Metacognition and Learning (2022) 17:375–398
https://doi.org/10.1007/s11409-021-09288-2

The underconfidence-with-practice effect in action memory: The contribution of retrieval practice to metacognitive monitoring

Veit Kubik[1] · Andreas Jemstedt[2] · Hassan Mahjub Eshratabadi[2] · Bennett L. Schwartz[3] · Fredrik U. Jönsson[2]

Received: 11 March 2020 / Accepted: 6 December 2021 / Published online: 20 January 2022

Abstract
When making memory predictions (judgments of learning; JOLs), people typically underestimate the recall gain across multiple study–test cycles, termed the underconfidence-with-practice (UWP) effect. This is usually studied with verbal materials, but little is known about how people repeatedly learn and monitor their own actions and to what extent retrieval practice via interim tests influence the progression of JOLs across cycles. Using action phrases (i.e., *squeeze the lemon*) as learning material, we demonstrated the UWP effect after both verbal and enactive encoding, although we did not get first-cycle overconfidence. As predicted, participants exhibited underconfidence in Cycles 2 and 3, as an error of calibrations. However, people's resolution of JOLs (i.e., ability to discriminate recalled from unrecalled items) increased across study–test cycles. Importantly, JOLs for study–test (relative to study–study) items increased faster across cycles suggesting that repeated study–test practice not only produces underconfidence across cycles, but also reduces underconfidence relative to study–study practice. We discuss these findings in terms of current explanations of the underconfidence-with-practice effect.

Keywords Underconfidence-with-practice effect · Judgments of learning · Study–test practice · Metacognitive monitoring · Enactment · Action phrases

✉ Veit Kubik
veit.kubik@uni-bielefeld.de

Andreas Jemstedt
andreas.jemstedt@psychology.su.se

Hassan Mahjub Eshratabadi
hassan.mahjub.eshratabadi@psychology.su.se

Bennett L. Schwartz
schwartb@fiu.edu

Fredrik U. Jönsson
fredrik.jonsson@psychology.su.se

[1] Department of Psychology, Bielefeld University, 33501 Bielefeld, Germany

[2] Stockholm University, S-106 91 Stockholm, Sweden

[3] Florida International University, Miami, FL 33199, USA

Metacognitive monitoring is central to human learning because it guides people's control of their study behavior (for reviews, see Bjork et al., 2013; Metcalfe & Finn, 2008; Rhodes, 2016; Rhodes & Castel, 2009; Schwartz & Jemstedt, 2021; Soderstrom & Bjork, 2014). However, miscalibrated metacognitive monitoring results in ineffective control (Kornell & Bjork, 2008; Soderstrom & Bjork, 2014). Thus, advancing our understanding of metacognitive monitoring and its role in control is important for both theory and application. In this paper, we direct our attention to a well-studied metacognitive judgment, judgments of learning (JOLs). Although people are often accurate in predicting subsequent memory performance (Dunlosky & Nelson, 1992; Koriat, 1997), they are prone to systematic biases such as the underconfidence-with-practice effect (UWP; Koriat et al., 2002; see also Ariel & Dunlosky, 2011; England & Serra, 2012; Finn, & Metcalfe, 2007, 2008; Hanczakowski et al., 2013; Koriat & Bjork, 2006; Rast & Zimprich, 2009; Scheck & Nelson, 2005; Tauber & Rhodes, 2012). The UWP effect refers to the observation that people tend to shift from overconfidence (mean JOL magnitude is higher than mean recall performance) to underconfidence (mean JOL magnitude is lower than mean recall performance) in subsequent learning cycles (e.g., Koriat et al., 2002). In addition to this shift in calibration between JOL magnitude and the overall recall level (i.e., calibration is also called absolute accuracy), the UWP is associated with an increase in resolution of JOLs with repeated study–test practice (Ariel & Dunlosky, 2011; Koriat, 1997; Koriat et al., 2002; Koriat & Bjork, 2006). Resolution (also known as relative accuracy) refers to people's ability to discriminate between the items that will and will not be recalled at a later occasion. Resolution is typically measured by the within-person JOL–recall gamma correlation (Nelson, 1984; but see Benjamin & Diaz, 2008).

In the present study, we pursued three main research aims. First, we examined whether the UWP effect generalizes to action phrases (i.e., phrases including one verb and one noun, such as *squeeze the lemon* or *break the pencil*) that are either verbally encoded or enactively encoded (i.e., by acting out the actions described in the to-be-remembered verb–noun phrases). Such knowledge is important because it contributes to our understanding of how learning can be effectively managed, especially because people typically monitor the learning of their own actions. Although the UWP effect is robust across experimental manipulations (for more details, see Finn & Metcalfe, 2007; Koriat et al., 2002), there is little research on how well we monitor learning of actions, in particular across multiple study and test phases. Our second and more important aim was to elucidate the role of retrieval versus study practice in the UWP effect, which also allowed us to evaluate predictions of the Memory-for-Past-Test account, the anchoring-and-adjustment-account, and the mnemonic-debiasing account. With our third aim, we examined the mnemonic benefits of and metacognitive sensitivity of enactive versus verbal encoding across three learning cycles.

Theoretical Accounts of the UWP Effect

Several accounts have been proposed to explain the basis of the UWP effect. According to the *mnemonic-debiasing account* (Koriat & Bjork 2006), the UWP effect occurs as a function of a foresight bias, which is alleviated by self-testing during practice. The *foresight bias* claims that people make inflated memory predictions regarding the recallability of information that is present during learning (e.g., both the cue and the target), but is absent during a subsequent test (Koriat & Bjork, 2005, 2006). Pertinent to this study,

this overconfidence is alleviated by repeated study–test practice, as an experience-based debiasing procedure (Koriat & Bjork, 2006). Study–test experience provides people with information about which items were recalled and which were not, and about item-specific retrieval fluency (Karpicke, 2009; Koriat & Bjork, 2006). These mnemonic cues are used to update people's memory predictions and, as a result, enhance both the relative and absolute accuracy of JOLs. That is, by and large, study–test experience sensitizes people to mnemonic cues that are diagnostic of future recall.

The *Memory for Past Test (MPT) account* proposes that JOLs result, in part, from memory for the last test (Finn & Metcalfe, 2007, 2008). People tend to give higher JOLs for previously-recalled than for previously-unrecalled items (e.g., Hanczakowski et al., 2013; Koriat et al., 2002). But, as some unrecalled items in Cycle 1 are newly learned in the subsequent cycle, underconfidence is produced. Learners will assign low JOLs to these previously unrecalled, but the newly learned items, for which recall performance will be 100%, will produce the underconfidence in Cycle 2. Furthermore, the MPT account explains the increase in resolution across the learning cycles (Finn & Metcalfe, 2007, 2008; see also Ariel & Dunlosky, 2011). In this view, this specific and salient cue of an item's past recall experience—the MPT heuristic—provides a reliable basis for predicting the future performance of this particular item relative to other items. Thus, based on this account JOLs' resolution increases across multiple learning cycles, even as the underconfidence emerges with repeated study–test practice.

Finally, the *anchoring-and-adjustment account* (England & Serra, 2012; Scheck & Nelson, 2005; Zhao & Linderholm, 2011) maintains that people set an anchoring point when making JOLs from which they insufficiently depart when monitoring their subsequent learning. That is, they do not adjust enough to account for new learning that occurs with repeated study. Numeric judgments tend to assimilate toward previously encountered numerical anchors (Yang, Potts, et al., 2017; Yang, Sun, et al., 2017). Initial overconfidence simply occurs when memory performance falls below the anchor point, which is typically set by experimenters' expectations at between 30%–50% (see Scheck & Nelson, 2005). Over time, underconfidence occurs because people inadequately adjust their JOLs upwards from their anchor during the progression of learning, which now exceeds that anchor. In the present study, we evaluated the findings in the context of these accounts.

The Relative Contribution of Retrieval versus Restudy Practice to the Underconfidence-With-Practice Effect

Retrieval practice has various beneficial effects on memory, specifically on long-term retention (Kubik et al., 2018, 2020; Roediger & Karpicke, 2006a; for comprehensive overviews, see McDermott, 2021; Roediger & Karpicke, 2006a; see also Kubik, Gaschler, et al., 2021). For current purposes, we focus on the indirect and metacognitive benefits of retrieval practice. The finding that taking an interim test enhances the efficiency of subsequent encoding of previously learned information is termed the indirect testing effect (i.e., *test-potentiated learning*; Izawa, 1966; see also Soderstrom & Bjork, 2014). For example, more items are newly retrieved from pre- to post-test when the number of tests prior to restudy is increased (Arnold & McDermott, 2013; see also Kubik et al., 2015; Tempel & Kubik, 2017). Beyond enhancing subsequent restudy of information, interim tests improve people's ability to accurately predict their own future learning (see Koriat et al., 2002). This may occur because retrieval practice exposes gaps in one's own knowledge and sensitizes

people to mnemonic cues—such as ease of learning and retrieval fluency—which are diagnostic for metacognitive monitoring and control of learning (Karpicke, 2009; Mitchum et al., 2016; Roediger & Karpicke, 2006b; Soderstrom & Bjork, 2014; Yang, Potts, et al., 2017; Yang, Sun, et al., 2017).

Given the current widespread interest in the benefits of interim tests, it is surprising that the relative contribution of study-versus-test experience in relation to the UWP effect has been investigated so few times (but see; England & Serra, 2012). Repeated study–test practice usually leads to a shift from over- to underconfidence with succeeding cycles, but it is less clear whether this bias derives from the study or test phases. Typically, repeated practice is instantiated as a repeated study–test condition (e.g., STSTST), but it is rarely compared to repeated restudy practice (e.g., SSSSSS) with regard to metacognitive judgments (but see Karpicke, 2009; Koriat & Bjork, 2006). In accordance with the mnemonic-debiasing account, interim tests—but not studying—provides information regarding the retrieval success and retrieval fluency of items. The test-related provision of and sensitization to these diagnostic mnemonic cues provide a basis for learners to update their memory predictions (Koriat & Bjork, 2006). Thus, rather than fostering underconfidence, test experience should reduce this metacognitive bias relative to repeated restudy, and therefore, test experience should lead to less underconfidence and increased resolution (see England & Serra, 2012). In contrast, following from the MPT account (Finn & Metcalfe, 2007), one may predict the underconfidence based on people's tendency to rely on MPT information across two cycles: that is, they continue to base their JOLs on the past test, when available, and not on new learning in the next study phase of the subsequent study–test cycle (cf. England & Serra, 2012). However, JOL magnitude should increase to the extent that past test performance also increased relative to an earlier study–test cycle. In this work, we aim to address the role of the relative contribution of study-versus-test experience to the progression of JOL magnitude across multiple learning cycles and test the predictions of the above mentioned accounts of the UWP effect.

Monitoring the Progression of the Enactment Effect across Multiple Learning Cycles

Memory is presumably biased, largely, to remember and monitor action-relevant information (Heuer et al., 2020). As such, a great deal of research has been devoted to memory for actions (Roediger & Zaromb, 2010). However, little is known about metacognition for the learning of action-related information as it progresses across several study–test cycles (but see Koriat et al., 2002). This knowledge is important for an understanding of effective management of learning, especially because of the claim that people sometimes have difficulties in monitoring their own actions.

Action memory research has largely focused on the *enactment effect*, the finding that motorically performed action phrases (i.e., enactive encoding) are better remembered later than the remembering that occurs after reading the same phrases (i.e., verbal encoding; for comprehensive reviews, see Engelkamp, 2001; Roediger & Zaromb, 2010; Steffens et al., 2015; for seminal studies, see Cohen, 1981; Engelkamp & Krumnacker, 1980; Saltz & Donnenwerth-Nolan, 1981). Researchers in this area widely agree that enactment promotes item-specific processing, that is, incidentally focusing attention on the individual features of the action phrase (Kubik, Obermeyer, et al., 2014; Li & Wang, 2018; Seiler & Engelkamp, 2003; Steffens et al., 2015), including the verb, the noun, and their association

(Koriat, 1995; Steffens et al., 2006, 2009). This contrasts with relational-processing which creates associations across items (e.g., linking *squeeze the lemon* with *pick up the fork*). This enhanced item-specific processing increases memory performance, leading to the enactment effect. However, there has been limited research that has focused on the learning of action phrases that occurs across multiple cycles of learning (Koriat & Pearlman-Avnion, 2003; Koriat et al., 1998; Kubik, Söderlund, et al., 2014).

More importantly, for the present study, the degree to which people accurately monitor their own actions has been examined in only a few studies, most of which employed a single study phase. According to this research, resolution (i.e., relative accuracy) of people's memory predictions is impaired by enactment (Cohen, 1983, 1988; Cohen et al., 1991; Koriat et al., 1991). These results, however, should be interpreted with caution because either no control conditions were provided (Cohen et al., 1991), or single words were used as control items that were presented for shorter durations than the enacted phrases.

With this background, it is of theoretical and practical importance to understand how people monitor the degree and progression of their own actions over several learning cycles. To our knowledge, only Koriat et al. (2002) has investigated the degree *and* progression of predicted and actual learning across multiple study–test cycles using learning material including action phrases in an enactment condition; the initial study phase of each cycle, participants had to learn 30 paired associates, consisting of a Tumai verb (an imaginary language, but in essence, a nonsense phrase) and a randomly paired Hebrew action phrase denoting its meaning. They were instructed to act the target action phrase (e.g., *smell the flower*) out and then say them aloud. Similarly, during the tests, participants were presented with Tumai verbs and were asked to recall their corresponding action phrases both by performing and saying them aloud. A clear UWP effect was demonstrated—that is, participants showed initial overconfidence in their ability to remember the action phrases, but became underconfident across subsequent study–test cycles. This research is an important starting point for more systematic investigations of how people monitor the degree and progression of their own actions over several learning occasions, using typical action phrases both comparing verbal and encoding conditions.

The Aims of the Study

First, our aim was to demonstrate the UWP effect with action phrases and to generalize it to performed actions. Based on all three theoretical accounts of the UWP effect, we expected to demonstrate a metacognitive bias in terms of the shift toward underconfidence as learning progresses for both verbal and enactive encoding. The MPT account specifically predicts that "forgotten but then recalled" items exhibit most prominently the underconfidence as these items are recalled but have low JOLs; recalled items on Cycle 1 should show no or less underconfidence; however, some underconfidence may occur because of JOLs' "downward variance from 100%" for recalled items (Finn & Metcalfe, 2007). In contrast, the anchoring-and-adjustment account explains underconfidence for both previously recalled and unrecalled items as JOLs are insufficiently adjusted from a low anchor to recall performance.

Furthermore, as a second characteristic of the UWP effect, we tested the prediction that resolution increases across cycles for verbal and enactive encoding. Based on the MPT and mnemonic debiasing accounts, we made this prediction on the assumption that past-test information or respectively diagnostic mnemonic cues in general accumulate with repeated

study–test practice with both cues being predictive of later recall performance. Furthermore, based on the MPT account, past-test information drives JOLs' magnitude. Consequently, past-correlations (i.e. the correlation of JOLs and recall performance in the previous learning cycle) on Learning Cycles 2 and 3 should be higher than resolution scores (i.e. the correlation of JOLs and recall performance in the current learning cycle), respectively. The anchoring-and-adjustment account does not make any predictions pertaining to resolution and past-test correlations.

Second, we sought to elucidate the role of tests in the UWP effect by comparing the effect of retrieval versus restudy experience on JOL magnitude. We tested the prediction that retrieval experience enhances JOLs' magnitude relative to restudy experience across learning cycles. This prediction is consistent with the mnemonic debiasing account assuming that retrieval experience, relative to restudy experience, provides more diagnostic mnemonic cues that more closely predict the increasing recall performance across cycles. However, the MPT account cannot explain the progress of JOL magnitude in study–study practice, as no past-test information is available. Furthermore, it predicts that the underconfidence largely stems from the "forgotten and then recalled" items, and not from previously recalled items. Based on the anchoring-and-adjustment account, we predict that the adjustment up from the initial anchor is larger for study–test than for study–study practice: with retrieval experience as a salient cue, learners may more likely overcome the stability bias in terms of discounting new learning (England & Serra, 2012). Consequently, retrieval experience should reduce rather than produce the underconfidence-with-practice effect relative to restudy experience—provided that study–study and study–test practice leads to similar levels of recall performance.

Third, we investigated actual and predicted cued-recall performance and memory predictions (JOLs) for enactive and verbal encoding across multiple learning cycles. Based on the notion that enactment elicits item-specific information, we predicted the enactment effect to remain stable across the entire learning phase. If learners use this item-specific information as a mnemonic cue for making their JOLs, they should also be sensitive to the mnemonic benefits of enactment across the learning session (see also Castel et al., 2013) Similar to prior research, we expected learners to have a generally impaired ability to monitor self-performed actions in terms of resolution (e.g., Cohen et al., 1991; Koriat et al., 1991).

Method

Participants

A sample size of 60 participants was pre-determined for this study, with $n = 30$ participants to be randomly assigned to both encoding groups. This sample-size estimation was based on prior research (Kubik, Söderlund, et al., 2014; Kubik et al., 2018, Exp. 3) without any a priori power calculation.

In actuality, we individually tested 61 Stockholm undergraduate students (*M* [*SD*] age, 24.34 [5.56]; 43 females), and data from 59 participants were included in the final analysis. Two participants were excluded for various reasons. One participant did not follow the instructions, and the data were not recorded for one participant due to a technical error. Participants from this convenience sample were all native Swedish speakers and participated voluntarily without compensation or in return for course credits or movie vouchers.

They were randomly assigned to the two groups of encoding type (enactive: $n = 30$; verbal: $n = 29$), with the restriction of obtaining a similar gender ratio (enactive: 22 females; verbal: 21 females). Participants did not vary in age as a function of encoding type (enactive: 24.23 [5.35] years; verbal: 24.45 [5.87] years), $W = 412.00$, $p = 0.733$, $r_{rb} = 0.05$.

Materials

Thirty-six Swedish action phrases (e.g., *squeeze the lemon*) were selected from the normative study of Molander and Arar (1998). The action phrases were two-to-four words long, were composed of one verb and one noun, and did not specify body parts as objects (e.g., *scratch the ear*).

Design

A mixed-factorial $2 \times 2 \times 3$ design was used, with *encoding type* (enactive vs. verbal) being a between-subjects variable and *study type* (study–study vs. study–test) and *cycle* (1, 2, vs. 3) being within-subject variables. The primary dependent measures were recall performance in the interim tests (measured as proportion correctly recalled targets) and JOL magnitude, collected in Cycles 1–3 of the learning session. JOL accuracy was assessed using both calibration and resolution measures.

Procedure

The experimental procedure consisted of a learning session that included three study–study or study–test cycles of learning action phrases (see Fig. 1). A final test session of verb-cued recall followed after 5 min and after 1 week. The results of the final test session are not reported here as they are not relevant to the present article on the UWP effect. This experimental procedure was run with E-Prime 2.0 professional software (Psychological Software Tools, Pittsburgh, PA; Schneider et al., 2002).

At the beginning of the learning session, we told participants that they would learn 36 action phrases in three cycles, each including two phases. Half of the action phrases were studied in both phases in each of the three cycles (i.e., SS SS SS) and are called *study–study items*; the other half were studied in the first phase and tested in the second phase in each of the three cycles (i.e., ST ST ST) and are called *study–test items*. Action phrases were presented for 7 s on the computer screen, during which time they were to be read or acted upon and were followed by a 1-s interstimulus interval. Depending on the encoding group, participants either read the action phases in the first phase of each cycle (i.e., *verbal encoding group*), or enacted the action phases (i.e., *enactive encoding group*). More specifically, participants in the enactive encoding group pantomimed the actions described in the to-be-remembered verb–noun phrases (i.e., motorically performed them) without any action-related object at hand. Experimenters monitored the learning to ensure that participants were in fact pantomiming the actions.

Importantly, during the initial study phase of each of the three cycles, item-based JOLs were collected, participants were asked how confident (0–100%) they were that they would remember the noun after several minutes if cued with the verb. During testing, the verb of a previously studied action phrase (e.g., *squeeze*) was displayed as a retrieval cue, one at

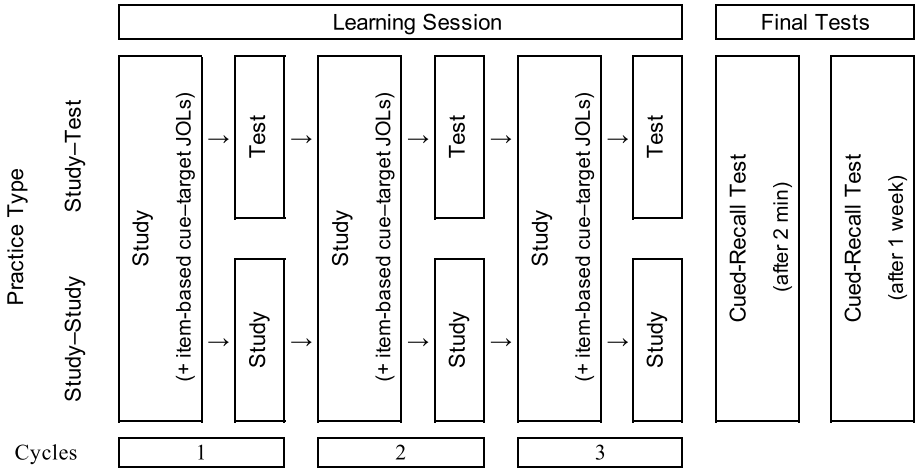

Fig. 1 The figure illustrates the experimental procedure of this study. Participants passed through a learning session, and received two additional final cued-recall tests, which were not considered in the present article on the Underconfidence-With-Practice effect. The critical learning session consisted of three learning cycles, each including an initial study phase with subsequent item-based, cue–target judgements of learning (i.e., JOLs) and an ensuing study versus interim test phase. Critically, half of the action phrases were studied in both phases in each of the three cycles (i.e., study–study items), and half of them were studied in the first phase and tested in the second phase in each of the three cycles (i.e., study–test items). Depending on the encoding type, participants either read the action phrases in the first phase of each cycle (i.e., *verbal encoding group*), or enacted the action phases (i.e., *enactive encoding group*)

a time, for 7 s, or until learners pressed the ENTER key to indicate that they remembered the respective target noun (e.g., *the lemon*). Learners were permitted a maximum of 10 s to type their responses on a computer keyboard. The items were randomly presented for each participant and for each study/test phase in each learning cycle. The study phases were separated from each other by a 30-s-long arithmetic filler task (i.e., evaluating the correctness of mathematical equations with varying difficulty; e.g., $16 \times 16 = 254$) to eliminate primary memory effects (Glanzer & Cunitz, 1966).

Scoring and Data Analyses

Participants' responses were scored as correct if the original noun was entered on the keyboard. Two measures of metacognitive accuracy were used. *Calibration* (i.e., absolute accuracy) of JOLs assesses how over- or underconfident learners are when predicting their own memory. It was calculated by subtracting recall performance (proportion correct) from mean item-based JOL for each participant in the study–test conditions. *Resolution* (i.e., relative accuracy) assesses the extent to which learners can discriminate between items that will or will not be recalled on a later retrieval occasion. It was calculated with the nonparametric Goodman–Kruskal within-participants gamma correlation between actual and predicted recall performance in the study–test conditions (Nelson, 1984). *Past-test correlations* were calculated as the nonparametric Goodman–Kruskal within-participants gamma correlations between actual recall performance in the prior learning cycle and the predicted recall performance in the current learning cycle in the study–test condition.

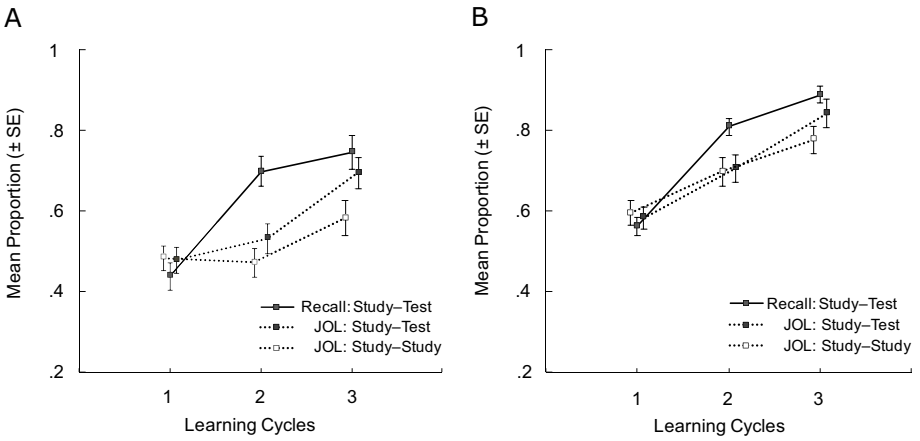

Fig. 2 Recall performance and JOL magnitude during the learning session for action phrases, as a function of cycle (1/2/3) and study type (study–study/study–test), separately shown for verbal encoding (Fig. 2A) and enactive encoding (Fig. 2B). Error bars represent standard errors (SEs) of the mean

An alpha-level of 0.05 was used. Adjusted, less-biased estimates for the population effect sizes were reported for analyses of variance (ANOVA; generalized omega squared [ω^2], Olejnik & Algina, 2003). In cases when the assumption of sphericity was violated, the reported numbers are calculated using the Huynh–Feldt correction. For planned comparisons between specific conditions or experimental groups, contrast analyses or Student *t*-tests were reported, using Cohen's *d*. If the assumptions of normality and/or homoscedasticity were violated, we reported equivalent nonparametric statistics, such as the Wilcoxon signed-ranks tests with the rank biserial correlation (r_{rb}). The dataset analyzed in the current study is publicly available at: https://osf.io/b4w26/.

Results

Figure 2 (Panel B) shows intermediate recall performance and JOL magnitude during the learning session for action phrases, as a function cycle and study type, separately for verbal encoding (Fig. 2A) and enactive encoding (Fig. 2B).

Recall Performance in Interim Tests

As predicted, enactive encoding topped verbal encoding in enhancing cued-recall performance. Learners recalled significantly more nouns after enactment than verbal encoding across all tests, $F(1, 57) = 12.66$, $p < 0.001$, $\varpi^2 = 0.091$. In addition, recall performance significantly increased with repeated study–test practice across cycles, $F(1.58, 89.86) = 130.01$, $p < 0.001$, $\omega^2 = 0.414$. These gains in memory performance did not differ between verbal and enactive encoding across the learning session, as indicated by the lack of a reliable Encoding Type × Cycle interaction, $F(1.58, 89.86) = 0.29$, $p = 0.695$, $\omega^2 < 0.001$.

Judgments of Learning

Magnitude JOL magnitude increased as learning progressed, as shown by a main effect of cycle, $F(1.38, 78.87) = 66.80$, $p < 0.001$, $\omega^2 = 0.153$. Overall, study–test items garnered higher JOLs than study–study items, $F(1, 57) = 30.79$, $p < 0.001$, $\omega^2 = 0.012$. Importantly, there was a significant Study Type × Cycle interaction, $F(1.49, 84.66) = 26.34$, $p < 0.001$, $\omega^2 = 0.013$, indicating that JOL magnitude did not differ in Cycle 1 between both study types, $t(58) = 0.92$, $p = 0.359$, $d = 0.12$, but then diverged in favor of study–test items across cycles. That is, JOL magnitude of study–test items significantly increased more than JOL magnitude of study–study items, both in Cycle 2, $t(58) = 4.40$, $p < 0.001$ (one-tailed), $d = 0.57$, and in Cycle 3, $t(58) = 6.40$, $p < 0.001$ (one-tailed), $d = 0.83$. Similarly, learners gave higher JOLs after enactive encoding than verbal encoding, $F(1, 57) = 13.08$, $p < 0.001$, $\omega^2 = 0.094$. Participants predicted action phrases to profit more from enactment with practice, as demonstrated by a significant Encoding Type × Cycle interaction, $F(1.38, 78.87) = 4.01$, $p = 0.036$, $\omega^2 = 0.008$. In addition, repeated study–test (compared to study–study) practice increased JOL magnitude more for verbal than for enactive encoding, as indicated by a significant Encoding Type × Study Type interaction, $F(1, 57) = 6.99$, $p = 0.011$, $\omega^2 = 0.002$. There was no significant Encoding Type × Study Type × Cycle interaction, $F(1.49, 84.66) = 1.64$, $p = 0.206$, $\omega^2 < 0.001$.

Calibration As also shown in Fig. 3A, calibration changed significantly (toward underconfidence) as learning progressed, $F(1.62, 92.13) = 16.16$, $p < 0.001$, $\omega^2 = 0.111$. JOLs exhibited only nonsignificant overestimation in Cycle 1 ($M = 0.03$, $SE = 0.02$), $t(58) = 1.45$, $p = 0.076$ (one-tailed), $d = 0.19$, but pronounced underconfidence in Cycle 2 ($M = –1.35$, $SE = 0.02$), $t(58) = 6.02$, $p < 0.001$ (one-tailed), $d = 0.78$, and in Cycle 3 ($M = –0.05$, $SE = 0.03$), $V = 488.00$, $p = 0.004$ (one-tailed), $r_{rb} = 0.41$. The disclosed results pattern related to the UWP effect did not vary as a function of encoding type. That is, action phrases were similarly calibrated for verbal and enactive encoding (verbal: $M = –0.06$, $SE = 0.03$; enactive: $M = –0.04$, $SE = 0.03$), $F(1, 57) = 0.22$, $p = 0.645$, $\omega^2 < 0.001$. This remained the case throughout the learning session, as reflected by the lack of a reliable Encoding Type × Cycle interaction, $F(1.62, 92.13) = 1.11$, $p = 0.323$, $\omega^2 = 0.001$.

The UWP effect is also characterized as a bias to underestimate recall gains across cycles with repeated practice (cf., Koriat & Bjork 2006). To evaluate the degree to which learners underestimate (denoted by negative scores) or overestimate (denoted by positive scores) the recall gains as learning progressed (cf. Serra & Dunlosky, 2005), we calculated calibration shift scores as follows: $\text{Shift Score}_{\text{C1-Cn}} = (\text{JOL}_{\text{C2}} - \text{Recall}_{\text{C2}}) - (\text{JOL}_{\text{C1}} - \text{Recall}_{\text{C1}})$. Shift scores amounted to –0.17 ($SE = 0.03$) between Cycles 1 and 2, to –0.08 ($SE = 0.03$) between Cycles 1 and 3, and to 0.09 ($SE = 0.03$) between Cycles 2 and 3. Thus, overall, calibration shifted significantly toward underconfidence, $F(1.65, 94.29) = 39.28$, $p < 0.001$, $\omega^2 = 0.178$, both from Cycle 1 to Cycle 2, $V = 167.00$, $p < 0.001$, $r_{rb} = 0.81$, and from Cycle 1 to Cycle 3, $V = 460.50$, $p = 0.001$, $r_{rb} = 0.48$ (see Fig. 3, Panel A). However, calibration shifted from Cycle 2 to Cycle 3 toward more accurate calibration, $V = 1370.00$, $p < 0.001$, $r_{rb} = 0.60$, reflecting the large preceding recall gain between Cycles 1 and 2. Again, encoding type did not significantly influence the size of the calibration shift, $F(1, 57) = 0.13$, $p = 0.722$, $\omega^2 < 0.001$. However, we note a trend that the change in the calibration shift across cycles was numerically reduced for enacted items, $F(1.65, 94.29) = 3.13$, $p = 0.058$, $\omega^2 = 0.012$.

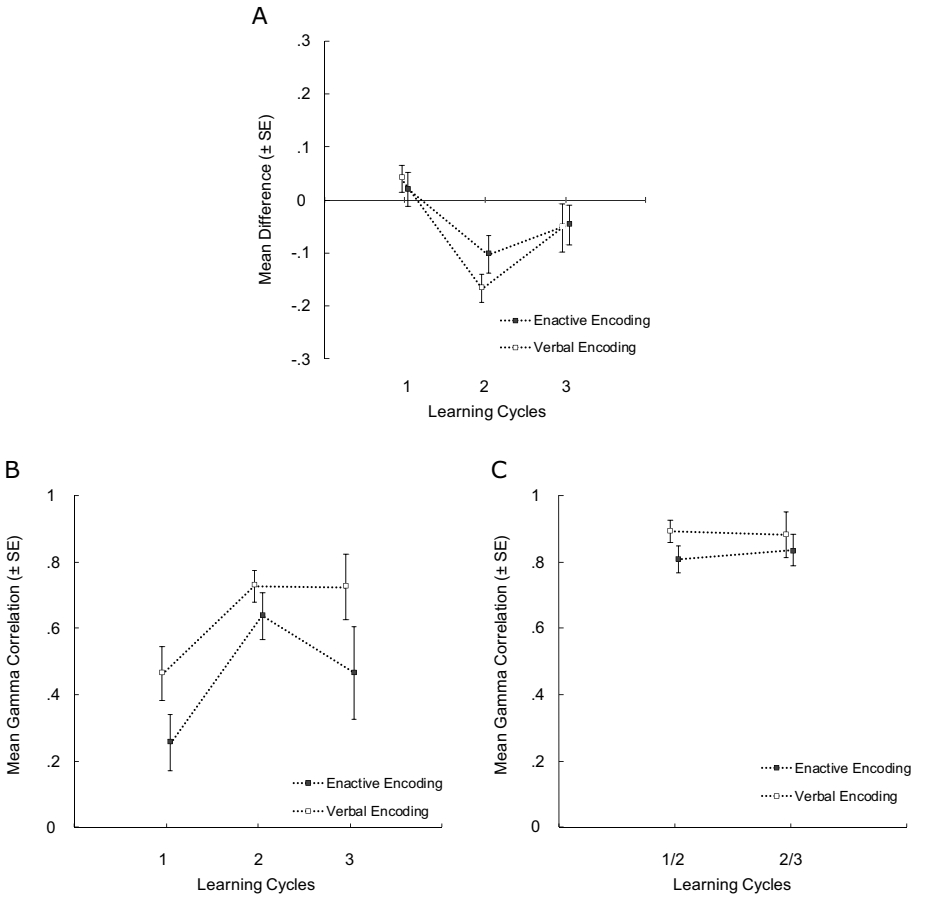

Fig. 3 Calibration (Panel A), resolution (Panel B), and past-test correlations (Panel C) for study–test items, as a function of cycle (1/2/3) and encoding type (verbal/enactive). Error bars represent standard errors (SEs) of the mean

Furthermore, we tested the relationship between shifts in calibration and recall gains between cycles. Results revealed a strong negative association between calibration shifts and recall gains from Cycles 1 and 2 ($r = –0.72$, $p < 0.001$) as well as Cycles 2 and 3 ($r = –0.86$, $p < 0.001$).

Anchoring To investigate the potential effects of anchoring and adjustment, we analyzed mean JOLs for Cycles 1 and 2 as a function of whether the items were correctly recalled or not during Cycle 1. We used JOLs for nonrecalled items in Cycle 1 as an estimate for the anchor that learners set prior to the experiment, likely indicating theory-based beliefs about the learning tasks. Table 1 illustrates mean JOLs, recall performance, and calibration as a function of study type and recall status (recalled vs. nonrecalled) on the preceding Cycle (n–1).

A 2 (recall $\text{status}_{\text{Cycle-1}}$: recalled vs. nonrecalled)×2 (cycle: Cycle-1 JOL vs. Cycle-2 JOL)×2 (encoding type: verbal vs. enactive) mixed ANOVA on JOL magnitude showed significant main effects of recall status (relatively higher JOLs for recalled

Table 1 Mean JOLs, Mean Recall, and Mean Calibration in Learning Cycles 1–3 for Study–Test Items as a Function of Recall Status (Recalled/Nonrecalled) on the preceding Cycle (n–1), separately shown for Verbal and Enactive Encoding

	Cycle-1 Recall Status		Cycle-2 Recall Status	
	Nonrecalled	Recalled	Nonrecalled	Recalled
Verbal Encoding				
Cycle-1 JOL	.43 (.18)	.56 (.19)		
Cycle-2 JOL	.38 (.19)	.76 (.19)	.36 (.20)	.62 (.17)
Cycle-2 Recall	.52 (.22)	.96 (.09)		
Cycle-2 Calibration	-.15 (.22)	-.20 (.19)		
Cycle-3 JOL			.42 (.25)	.84 (.17)
Cycle-3 Recall			.38 (.29)	.97 (.05)
Cycle-3 Calibration			.02 (.35)	-.13 (.18)
Enactive Encoding				
Cycle-1 JOL	.54 (.17)	.62 (.16)		
Cycle-2 JOL	.57 (.20)	.83 (.19)	.52 (.25)	.75 (.19)
Cycle-2 Recall	.63 (.21)	.97 (.05)		
Cycle-2 Calibration	-.06 (.28)	-.15 (.18)		
Cycle-3 JOL			.58 (.28)	.89 (.19)
Cycle-3 Recall			.63 (.31)	.96 (.08)
Cycle-3 Calibration			-.01 (.45)	-.06 (.20)

Note. The standard deviation (*SD*) is shown in parentheses

phrases), cycle (JOLs increased across cycles), and encoding type (relatively higher JOLs for enacted phrases), as well as several significant two-way interactions between the three factors ($ps < 0.001$). Specifically, JOLs were higher for items that were recalled on Cycle 1 compared to those which were not recalled, as indicated by a main effect of recall status, $F(1, 57) = 154.56$, $p < 0.001$, $\omega^2 = 0.300$, and JOLs increased from Cycle 1 to Cycle 2, as indicated by a main effect of cycle, $F(1, 57) = 28.07$, $p < 0.001$, $\omega^2 = 0.074$. Importantly, JOL magnitude increased from Cycle 1 to Cycle 2 for previously recalled items, whereas they decreased for nonrecalled items, as indicated by the Recall Status × Cycle Interaction, $F(1, 57) = 90.67$, $p < 0.001$, $\omega^2 = 0.107$. In addition, learners gave higher values to enacted action phrases than to read action phrases, as indicated by a main effect of encoding type, $F(1, 57) = 7.65$, $p = 0.008$, $\omega^2 = 0.054$. However, JOL magnitude increased from Cycle 1 to Cycle 2 to the same degree for enactive and verbal encoding, as indicated by a nonsignificant Encoding Type × Cycle interaction, $F(1, 57) = 1.32$, $p = 0.256$, $\omega^2 < 0.001$. Critically, a significant Encoding Type × Recall Status interaction was revealed, $F(1, 57) = 6.82$, $p = 0.012$, $\omega^2 = 0.016$, indicating that the predicted enactment effect was larger for nonrecalled items than recalled items in Cycle 1 (see Table 1). Thus, the encoding-related anchor difference was even larger than the JOL difference for recalled items on Cycle 1. No significant Encoding Type × Recall Status × Cycle interaction was observed, $F(1, 57) = 2.55$, $p = 0.116$, $\omega^2 = 0.002$.

We used the same analysis for the mean JOLs in Cycles 2 and 3 as a function of the Cycle-2 recall status. A 2 (recall $\text{status}_{\text{Cycle-2}}$: recalled vs. not recalled) × 2 (cycle: Cycle-2

JOL vs. Cycle-3 JOL)×2 (encoding type: verbal vs. enactive) mixed ANOVA showed significant main effects of recall status and cycle, as well as a significant interaction effect between both factors (*p*s < 0.001). JOLs magnitude were higher for items that were recalled than for items that were not recalled in Cycle 2, $F(1, 57) = 141.45$, $p < 0.001$, $\omega^2 = 0.358$. JOLs increased from Cycles 2 to 3, $F(1, 57) = 107.60$, $p < 0.001$, $\omega^2 = 0.104$, and importantly, they increased more steeply for recalled (relative to nonrecalled) items in Cycle 2, indicated by a Recall Status×Cycle Interaction, $F(1, 57) = 26.66$, $p < 0.001$, $\omega^2 = 0.026$. In addition, learners assigned significantly higher values for enacted compared than for read action phrases, $F(1, 57) = 7.22$, $p = 0.009$, $\omega^2 = 0.051$, and JOL magnitude increased similarly across cycles for both encoding types, $F(1, 57) = 2.42$, $p = 0.125$, $\omega^2 = 0.002$. There was no significant Encoding Type×Recall Status interaction, $F(1, 57) = 2.08$, $p = 0.154$, $\omega^2 = 0.004$, indicating that the predicted enactment effect did not reliably differ between recalled and nonrecalled items in Cycle 2 (see Table 1). No significant Encoding Type×Recall Status×Cycle interaction was observed, $F(1, 57) = 3.02$, $p = 0.087$, $\omega^2 = 0.002$.

Resolution We calculated resolution for the study–test items as the mean within-participant gamma correlation between JOL magnitude of a given study phase and recall performance of the ensuing test phase within the same learning cycle. Figure 3 (Panel B) illustrates the resolution scores as a function of cycle and encoding type. Note that gamma correlations could not be calculated for several participants ($n = 2$: enactive encoding in Cycle 2; $n = 8$: enactive encoding in Cycle 3; $n = 2$ in verbal encoding in Cycle 3) because their recall and/or JOLs reached an average proportion of 1 or zero. As indicated by a main effect of cycle, $F(1.76, 82.86) = 8.50$, $p < 0.001$, $\omega^2 = 0.078$, learners improved their resolution scores from Cycle 1 ($M = 0.37$, $SE = 0.06$) to Cycle 2 ($M = 0.69$, $SE = 0.06$), $t(94) = 3.99$, $p < 0.001$, but not from Cycle 2 to Cycle 3 ($M = 0.60$, $SE = 0.06$), $t(94) = 1.08$, $p = 0.285$. This change in resolution across the learning session was similar for both encoding types, indicated by the lack of a reliable Encoding Type×Cycle interaction, $F(1.76, 82.86) = 0.58$, $p = 0.543$, $\omega^2 < 0.001$ (see Fig. 3, Panel C). As indicated by a significant main effect, enactment ($M = 0.46$, $SE = 0.06$) significantly hampered the resolution of JOLs compared to verbal encoding ($M = 0.65$, $SE = 0.06$), $F(1, 47) = 4.68$, $p = 0.036$, $\omega^2 = 0.037$.

Past-test correlations We calculated past-test correlations as the mean within-participant gamma correlation between JOL magnitude of a given study phase and recall performance of the *preceding* test phase. Figure 3 (Panel C) illustrates the past-test correlations as a function of cycle and encoding type. Note that past-test correlations could not be calculated for several participants ($n = 3$: enactive encoding in Cycle 3) because their recall and/or JOLs reached an average proportion of 1 or zero. A 2 (measure: resolution vs. past-test correlation)×2 (cycle: Cycle-2 JOLs vs. Cycle-3 JOLs)×2 (encoding type: enactive vs. verbal) mixed ANOVA revealed that past-test correlations ($M = 0.85$, $SE = 0.04$) were significantly higher than the corresponding resolution scores ($M = 0.65$, $SE = 0.04$), $F(1, 47) = 20.02$, $p < 0.001$, $\omega^2 = 0.124$, and that both measures were significantly lower for enactive than for verbal encoding, $F(1, 47) = 4.77$, $p = 0.034$, $\omega^2 = 0.038$. No other main effect or interactions were reliable (*p*s > 0.155; see Fig. 3, Panels B and C).

General Discussion

The present study had three major aims. First, we made predictions from competing theories and then tested the generality of the UWP effect by using verbally and enactively encoded action phrases. Second, we explored the specific role that retrieval practice versus restudy practice plays in the progression of JOLs' magnitude suggesting that repeated study–test practice not only produces underconfidence across cycles, but also reduces underconfidence relative to study–study practice. Third, we examined the mnemonic benefits of and metacognitive sensitivity to enactive versus verbal encoding across three learning cycles. We predicted the enactment effect to occur across learning cycles, and that people are sensitive to this recall benefit in terms of JOL magnitude, but we also predicted that enactment decreases JOLs' resolution.

The Underconfidence-With-Practice-Effect in Memory for Actions

The UWP effect refers to the shift from over- to underconfidence across learning cycles. This pattern was only partially shown using simple action phrases, that is, we showed underconfidence in Cycles 2 and 3. More specifically, participants shifted their memory predictions toward underconfidence, from both Cycle 1 to Cycle 2, and from Cycle 1 to Cycle 3 when action phrases were verbally and enactively encoded during the study phases. Thus, the tendency to underestimate memory performance with practice generalizes to action events, replicating prior research using paired associates and action phrases (Koriat et al., 2002), single words (e.g., Koriat et al., 2002) as well as word pairs (e.g., Finn & Metcalfe, 2007, 2008; Hanczakowski et al., 2013, Exp. 1; Serra & Dunlosky, 2005; Tauber & Rhodes, 2012). Thus, this generalizability suggests that the UWP is a general feature of learning. The shift toward underconfidence is consistent with both the MPT account, embarking on learners' tendency to underestimate the amount of new learning (Finn & Metcalfe, 2007, 2008), and the anchoring-and-adjustment account, referring to the incomplete adjustment up from a prior psychological anchor as learning progresses (England & Serra, 2012; Scheck & Nelson, 2005). The mnemonic debiasing account (Koriat & Bjork, 2005, 2006) cannot account for the current data because it cannot explain why underconfidence on the second cycle occurs. However, the mnemonic debiasing account can explain how the provision of mnemonic cues in terms of recall success or ease can enhance JOL calibration toward underconfidence and thereby decrease the overconfidence in Cycle 1.

Additional analyses revealed that JOLs shifted toward more accurate calibration from Cycle 2 to Cycle 3. This result pattern was equivalent for verbal and enactive encoding of action phrases. This result is consistent with other findings of the UWP literature, which also show little or no underconfidence in later cycles (e.g. Hanczakowski et al., 2013, Exp. 4). However, such a finding is rather inconsistent with the majority of research reporting that the magnitude of underconfidence stays the same from Cycle 2 onwards (see also Hanczakowski et al., 2013; Exp. 1; Koriat et al., 2002). This shift toward more accurate calibration from Cycle 2 to Cycle 3 cannot easily be accommodated with the MPT account because past-test information of study–test items should not reduce underconfidence but actually produce underconfidence (see also England & Serra, 2012). However, as a post-hoc explanation, we argue that the recall gain between Cycles 2 and 3 was relatively small compared to the large recall gain from Cycle 1 to Cycle 2, and that JOLs reflect in terms of memory for past information. This assumption is supported by strong negative associations

between calibration shifts and recall gains between Cycles 1 and 2 as well as Cycles 2 and 3. The mnemonic-debiasing account and anchoring-and-adjustment account can explain the general pattern of the UWP effect leading to JOLs' underconfidence as learning progresses, but not the specific finding of a shift toward more accurate calibration. Future studies should systematically compare the UWP effect as a function of learning materials, number of cycles, and, importantly, item characteristics.

Notably, some of the present results departed from typical pattern seen in the UWP effect. Only a slight and nonsignificant overconfidence in Cycle 1 was exhibited with action-phrase materials for both verbal and enactive encoding. This finding was not predicted and is somewhat surprising, but a lack of initial overconfidence in the UWP pattern has been reported in other studies (e.g., Hanczakowski et al., 2013, Exp. 1; Koriat, 1997; Exp. 1; Koriat et al., 2002). Nonetheless, UWP studies typically revealed JOLs' overconfidence in Cycle 1 (e.g., Koriat et al., 2002). Many factors may contribute to this finding. One factor may be that participants set a more accurate anchor point, possibly because of the increased familiarity of action-related concepts in everyday life relative to paired associates. As a result, the typical overconfidence bias in Cycle 1 was reduced. Another factor relates to the idea that item characteristics such as the backward association strength from the target word (e.g., *cheese*) toward the cue word (e.g., *gouda*) moderates the occurrence of the overconfidence effect. Prior research with paired associates (e.g., *gouda–cheese*) showed that a rather low backward association strength leads to a well-calibrated JOL–recall correspondence (Koriat & Bjork, 2005, 2006), similar to the present results with action phrases. However, high backward association strength leads to inflated JOL and an illusion of knowledge during study. In JOLs, in which both the cue and target being are available, participants discount that only the cue will be available at the time of the test. Thus, at the time of test, the backward association is absent and cannot support the learner (Koriat & Bjork, 2005, 2006; for a similar explanation for the overconfidence in retrospective judgements, see Juslin et al., 2000). In the current study, the selected set of action phrases in this study may have on average a low backward association strength from the noun target (e.g., *lemon*) to the verb cue (e.g., *squeeze*) because the target noun is also associated with many alternative actions (e.g., *to eat, to smell, to pickle, to suck, to slice, to peel a lemon*). Future studies can explore this possibility by assessing the verb–noun association strength and examining the size of the overconfidence as a function of it. Thus, the occurrence of overconfidence during any test may hinge on numerous methodological factors, whereas the shift pattern of JOLs toward underconfidence is a more pervasive and critical feature of the UWP effect (cf. Koriat et al., 2002).

A second feature of the UWP effect is the increased resolution with repeated study–test practice (Koriat et al., 2002; see also Ariel & Dunlosky, 2011). In our study, we found that JOLs' resolution increased across cycles, specifically between Cycles 1 and 2, and the past-test correlations were higher than the corresponding resolution scores for both encoding types. Both findings accord with the MPT account that the past-test information for each item partially drives the resolution increases of JOLs (see Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007, 2008). Similarly, the results pattern is consistent with the mnemonic debiasing account; it assumes that study–test practice provides generally more diagnostic mnemonic cues on an item-by-item basis, such as retrieval experience and fluency, and in turn increases the participants' resolution scores (Koriat & Bjork 2006). In contrast, the anchor-and-adjustment account does not address these findings.

The Relative Contribution of Retrieval versus Restudy Practice to the Underconfidence-With-Practice Effect

Several findings related to retrieval practice contribute to the understanding of the UWP effect. Typically, researchers report that repeated study–test practice leads to underconfidence as learning progresses, but few studies have compared the interim-test condition to repeated study–study practice (but see England & Serra, 2012; Karpicke, 2009). In our study, we demonstrated that JOL magnitude for action phrases did not increase between Cycles 1 and 2 with repeated study–study practice, which is consistent with prior work (England & Serra, 2012; Karpicke, 2009). Furthermore, JOL magnitude accelerated faster across cycles when interim tests were provided, that is, JOLs had practically the same magnitude in Cycle 1, but significantly increased for study–test (relative to study–study) items in Cycles 2 and 3. This suggests that study–study items would produce even more underconfidence, provided that they lead to a similar recall performance as study–test items, as previously shown for action phrases (Kubik et al., 2015, 2016) and also shown in the final immediate test of the present study ($ps > 0.10$; reported in Kubik, Soderstrom, et al., 2021). Together, these results are consonant with prior findings (see England & Serra, 2012; Karpicke, 2009).

From the perspective of the *mnemonic-debiasing account*, interim test experience provides mnemonic cues, such as retrieval success or retrieval fluency, and sensitizes learners to them across the learning session. These cues are more diagnostic of recall. As a result, study–test items should better reflect the recall gains across the learning session, and thereby increase faster than study–study items. More specifically, people continue with repeated study–study practice to base JOLs on internal cues (e.g., perceived intrinsic difficulty of items), whereas with study–test practice they shift successively to the more diagnostic mnemonic cue of encoding fluency (e.g., measured by self-paced study time; Karpicke, 2009; Koriat, 1997; Koriat & Bjork 2006) and retrieval fluency (e.g., Koriat & Ma'ayan, 2005) as learning progresses (Koriat & Bjork, 2006).

The anchoring-and-adjustment and the MPT accounts do not provide any specific mechanism to accommodate the finding that JOLs increase more moderately across cycles for repeated study–study compared to study–test items (cf. England & Serra, 2012). Nonetheless, the *anchoring-and-adjustment account* is relevant in that we argue that participants adjust their JOLs more effectively from the anchor point based on the salient cue of retrieval experience. That is, participants match the item-specific JOLs to their previous recall performance. However, as study–study items lack this test experience, participants make rather small adjustments (England & Serra, 2012).

The *MPT account* does not explain underconfidence for repeated study–study practice because learners cannot use the memory of a past test as a heuristic to determine their JOLs. Furthermore, past-test information was primarily hypothesized to produce underconfidence in immediate JOLs, rather than reducing it. Against the predictions of the MPT account, the data suggest that past-test information is not critical for this feature of the UWP effect (Koriat et al., 2002) or even detrimental to it (cf. England & Serra, 2012).

Another relevant finding to the theoretical discussion of the MPT account is that we observed that, within the set of study–test items, an increased underconfidence bias for previously recalled relative to nonrecalled items, and in Cycle 3 no underconfidence at all for nonrecalled items on Cycle 2, supporting previous research (England & Serra, 2012). Taken together with prior findings of a similarly sized UWP effect for previously recalled and unrecalled items (Koriat et al., 2002), this pattern of results is inconsistent

with the notion that items' past-test information accounts for the UWP effect. Although some level of underconfidence for recalled items may result from JOLs' downward variance from 100% (Finn & Metcalfe, 2007), the MPT account has difficulties explaining the present finding of increased underconfidence for recalled items. Based on the latter, the opposite results pattern is predicted such that "forgotten and then recalled items" should disproportionally contribute to the underconfidence, whereas JOLs for previously recalled items should be more accurate, as has been previously reported in a single study (Finn & Metcalfe, 2007). These observed UWP findings can be better explained by the anchoring-and-adjustment account, such that people insufficiently adjust from a psychological anchor point to match recall performance, and this matching occurs independently of items' prior recall status (cf. England & Serra, 2012).

However, because we did not experimentally manipulate the anchor point, the evidence for the anchoring-and-adjustment account in this study is correlational. Nonetheless, previous research has successfully manipulated participants' initial anchor or in general their metacognitive judgements without affecting recall performance by providing different cover stories (e.g., the task is easy vs. difficult; England & Serra, 2012), or by framing JOLs in terms of forgetting versus remembering information over a delay (England et al., 2017). These manipulations likely change the usage (or selection) of cues or the scale of JOLs (i.e., how confidence is translated into a numeric JOL response). There is further experimental evidence for anchoring and adjustment effects in research on decision making, how participants utilize experimenter- versus self-generated anchor points as the starting point from which they adjust their judgements (see Epley & Gilovich, 2001; Tversky & Kahneman, 1974).

Taken together, the UWP effect cannot be explained by any single account. Based on the data, it is likely that the UWP arises from a complex pattern in terms of JOL magnitudes, calibration, calibration shifts as well as resolution scores across learning cycles. Thus, the UWP effect is likely determined by several underlying mechanisms that act in concert.

Monitoring the Progression of the Enactment Effect across Multiple Learning Cycles

As predicted, enacting action phrases during encoding led to better recall performance than verbal encoding in enhancing cued-recall performance. More importantly, we showed that the enactment effect remained constant throughout the learning session. This finding is consistent with the notion that motor enactment relies on item-specific processing of the whole action phrase, in particular, on the binding of verb and noun within action phrases (see Kormi-Nouri, 1995; Kormi-Nouri & Nilsson, 2001). This finding accords with previous findings showing that when using free recall, which research shows is dependent on item-relational processing, the mnemonic effects of enactment are less reliable (Earles & Kersten, 2000; Knopf, 1995; Steffens, 1999) and even decrease further across multiple study–test cycles (Koriat & Pearlman-Avnion, 2003). Successful free-recall performance requires not only item-specific, but also item-relational information that is less strongly fostered by enactment (e.g., Steffens et al., 2006, 2009). Notably, we observed smaller error bars for enacted compared to nonenacted phrases, likely indicating that enactment is also be the result of the specific study instruction that we gave participants, whereas verbal encoding allows for a greater variety of learning strategies being selected between and within participants.

Despite the remarkable body of action memory research, there has not been any systematic evaluation of how accurately learners monitor their learning and memory for actions.

To this end, the present results make a novel contribution, showing that participants are sensitive to the memory-enhancing effects of enactment and that participants can predict the enactment effect to enlarge with repeated study–test practice. The finding that JOLs reflect the enactment effect is quite remarkable because extrinsic cues, which include encoding manipulations, are generally discounted in JOLs, especially when manipulated between-subjects (Koriat, 1997). Typically, JOLs reflect encoding effects only in within-subject designs (e.g., interactive imagery, Begg et al., 1989; levels-of-processing, Shaw & Craik, 1989; verbal production, Castel et al., 2013). Thus, the present results add to previous research by suggesting that even in between-subject designs, JOLs can be sensitive to different encoding manipulations (e.g., generated vs. read words: Begg et al., 1991; Mazzoni & Nelson, 1995). Thus, these results contribute to the literature as prior studies did not systematically investigate to what extent learners are sensitive to the enactment effect. When inspecting the descriptive results of prior studies exploratively, learners seemed to provide at least numerically higher JOLs for enacted phrases, particularly when encoding type is manipulated within-subject (Cohen, 1988, Experiment 2), but also when manipulated between subjects (Cohen, 1988, Experiment 1; Cohen et al., 1991). Note, however, that some of these studies had methodological shortcomings. For example, enacted action phrases were often compared to words that are simply read (Cohen et al. 1991), and they were in part presented for shorter amounts of time (Cohen, 1988; Cohen et al., 1991). Presumably, enactment as a specific encoding type (i.e., motoric performance) draws one's attention to the items' individual characteristics and increases their distinctiveness, which enhances recall for these enacted items. Thus, despite manipulating encoding type between participants, learners acknowledged its memorial benefits, as they probably based JOLs partially on item-specific information as mnemonic cues being diagnostic of future recall (for a similar suggestion, see Castel et al., 2013, explaining the metacognitive sensitivity of the production effect, i.e., the mnemonic benefit of saying words aloud versus silent encoding).

Furthermore, there has been little systematic research on anchoring effects in action memory until our study. The present study reveals that the predicted enactment effect in Cycle 1, as reflected in JOL magnitude, was even larger for items that were unrecalled. This finding suggests that learners have a theory-based belief that enactment is a more powerful encoding technique than verbal encoding. However, it will require further research to examine how beliefs are utilized when participant solicit JOLs (cf. Mueller et al., 2013), and how they interact with experience-based cues (cf. England et al., 2017; Serra & England, 2012). As a consequence, in the current study, JOLs reflected the mnemonic benefit of enactment and thereby calibration scores between recall and JOLs did not differ between the two encoding types. In contrast to JOL magnitudes and calibration scores, the present results showed that enactive encoding hampered JOLs' resolution across cycles, compared to verbal encoding (e.g., Cohen, 1983, 1988; Cohen et al. 1991; Cohen et al., 1991). They extend prior results showing that resolution for future free-recall performance is fairly accurate for words, but not for enacted action phrases. We speculate that if all enacted action phrases become highly distinctive as a result of such item-specific processing then resolution, which relies on cross-item comparisons, might suffer. For verbally-encoded action phrases, on the other hand, the amount of distinctiveness between action phrases varies more and thus the distinctiveness of one action phrase relative to another should be more salient, providing learners with more diagnostic information upon which their JOLs can be based. Thus, although

enactment bolsters memory performance, it is associated with poor metacognitive accuracy in terms of resolution.

Practical Applications

These data have implications for practical aspects of learning. There has been considerable research that documents how important retrieval practice is for efficiently learning in educational settings (cf., Roediger & Karpicke, 2006b; see also Kubik, Gaschler, et al., 2021). Equally important for learning in school and elsewhere is being able to use metacognitive resources to study the items that require further study (see e.g., Efklides, 2014; Zhao & Linderholm, 2011). This study adds something new. When people use retrieval practice across study cycles, not only do they boost the efficiency of their learning, but it also enhances the magnitude of JOL over the progression of learning and thereby reduces the amount of underconfidence that occurs. With less underconfidence, one implication is that efficient learners can direct their attention to new learning, rather than repeat items that they have already mastered. Accurate self-evaluations are critical for effective self-regulated learning (cf. Dunlosky & Rawson, 2012; Dunlosky & Thiede, 2013; see also Roelle et al., 2017), and this is true regardless of whether the learning is verbal or motoric. This metacognitive benefit of retrieval practice may be useful in learning physical trades or skills, multimedia learning (see Eitel, 2016) as well as more academic learning in general.

Concluding Comments

Study strategies, such as retrieval practice, restudy practice, as well as enactment, not only enhance memory performance but also affect memory predictions across repeated learning occasions. This has implications for the theoretical understanding of how to effectively regulate one's own learning, but also educational implications. It is upon such metacognitive monitoring that people base their decisions to continue or stop studying and upon which items to focus (Nelson & Leonesio, 1988). Any systematic dissociation between subjective and objective learning curves can be detrimental to learning because ineffective study strategies or study time allocation may result. Future research is encouraged to investigate over- and underconfidence in self-evaluations and its impact on self-regulated learning and academic achievement in educational real-world contexts (such as example-based and multimedia learning scenarios; cf. Eitel 2016; Roelle et al., 2017).

Contributions

Veit Kubik was the main contributor to the study conception and study design. Material preparation, data collection and analysis were performed by Veit Kubik. The first draft of the manuscript was written by Veit Kubik, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they do not have any conflict of interest.

Ethics Approval The study was carried out in accordance with the recommendations of the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct. All subjects gave written informed consent in accordance with the Declaration of Helsinki (World Medical Association 2013) before participating in the study, with the understanding that they could quit at any time. The Regional Ethics Review Board, Stockholm (http:/www.epn.se) concluded that there are no ethical concerns regarding the proposed experiments on the testing effect, including the current study, needed to be further reviewed.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*, 171–184. https://doi.org/10.3758/s13421-010-0002-y

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940–945. https://doi.org/10.1037/a0029199

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*, 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition, 19*(5), 487–497. https://doi.org/10.3758/BF03199571

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Memory and Metamemory* (pp. 73–94). Psychology Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41*, 28–35. https://doi.org/10.3758/s13421-012-0249-6

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology, 22*, 267–281. https://doi.org/10.1111/j.1467-9450.1981.tb00402.x

Cohen, R. L. (1983). The effect of encoding variables on the free recall of words and action events. *Memory & Cognition, 11*, 575–582. https://doi.org/10.3758/BF03198282

Cohen, R. L. (1988). Metamemory for words and enacted instructions: Predicting which items will be recalled. *Memory & Cognition, 16*, 452–460. https://doi.org/10.3758/BF03214226

Cohen, R. L. (1989). Memory for action events: The power of enactment. *Educational Psychology Review, 1*, 57–80. https://doi.org/10.1007/BF01326550

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology, 45*, 523–538. https://doi.org/10.1037/h0084303

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. https://doi.org/10.3758/BF03210921

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*, 58–61. https://doi.org/10.1016/j.learninstruc.2012.05.002

Earles, J. L., & Kersten, A. W. (2000). Adult age differences in memory for verbs and nouns. *Aging, Neuroscience, and Cognition, 7*, 130–139.

Efklides, A. (2014). How does metacognition contribute to the regulation of learning? An integretative approach. *Psychological Topics, 23*, 1–30.

Eitel, A. (2016). How repeated studying and testing affects multimedia learning: Evidence for adaptation to task demands. *Learning and Instruction, 41*, 70–84. https://doi.org/10.1016/j.learninstruc.2015.10.003

Engelkamp, J., & Krumnacker, H. (1980). Image- and motor-processes in the retention of verbal materials. *Zeitschrift Für Experimentelle Und Angewandte Psychologie, 27*(4), 511–533.

Engelkamp, J., & Dehn, D. M. (2000). Item and order information in subject performed tasks and experimenter-performed tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 671–682. https://doi.org/10.1037/0278-7393.26.3.671

Engelkamp, J. (2001). Action memory: A system-oriented approach. In H. D. Zimmer, R. Cohen, M. Guynn, J. Engelkamp, R. Kormi-Nouri, & M. N. Foley (Eds.), *Memory for Action: A Distinct Form of Episodic Memory* (pp. 49–96). Oxford University Press.

England, B. D., Ortegren, F. R., & Serra, M. J. (2017). Framing affects scale usage for judgments of learning, not confidence in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(12), 1898–1908. https://doi.org/10.1037/xlm0000420

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review, 19*, 715–722. https://doi.org/10.3758/s13423-012-0237-7

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*, 391–396. https://doi.org/10.1111/1467-9280.00372

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36*, 813–821. https://doi.org/10.3758/MC.36.4.813

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 238–244. https://doi.org/10.1037/0278-7393.33.1.238

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19–34.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–552. https://doi.org/10.1037/0033-295X.98.4.506

Glanzer, M., & Cunitz, R. A. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*, 351–360. https://doi.org/10.1016/S0022-5371(66)80044-0

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language, 69*(3), 429–444. https://doi.org/10.1016/j.jml.2013.05.003

Heuer, A., Ohl, S. & Rolfs, M. (2020). Memory for action: A functional view of selection in visual working memory. *Visual Cognition, 28*(5–8), 388–400. https://doi.org/10.1080/13506285.2020.1764156

Izawa, C. (1966). Reinforcement-test sequences in paired-associated learning. *Psychological Reports, 18*(3), 879–919. https://doi.org/10.2466/pr0.1966.18.3.879

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied, 22*(3), 305–318. https://doi.org/10.1037/a0019902.supp

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review, 107*, 384–396.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486. https://doi.org/10.1037/a0017341

Knopf, M. (1995). Memory for action events: Structure and development in adulthood. In F. E. Weinert & W. Schneider (Eds.), *Memory Performance and Competencies Issues in Growth and Development* (pp. 127–138). Mahwah (NJ): Erlbaum.

Koriat, A., Ben-Zur, H., & Druch, A. (1991). The contextualization of memory for input and output events. *Psychological Research Psychologische Forschung, 53*, 260–270. https://doi.org/10.1007/BF00941396

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311–333. https://doi.org/10.1037/0096-3445.124.3.311

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. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., Pearlman-Avnion, S., & Ben Zur, H. (1998). The subjective organization of input and output events in memory. *Psychological Research Psychologische Forschung, 61*, 295–307. https://doi.org/10.1007/s004260050034

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162. https://doi.org/10.1037/0096-3445.131.2.147

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy: Insights from the processes underlying judgments of learning in children. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, Function, and Use* (pp. 1–17). Kluwer.

Koriat, A., & Pearlman-Avnion, S. (2003). Memory organization of action events and its relationship to memory performance. *Journal of Experimental Psychology: General, 132*, 435–454. https://doi.org/10.1037/0096-3445.132.3.435

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187–194. https://doi.org/10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval fluency. *Memory & Cognition, 34*, 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478–492. https://doi.org/10.1016/j.jml.2005.01.001

Kormi-Nouri, R. (1995). The nature of memory for action events: An episodic integration view. *European Journal of Cognitive Psychology, 7*(4), 337–363. https://doi.org/10.1080/09541449508403103

Kormi-Nouri, R., & Nilsson, L. G. (2001). The motor component is not crucial! In H.D. Zimmer, R. L. Cohen, M. J. Guynn, J. Engelkamp, R. Kormi-Nouri, & M. A. Foley (Eds.), *Memory for action: A distinct from episodic memory*? (pp. 97–111). Oxford: Oxford University Press.

Kornell, N., & Bjork, R. A. (2008). Optimizing self-regulated study: The benefits and costs of dropping flashcards. *Memory, 16*, 125–136.

Kornell, N., & Rhodes, M. G. (2013). Feedback reduces the metacognitive benefit of tests. *Journal of Experimental Psychology: Applied, 19*, 1–13. https://doi.org/10.1037/a0032147

Kubik, V., Gaschler, R., & Hausman, H. (2021a). Enhancing student learning in research and educational practice: The power of retrieval practice and feedback. *Psychology Learning & Teaching, 20*(1), 1–20. https://doi.org/10.1177/1475725720976462

Kubik, V. Soderstrom, N., Jemstedt, J. K., & Jönsson, F. U. (2021b). *Metacognition in memory for actions: Predicting the mnemonic effects of enactment and testing*. Manuscript in preparation.

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting testing into action. Enacted retrieval practice benefits long-term retention more than covert retrieval retention. *Quarterly Journal of Experimental Psychology, 73*(12), 2093–2105. https://doi.org/10.1177/1747021820945560

Kubik, V., Jönsson, F. U., Knopf, M., & Mack, W. (2018). The direct testing effect is pervasive in action memory: Analyses of recall accuracy and recall speed. *Frontiers in Psychology., 9*, 1632. https://doi.org/10.3389/fpsyg.2018.01632

Kubik, V., Nilsson, L.-G., Olofsson, J. K., & Jönsson, F. U. (2015). Testing effects on subsequent restudy and forgetting of action phrases. *Scandinavian Journal of Psychology, 56*(5), 475–481. https://doi.org/10.1111/sjop.12238

Kubik, V., Olofsson, J. K., Nilsson, L.-G., & Jönsson, F. U. (2016). Putting action memory to the test: Testing affects subsequent restudy but not long-term forgetting of action events. *Journal of Cognitive Psychology, 28*(2), 209–219. https://doi.org/10.1080/20445911.2015.1111378

Kubik, V., Obermeyer, S., Meier, J., & Knopf, M. (2014a). The enactment effect in a multi-trial free-recall paradigm. *Journal of Cognitive Psychology, 26*, 781–787. https://doi.org/10.1080/20445911.2014.959018

Kubik, V., Söderlund, H., Nilsson, L.-G., & Jönsson, F. U. (2014b). Individual and combined effects of enactment and testing on memory for action phrases. *Experimental Psychology, 61*, 347–355. https://doi.org/10.1027/1618-3169/a000254

Li, G., & Wang, L. (2018). The role of item-specific information for the retrieval awareness of performed actions. *Frontiers in Psychology, 9*, 1325. https://doi.org/10.3389/fpsyg.2018.01325
Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1263–1274. https://doi.org/10.1037/0278-7393.21.5.1263
McDermott, K. B. (2021) Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*, 609–633. https://doi.org/10.1146/annurev-psych-010419-051019
Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174–179. https://doi.org/10.3758/pbr.15.1.174
Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200–219. https://doi.org/10.1037/a0039923
Molander, B., & Arar, L. J. (1998). Norms for 439 action events: Familiarity, emotionality, motor activity, and memorability. *Scandinavian Journal of Psychology, 39*, 275–300. https://doi.org/10.1111/1467-9450.00087
Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*, 378–384. https://doi.org/10.3758/s13423-012-0343-6
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133. https://doi.org/10.1037/0033-2909.95.1.109
Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 676–686. https://doi.org/10.1037/0278-7393.14.4.676
Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*(4), 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x
Nilsson, L., Nyberg, L., Kormi-Nouri, R., & Rönnlund, M. (1995). Dissociative effects of elaboration on memory of enacted and non-enacted events: A case of negative effect. *Scandinavian Journal of Psychology, 36*(2), 225–231. https://doi.org/10.1111/j.1467-9450.1995.tb00981.x
Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434
Rast, P., & Zimprich, D. (2009). Age differences in the underconfidence-with-practice effect. *Experimental Aging Research, 35*, 400–431. https://doi.org/10.1080/03610730903175782
Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. New York: Oxford University Press.
Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550–554. https://doi.org/10.3758/PBR.16.3.550
Roediger III, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/2Fj.1467-9280.2006.01693.x
Roediger III, H. L.& Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/2Fj.1745-6916.2006.00012.x
Roediger III, H. L. & Zaromb, F. M. (2010). Memory for actions: How different? In L. Bäckman & L. Nyberg (Eds.), *Aging, Memory, and the Brain: Essays in Honor of Lars-Göran Nilsson* (pp. 24–52). Psychology Press.
Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology, 109*, 99–117. https://doi.org/10.1037/edu0000132
Saltz, E., & Donnenwerth-Nolan, S. (1981). Does motoric imagery facilitate memory for sentences? A selective interference test. *Journal of Verbal Learning and Verbal Behavior, 20*, 322–332. https://doi.org/10.1016/S0022-5371(81)90472-2
Schacter, D. L., & Szpunar, K. K. (2015). Enhancing attention and memory during video-recorded lectures. *Scholarship of Teaching and Learning in Psychology, 1*(1), 60–71. https://doi.org/10.1037/stl0000011
Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with- practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*(1), 124–128. https://doi.org/10.1037/0096-3445.134.1.124

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. Psychology Software Tools Inc.

Schwartz, B. L., & Jemstedt, A. (2021). The role of fluency and dysfluency in metacognitive experiences. In P. Metallidou & D. Moraitou (Eds.), *Trends and Prospects in Metacognition Research across the Lifespan – A Tribute to Anastasia Efklides* (pp. 25–40). Springer.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1258–1266. https://doi.org/10.1037/0278-7393.31.6.1258

Serra, M.J., & England, B.D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *Quarterly Journal of Experimental Psychology, 65*, 2231–57. https://doi.org/10.1080/17470218.2012.685081

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4*(2), 131–135. https://doi.org/10.1037/0882-7974.4.2.131

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99–115.

Steffens, M. C. (1999). The role of relational processing in memory for actions: A negative enactment effect in free recall. *The Quarterly Journal of Experimental Psychology, 52*(4), 877–903. https://doi.org/10.1080/713755860

Steffens, M. C., Jelenec, P., Mecklenbräuker, S., & Thompson, E. M. (2006). Decomposing retrieval and integration in memory for actions: A multinomial modelling approach. *The Quarterly Journal of Experimental Psychology, 59*(3), 557–576. https://doi.org/10.1080/02724980443000764

Steffens, M. C., Jelenec, P., & Mecklenbräuker, S. (2009). Decomposing the memory processes contributing to enactment effects by multinomial modeling. *European Journal of Cognitive Psychology, 21*(1), 61–83. https://doi.org/10.1080/09541440701868668

Steffens, M. C., von Stülpnagel, R., & Schult, J. C. (2015). Memory recall after "learning by doing" and "learning by viewing": Boundary conditions of an enactment benefit. *Frontiers in Psychology, 6*, 1907. https://doi.org/10.3389/fpsyg.2015.01907

Szpunar, K. K., McDermott, K. B., & Roediger III, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1392–1399.

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27*(2), 474–483. https://doi.org/10.1037/a0025246

Tempel, T., & Kubik, V. (2017). Test-potentiated learning of motor sequences. *Memory, 25*(3), 326–334. https://doi.org/10.1080/09658211.2016.1171880

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

World Medical Association. (2013). World Medical Association Declaration of Helsinki ethical principles for medical research involving human subjects. *JAMA: Journal of the American Medical Association, 310*(20), 2191–2194.https://doi.org/10.1001/jama.2013.281053

Yang, C., Potts, R., & Shanks, D. R. (2017a). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*(3), 263–277. https://doi.org/10.1037/xap0000122

Yang, C., Sun, B., & Shanks, D. R. (2017b). The anchoring effect in metamemory monitoring. *Memory & Cognition, 46*(3), 384–397. https://doi.org/10.3758/s13421-017-0772-6

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence, not subjective probability. *Memory & Cognition, 43*, 1168–1179.

Zawadzka, K., & Higham, P. A. (2016). Recalibration effects in judgments of learning: A signal detection analysis. *Journal of Memory and Language, 90*, 161–176. https://doi.org/10.1016/j.jml.2016.04.005

Zhao, Q., & Linderholm, T. (2011). Anchoring effects on prospective and retrospective metacomprehension judgments as a function of peer performance information. *Metacognition and Learning, 6*, 25–43. https://doi.org/10.1007/s11409-010-9065-1