and thus more attentionally demanding), but not when they are easier (and less attentionally demanding). Thus, the mechanism whereby to-be-remembered items are displaced from primary memory when attentional resources shift to a secondary task is not an all-or-none one. Indeed, a similar view is also shared by the time-based resource sharing model (TBRS; Barrouillet, Bernardin, & Camos, 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007). This model proposes that individuals use attentional resources to keep the to-be-remembered information accessible in working memory, and that these memory traces suffer from a time-based decay when attention is shifted away to a secondary task. However, these traces can be kept accessible by a refreshing mechanism, which requires individuals to switch attention from the secondary task back to the to-be-remembered information. This switching is thought to take place during brief periods of time in which the secondary task does not require attention. Thus, it becomes more difficult to refresh the memory trace of the to-be-remembered information during more attentionally demanding secondary tasks. Thus, along with Rose and Craik (2012), the TBRS model suggests that the influence of a secondary task depends on how attentionally demanding it is.

With regards to the current study, it may be the case that recalls immediately following less attentionally demanding secondary tasks may not require secondary memory. Additionally, it may be the case that the accuracy of displaced JOLs depends on how demanding the secondary task is. More specifically, the accuracy of displaced JOLs should be similar to delayed JOLs when the secondary task is more demanding (i.e., more difficult and/or more time-consuming task). Following this logic, we increased secondary task demands in Experiments 2–4 by adding an additional intervening math problem prior to the JOLs in the displaced condition. Assuming that task demands are sufficient to increase dependence on secondary memory, one would expect to see not only similar metacognitive accuracy for displaced and delayed JOLs, but also greater accuracy for displaced than maintenance JOLs.

Experiment two

Method

Sixty undergraduate students from Washington University in St. Louis (30 females; *M* age = 19.3 years, *SD* = 1.3) participated in the study as partial course fulfilment, and reported being proficient English speakers. The design, materials, and procedures used in this experiment were identical to Experiment 1 except for a few important changes. For target items assigned to the displaced condition, there were now two intervening math problems for participants to complete before they made their JOL. Therefore, for target items in the maintenance condition participants were shown two intervening word-pairs before making their JOL. JOLs in the delayed condition were made after eight trials that again loosely consisted of word-pairs, math problems, and JOLs for target items assigned to one of the other two conditions. Each discrete trial (word-pair, math problem, JOL) was 5 s long. As was the case in Experiment 1, the study phase was structured such that filler items and math problems made it difficult for participants to anticipate when they would be making JOLs. Since these modifications added additional trials to the study phase, we reduced the number of word-pairs assigned to each of the three conditions from 10 to 8, keeping the length of the study phase comparable to that of Experiment 1.

Results

Table 1 provides descriptive statistics for the three conditions. One-way repeated measure ANOVAs revealed no effect of condition on JOLs, $F(2, 177) = 0.31$, $p = .737$, $\eta^2 = .00$, recall, $F(2, 177) = 0.97$, $p = .382$, $\eta^2 = .01$, or calibration (mean JOL minus mean recall), $F(2, 177) = 1.66$, $p = .193$, $\eta^2 = .02$. Although calibration in the delayed ($M = .02$) and displaced ($M = .03$) conditions was numerically smaller than calibration in the maintenance condition ($M = .08$), this did not reach significance. The calibration curve (Figure 2 (B)) depicts three important points: the line for delayed and displaced JOLs (1) are close to the (perfect calibration) diagonal line, (2) do not appear to differ from one another, and (3) are closer to the diagonal than the line for maintenance JOLs. Linear regression equations support these observations: The slopes for the delayed condition ($y = 0.62x + .05$) and the displaced condition ($y = 0.61x + 0.03$) were significantly greater than the slope in the maintenance condition ($y = 0.48x + .09$), p's < .035; though the slopes in the delayed and displaced condition did not differ from one another, $p = .805$.

With regards to resolution (gamma correlations), the expectation-maximisation procedure was again used to estimate the missing values, which represented 8% of the data in this analysis. A one-way ANOVA revealed an effect of condition, $F(2, 177) = 3.91$, $p = .022$, $\eta^2 = .04$. Follow-up comparisons indicated that resolution in the delayed condition ($M = .82$) was better than the maintenance condition ($M = .65$), $t(59) = 2.39$, $p = .020$, $d = 0.45$, but no different from the displaced condition ($M = .81$), $t(59) = 0.16$, $p = 1.00$, $d = 0.03$. Importantly, resolution in the displaced condition was better than the maintenance condition, $t(59) = 2.17$, $p = .034$, $d = 0.39$.

Discussion

Our results partially replicated the findings from Experiment 1, showing that resolution for delayed JOLs was better than maintenance JOLs (e.g., Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011). Most important, and consistent with our predictions, when the secondary task was more demanding (i.e., including an additional math problem), resolution for displaced JOLs was similar to delayed JOLs, and better than maintenance JOLs. Before discussing the implications of these findings, however, we wanted to not only replicate these findings, but to do so with a more diverse sample.

Experiment three

Method

Sixty participants (39 females; *M* age = 31.8 years, *SD* = 8.2), were recruited from Amazon's Mechanical Turk web site to take part in this study for monetary compensation ($0.60), and reported to be proficient English speakers. JOL type (delayed vs. displaced vs. maintenance) was manipulated within participants. The materials and procedure were identical to Experiment 2.

Results

Table 1 provides descriptive statistics for the three conditions. No effects of condition on JOLs, $F(2, 177) = 0.10$, $p = .904$, $\eta^2 = .00$, or recall were observed, $F(2, 177) = 1.98$, $p = .141$, $\eta^2 = .02$. There was, however, a significant effect of condition on calibration (mean JOL minus mean recall), $F(2, 177) = 3.04$, $p = .049$, $\eta^2 = .04$, such that calibration was better in the delayed condition than in the maintenance condition, $t(59) = 3.60$, $p < .001$, $d = 0.44$, but not different from the displaced condition, $t(59) = 0.63$, $p = 0.53$, $d = 0.09$. Importantly, calibration in the displaced condition was better than in the maintenance condition, $t(59) = 2.51$, $p = .007$, $d = 0.33$. As in Experiment 2, the lines for delayed and displaced JOLs in the calibration curve depicted in Figure 2(C): (1) are close to the (perfect calibration) diagonal line, (2) do not appear to differ from one another, and (3) are closer to the diagonal than the line for maintenance JOLs. Indeed, linear regression equations support these observations: the slopes for the delayed condition ($y = 0.62x + .03$) and the displaced condition ($y = 0.61x + 0.05$) were greater than the slope in the maintenance condition ($y = 0.44x + .05$), p's $< .01$. Importantly, the slopes in the delayed and displaced conditions did not differ from one another, $p = .42$.

With regards to resolution (gamma correlations), the expectation-maximisation procedure was used to estimate the missing values (9% of the data).[3] A one-way ANOVA revealed an effect of condition, $F(2, 177) = 3.89$, $p = .022$, $\eta^2 = .04$. Follow-up tests indicated that resolution in the delayed condition was better than in the maintenance condition, $t(59) = 2.31$, $p = .012$, $d = 0.43$, but no different from the displaced condition, $t(59) = 0.14$, $p = .893$, $d = 0.02$. Finally, resolution in the displaced condition was better than the maintenance condition, $t(59) = 2.32$, $p = .012$, $d = 0.40$.

Discussion

The results of Experiment 2 and 3 indicate that calibration (difference scores and regression slopes) and resolution (gamma correlations) measures were best when JOLs were made after a longer temporal interval (delayed), or shortly after learning with an intervening secondary task (displaced), compared to when JOLs were made shortly after learning without an intervening secondary task (maintenance). Most notably, the level of accuracy did not differ between the delayed and displaced conditions. The fact that metacognitive accuracy for the displaced condition in Experiments 2 and 3 was improved compared to the displaced condition in Experiment 1 is likely attributable to increased secondary task difficulty (i.e., two math problems). However, Experiment 4 directly tested this hypothesis by manipulating whether participants solved one versus two math problems in the interval between learning the word-pairs and making a JOL.

Experiment four

One important thing to note is that we did not use a true immediate condition in Experiments 1–3. As mentioned at the beginning of this paper, our interpretation of "immediate" is based on the assumption that the target item on which the JOL is being made is in primary memory, and that JOLs made after no intervening trials are likely to be inaccurate for the same reasons as JOLs made after one or two intervening trials in our maintenance conditions. To provide more leverage for the argument presented above, Experiment 4 included maintenance and displaced conditions similar to those used in Experiments 1–3 and, most importantly, compared them to a true immediate JOL condition (Table 2).

Method

Forty participants (25 females; M age = 32.4 years, $SD = 7.7$), were recruited from Amazon's Mechanical Turk website to take part in this study for monetary compensation ($0.60), and reported being proficient English speakers. The materials and procedure were identical to Experiment 3, except participants made JOLs in the following five conditions: (1) the *true immediate condition* included no intervening items between study and JOL, (2) the *maintenance-one condition* had one intervening word pair, (3) the *maintenance-two condition* had two intervening word-pairs, (4) *the displaced-one condition* had one intervening math problem, and finally (5) the *displaced-two condition* had two intervening math problems. These conditions were manipulated within participants.

Results and discussion

A repeated-measures ANOVA revealed a significant effect of condition on JOLs, $F(4, 195) = 3.39$, $p = .011$, $\eta^2 = .06$. Follow-up comparisons indicated that JOLs in the true immediate condition were significantly higher than those in the other four conditions. There was no effect of condition on recall, $F(4, 195) = 0.92$, $p = .454$, $\eta^2 = .01$.

Another repeated-measures ANOVA revealed a significant effect of condition on calibration, $F(4, 195) = 3.88$, $p = .005$, $\eta^2 = .07$. *Post-hoc* comparisons indicated that only the true immediate (no intervening trials; $M = 26.16$) and displace-two conditions (two intervening math problems; $M = 3.57$) differed significantly, the calibration of the remaining three conditions falling somewhere in the middle (maintenance-one: 11.99, maintenance-two: 10.82, displace-one: 14.46).

There was a significant effect of condition on resolution, $F(4, 195) = 4.02$, $p = .004$, $\eta^2 = .08$. *Post-hoc* comparisons indicated that the resolution in the displaced-two condition ($M = 0.81$) was significantly better than the

Table 2. Descriptive statistics for JOL, recall, calibration, and resolution as a function of condition for Experiment 4.

		Displaced-2	Displaced-1	Maintenance-2	Maintenance-1	Immediate	*F*	*p*	η^2
Experiment Four	**JOL:**	.38 (24)	.41 (.26)	.46 (.23)	.42 (.27)	.57 (.31)	3.39	.011	.06
n = 40	**Recall:**	.34 (.19)	.27 (.24)	.35 (.22)	.31 (24)	.32 (.21)	0.92	.454	.01
	Calibration:	.04 (.20)	.14 (.27)	.11 (.22)	.11 (.28)	.26 (.33)	3.88	.005	.07
	Resolution:	.81 (.30)	.56 (.48)	.46 (.54)	.51 (.55)	.42 (.53)	4.02	.004	.08

Note: Standard deviations are in parenthesis. JOL scores were converted from percentage scores to proportion to allow for a direct comparison against recall performance (proportion correct). Calibration scores are calculated as the difference between JOL and recall scores, and resolution scores are presented as gamma correlations.

resolution in the true immediate (*M* = .42), maintenance-one (*M* = .51) and maintenance-two (*M* = .46) conditions, which did not differ from one another. Resolution in the displace-one condition (*M* = .56) did not differ from that of the displace-two nor the other three conditions.

Most important, metacognitive measures for the true immediate JOL condition did not differ significantly from our maintenance conditions used in the earlier experiments; however, calibration and resolution were significantly worse for the true immediate condition as compared to our displace-two condition. Such evidence indicates that JOLs are less accurate when the to-be-remembered information is maintained in primary memory as compared to when it must be retrieved from secondary memory.

General discussion

The goal of this study was to evaluate whether intervening secondary task demands influence metacognitive accuracy. Our results demonstrated that JOLs made shortly after learning were just as accurate as delayed JOLs, but only when the secondary task was sufficiently demanding to displace the target item from primary to secondary memory (Experiments 2–4). More specifically, in the first experiment, when JOLs were made after one intervening math problem (displaced), calibration and resolution were no different than if JOLs were made after an intervening study trial (maintenance), and less accurate than JOLs made after a larger number of intervening trials (delayed). In the final three experiments, when secondary task demands were increased, JOLs made after two intervening math problems (displaced) were not only just as accurate as JOLs made after a larger number of intervening trials (delayed), but critically, more accurate than JOLs made after two intervening trials that involved study only (maintenance). However, across all four experiments, we provide evidence that JOLs made shortly after study without any intervening secondary tasks are less accurate than JOLs made after a delay, replicating previous findings (Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011).

It is important to note that the temporal interval between study and JOLs for the displaced and maintenance conditions were the same. Thus, the presumed critical difference was that JOLs in the displaced condition required retrieval of information from secondary memory, whereas JOLs in the maintenance condition did not. As suggested by Nelson and Dunlosky (1991), to the extent that JOLs are made by retrieving information from secondary memory, they will be more accurate than JOLs made by retrieving of information from less diagnostic sources such as primary memory. Thus, immediate JOLs are typically not as accurate as delayed JOLs because the source of information used to drive metacognitive processes is likely coming from primary memory. In contrast, the passage of time that occurs for delayed JOLs displaces information from primary memory into secondary memory. By the time JOLs are made, information is coming from a much more diagnostic source (secondary memory). This seemed to be the case in Experiments 2–4, and provided evidence that the delayed-JOL effect is due primarily to retrieval demands from a more diagnostic source (i.e., secondary memory).

Comparing calibration and resolution between the displaced and delayed conditions, the results from Experiments 2–4 provide support for the idea that the secondary memory component that Unsworth and Engle (2007) hypothesised underlies performance on working memory tasks is the same as that underlying long-term episodic memory tests. In the context of Unsworth and Engle's model of working memory, it seems then that when information can no longer be maintained in primary memory, successful recall relies at least in part on the same secondary memory component that supports long-term episodic memory. Compared to earlier conceptualisations of short-term memory as distinct from long-term memory (e.g., Atkinson & Shiffrin, 1968), this paints a much more intimate relationship between primary memory and secondary memory.

Taken together, it appears as though the observed differences in metacognitive accuracy between Experiment 1 and Experiments 2–4 are due to the additional intervening trial. More specifically, the increase in the number of math problems to be completed before making a JOL likely forced participants to retrieve the target item from secondary memory. This interpretation is consistent with the hypothesis that whether items must be retrieved from secondary memory at the time of recall depends on the extent to which resources have been diverted away from maintaining those to-be-remembered items in primary memory (Rose & Craik, 2012). As Rose and Craik point out, if secondary task demands are low, items can still be maintained in primary memory, and thus experimental manipulations that produce robust effects on long-term episodic memory tests (which primarily tap secondary memory) may have little effect on short-term memory tests.

As secondary task demands increase, however, so should dependence on secondary memory. In such cases, patterns of results like those normally observed on long-term episodic memory tasks also will be more likely to be observed on short-term memory tests. Aggregating results from across several different studies, Rose and Craik (2012) showed that levels-of-processing effects were observed on working memory tasks when the processing time required by the secondary task was longer, suggesting that the processing was more demanding than when processing times were shorter and thus processing was presumably less demanding. For example, processing times in deep processing conditions were approximately 5000 ms in Loaiza et al. (2011, Experiment 1) in which a levels-of-processing effect was observed, but only about 1000 ms in Rose et al. (2010), in which they observed no levels-of-processing effect. Finally, Rose et al. (2014) obtained more direct evidence in which levels-of-processing effects on working memory were observed only when the secondary task was more difficult, and not when it was easier. In light of these previous results, it may not be surprising that when task demands were made greater in Experiments 2–4 of the present paper, resulting in processing times similar to those in Loaiza et al. (across experiments, the average for was 6183 ms compared to 3080 ms in Experiment 1), JOL accuracy in the displaced condition increased.

One thing to consider is whether new memoranda (the presentation of additional word-pairs after the target word-pair; maintenance condition) invokes similar demands as a secondary task (the verification of math problems; displace condition), as may be implied by the Unsworth and Engle (2007) framework. Indeed, Unsworth and Engle (2006) provide evidence that although performance on complex span tasks and longer trials of simple span tasks share some variability, they also account for independent variance in fluid intelligence. Furthermore, Loaiza and McCabe (2012) showed no benefit on a delayed memory test on longer trials of a simple span task compared to shorter trials, contrary to what is typically observed with complex span tasks (the "McCabe effect"; cf. McCabe, 2008; see also Loaiza et al., 2015). With regards to this study, a secondary task may be more attentionally demanding than incoming memoranda, and as a result, some degree of asymmetry between the two in terms of displacement to secondary memory.

It may also be noted that the small number of JOL observations in each condition of this study made the gamma correlations (resolution) somewhat volatile across participants. Indeed, with only eight possible observations in each condition, interpretation of our results regarding resolution should be made somewhat cautiously (see Spellman, Bloomfield, & Bjork, 2008). Additionally, in Experiments 1–3, the fact that JOLs were not provided on all trials occasionally created situations where less than eight observations were provided (an issue we resolved in the experiment described in Footnote 3), further creating less stable estimates of resolution. However, resolution in the displaced condition was better than in the maintenance condition and no different from the delayed condition across three experiments (Experiments 2–4), which should presumably alleviate any concerns of stability and allow us to make appropriate interpretations of our findings. Moreover, across all four experiments, resolution was better in the delayed condition compared to the maintenance condition, consistent with prior literature. Finally, it should be noted that these patterns of results are largely similar to what we obtain with our calibration measures, providing converging evidence that metacognitive accuracy is partly mediated by secondary task demands.

Closing thoughts

The results of these studies also have important implications for self-regulated learning, as JOLs play an important role in guiding behaviour during learning. Although there may be several ways to improve the accuracy of JOLs, one of the more popular prescriptions for yielding accurate JOLs involves making JOLs with a long temporal delay after learning, as this technique merely requires altering a learning schedule and does not require using a new cognitive strategy (which may take time to learn). The results from the current experiments indicate a new method of improving JOL accuracy that is not only more effective than JOLs made immediately after learning, but is just as effective as (and perhaps more efficient than) delayed JOLs. Indeed, engaging in another task not related to the JOL itself, even for shorter periods of time, may be sufficient to facilitate retrieval from more diagnostic sources (secondary memory). This provides an alternative strategy for students to use when learning new information, especially during situations where time may be limited. Moreover, though there are circumstances in which immediate JOLs can be relatively accurate, these situations are not practical to the extent that students are not able to choose the material that they need to learn. In contrast, our results provide evidence for an effective method that is more practical for students to implement. Specifically, participants could engage in some unrelated task for a short period of time (in our case, solving math problem, but one could perhaps play a challenging game on one's phone for a short period of time) and then evaluating how well they know a given piece of information before moving on to the next to-be-learned item.

In summary, the current experiments demonstrate that JOLs are diagnostic of later performance if the target information is retrieved from secondary memory. Importantly, this occurs not only with a temporal delay between learning and the JOL, but also when there is a sufficiently demanding intervening task that prevents the target information from being maintained in primary memory. In both circumstances, metacognitive accuracy is greater compared to instances where JOLs do not rely on retrieval from secondary memory, which highlights the importance of secondary memory (see also, Nelson & Dunlosky, 1991) for making accurate predictions of later performance, and also supports predictions from a dual-component viewpoint of working memory (Unsworth & Engle, 2007).

Notes

1. It is useful at this point to define terms that, although often used interchangeably, can also be a source of confusion. A lack of consistency arises because primary memory is often used interchangeably with short-term memory, and secondary memory with long-term memory. However, as we will discuss shortly, it may not be accurate to assume that such terms are transposable. Instead, we will adopt Craik and Lockhart's (1972) suggestion that we should think about primary and secondary memory as systems, and short/long-term memory as referring to the tasks that measure the contribution of these systems.
2. The means provided in Table 1 and used in the reported analyses are those calculated with the estimated data provided by the expectation-maximisation procedure. The means and standard deviations obtained from this procedure are not substantially different from the corresponding values obtained if we removed the participants who did not have gamma correlations for all three conditions. This was the case in the analyses of gamma correlations across all four experiments reported in this paper. However, such a listwise deletion method would have reduced power, and the likelihood of detecting an effect of condition. It should be noted that when using data only from participants who provided gamma correlations for all three conditions, a significant effect of condition was observed for all the reported experiments except for Experiment 2.
3. One potential concern is that participants did not provide a JOL on all of the target trials (Experiment 1: 8% of trials without JOL; Experiment 2: 9%; Experiment 3: 7%), which may influence metacognitive accuracy. However, in a separate experiment we obtained the same pattern of results when participants were allowed as much time as they needed to make their JOLs. On average, participants spent 5605.28 ms ($SD = 1181.73$) on a trial providing a JOL, which was significantly longer than the 5000 ms allotted for the previous three experiments, $p = .002$. However, metacognitive accuracy (as measured by calibration and resolution) was equivalent for the displaced and delayed JOL conditions. More importantly, metacognitive accuracy for the displaced condition was higher than metacognitive accuracy for the maintenance condition.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*, 432.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89–195.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*, 83–100.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 570–585.

Benjamin, A. A., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*, 567–572.

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin & Review*, *20*, 341–347.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge: Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Duchek, J. M., Balota, D. A., Tse, C. S., Holtzman, D. M., Fagan, A. M., & Goate, A. M. (2009). The utility of intraindividual variability in selective attention tasks as an early marker for Alzheimer's disease. *Neuropsychology*, *23*, 746–758.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Beverly Hills, CA: Sage.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374–380.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271–280.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906–911.

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and direction. *Review of Educational Research*, *77*, 334–372.

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

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*, 36–69.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, UK: Cambridge Univeristy Press.

Loaiza, V. M., Duperreault, K. A., Rhodes, M. G., & McCabe, D. P. (2015). Long-term semantic representations moderate the effect of attentional refreshing on episodic memory. Psychonomic Bulletin & Review.

Loaiza, V. M., & McCabe, D. P. (2012). Temporal contextual processing in working memory: Evidence from delayed cued recall and delayed free recall tests. *Memory & Cognition*, *40*, 191–203.

Loaiza, V., McCabe, D., Youngblood, J., Rose, N. S., & Myerson, J. (2011). The influence of levels of processing on recall from working memory and delayed recall tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1258–1263.

Loaiza, V. M., Rhodes, M. G., & Anglin, J. (2013). The influence of age-related differences in prior knowledge and attentional refreshing opportunities on episodic memory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *70*(5), 729–736.

Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology and Aging*, *26*, 661–670. April 4.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition*, *18*, 196–204.

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language*, *58*, 480–494.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174–179.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*, 530–542.

Moulin, C. J. A., Perfect, T. J., Akhtar, S., Williams, H. L., & Souchay, C. (2011). Judgements of learning and study time allocation: An illustration from neuropsychology. In J. Leboe & P. A. Higham (Eds.), *Constructions of remembering and metacognition* (pp. 167–181). Basingstoke: Palgrave Macmillan.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOL) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science*, *2*, 267–270.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*, 207–213.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu

Nelson . T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York, NY: Academic Press.

Rawson, K. A., O'Neil, R. L., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied*, *17*, 288–302.

Rhodes, M. G., & Tauber, S. K. (2010). Does the amount of material to be remembered influence judgments of learning (JOLs)? *Memory (Hove, England)*, *18*, 351–362.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*, 131.

Roediger, H. L., & Crowder, R. G. (1975). The spacing of lists in free recall. *Journal of Verbal Learning and Verbal Behavior*, *14*, 590–602.

Rose, N. S., Buchsbaum, B. R., & Craik, F. I. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition*, *42*, 689–700.

Rose, N. S., & Craik, F. I. M. (2012). A processing approach to the working memory/long-term memory distinction: Evidence from a levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1019–1029.

Rose, N. S., Myerson, J., Roediger, H. L., & Hale, S. (2010). Similarities and differences between working memory and long-term memory: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 471–483.

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning task. *Cognitive Development*, *15*, 115–135.

Sikstrom, S., & Johnsson, F. (2005). A model for stochastic drift in memory strength to account for judgments of learning. *Psychological Review*, *112*, 932–950.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 204–221.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, *33*, 1116–1129.

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general) and using gamma (in particular) to do so. *Handbook of Metamemory and Memory*, 95–114.

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology*, *62*, 254–263.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review*, *6*, 662–667.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*, 66–73.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language*, *28*, 127–154.

Unsworth, N. (2016). The many facets of individual differences in working memory capacity. In B. Ross (Ed.), *The psychology of learning and motivation*, *65*, 1–46.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*, 68–80.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory. *Cognitive Psychology*, *71*, 1–26.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychology Science Quarterly*, *4*, 388–402.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in Educational Theory and Practice*, *93*, 27–30.