Judgments of Learning Index Relative Confidence Not Subjective Probability

Katarzyna Zawadzka and Philip A. Higham

University of Southampton

Author Notes

Katarzyna Zawadzka and Philip A. Higham, Psychology, University of Southampton, UK.

Katarzyna Zawadzka is now at School of Psychology, Cardiff University.

We thank Maciej Hanczakowski for his helpful comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Katarzyna Zawadzka, School of Psychology, Cardiff University, Cardiff CF10 3AT, UK. Email: ZawadzkaK@cardiff.ac.uk

Article published in Memory & Cognition.
The final publication is available at Springer via http://dx.doi.org/10.3758/s13421-015-0532-4

Abstract

The underconfidence-with-practice (UWP) effect is a common finding in calibration studies concerned with judgments of learning (JOLs) elicited on a percentage scale. The UWP pattern is present when, in a procedure consisting of multiple study-test cycles, mean scale JOLs underestimate mean recall performance on cycle 2 and beyond. Although this pattern is present both for items recalled and unrecalled on the preceding cycle, to date research has concentrated mostly on the sources of UWP for the latter type of items. The present study aimed at bridging this gap. In three experiments, we examined calibration on the third of three cycles. The results of Experiment 1 demonstrated the typical pattern of higher recall and scale JOLs for previously recalled items compared to unrecalled ones. More important, they also revealed that even though the UWP effect was found for both items previously recalled once and twice, its magnitude was greater for the former class of items. Experiments 2 and 3, which employed a binary betting task and a binary 0/100% JOL task, respectively, demonstrated that people can accurately predict future recall for previously recalled items with binary decisions. In both experiments, the UWP effect was absent both for items recalled once and twice. We suggest that the sensitivity of scale JOLs, but not binary judgments, to the number of previous recall successes strengthens the claim of Hanczakowski, Zawadzka, Pasek, and Higham (2013) that scale JOLs reflect confidence in, rather than the subjective probability of, future recall.

Judgments of Learning Index Relative Confidence Not Subjective Probability

Metacognitive theorists use a variety of different judgments to investigate how people assess their own memory processes. One common one is the judgment of learning (JOL) for which people assess their future memory performance. In a typical experiment employing JOLs, participants study a list of single words or word pairs. After the presentation of each item, a prompt appears instructing participants to rate the likelihood of future recall of that item on a scale from 0% to 100% - the JOL.[1] After the study phase, a recall test for the whole list follows. By comparing JOLs to recall performance, two measures can be calculated. First*, resolution* is the degree to which JOLs distinguish between items that will and will not be recalled at test. In order for resolution to be maximized, later recalled versus later unrecalled items should be assigned high versus low JOLs, respectively. Second, *calibration* is the difference between mean JOLs and mean recall performance. If the two measures are equal, assessments of future recall are said to be *realistic*. Mean JOLs lower versus higher than recall performance indicate *underconfidence* versus *overconfidence*, respectively.

Although most JOL studies reveal overconfidence (e.g., see Koriat, 2012), there are exceptions. For example, the underconfidence-with-practice (UWP) effect (e.g., Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002) is a common finding in JOL research involving repeated study and recall of the same list over at least two cycles. In most UWP studies, recall performance increases with each additional study-test cycle, as

[1] Although the probabilistic 0-100% scale is most commonly used for eliciting predictions of future memory performance, alternative judgment formats (e.g., Hanczakowski, Zawadzka, Pasek, & Higham, 2013; McCabe & Soderstrom, 2011; McGillivray & Castel, 2011) and framings of the JOL prompt (e.g., Finn, 2008; Serra & England, 2012) have been employed by researchers as well.

does resolution. However, although JOLs are typically similar to recall on the first cycle, they do not increase as much as recall on subsequent cycles, causing calibration to worsen with practice - the UWP effect.

One explanation of the UWP pattern is based on people's memory for past test performance (e.g., Finn & Metcalfe, 2007, 2008; Tauber & Rhodes, 2012; for alternative accounts of UWP see, e.g., England & Serra, 2012; Koriat, 1997; Scheck & Nelson, 2005). According to the memory-for-past-test (MPT) account of the UWP effect, after cycle 1, people base their immediate JOLs on their performance on the last test. Previously recalled items tend to get high JOLs, as their future recall seems very likely. Conversely, previously unrecalled items are assigned low JOLs, as people remember their failed recall attempt. What people fail to appreciate, though, is that additional learning occurs between the two tests. This additional learning means that some of these previously unrecalled items are recalled on a subsequent test, increasing the discrepancy between mean JOLs and mean recall performance, thus producing UWP.

The MPT account localizes the UWP effect mostly in unduly low judgments assigned to previously unrecalled items. However, Koriat et al. (2002), Finn and Metcalfe (2007), and Hanczakowski et al. (2013) reported the presence of the UWP pattern for previously recalled items as well. Finn and Metcalfe argued that these items may contribute to the UWP effect because of variability present in JOLs. Subsequent recall of items that were successfully recalled on a previous cycle(s) is typically excellent and they attract very high JOLs. However, because the JOL scale ends at 100%, any variability is necessarily downward, resulting in mean JOLs that underestimate mean recall performance (i.e., underconfidence). But what produces this downward variability? Finn and Metcalfe remain agnostic of its source. One

option is that it may be simply random, not stemming systematically from any characteristics of the rated items. For example, people may be reluctant to use the 100% rating too often, therefore assigning lower ratings to some items even though they believe that they are extremely likely to be later recalled. It is equally plausible, however, that the JOL variance for recalled items depends on item-specific information (see Dougherty, Scheck, Nelson, & Narens, 2005, for similar considerations regarding random and systematic sources of variability in delayed scale JOLs).

If item-specific information is responsible for the downward variability present in scale JOLs, then it is important to consider what cues might influence the assignment of these judgments. One type of information participants may rely on when predicting future recall of the previously recalled items is how many times these items were successfully retrieved. If JOLs are collected in a procedure employing at least three study-test cycles, then it stands to reason that items that were successfully retrieved twice, on both cycles 1 and 2, will receive higher JOLs than items that were successfully retrieved once, on either cycle 1 or cycle 2. As Vaughn and Rawson (2011) have shown, the number of successful retrievals in a learning-to-criterion procedure determines the magnitude of JOLs. Crucially, it seems also likely that such a systematic source of variance in JOLs for already recalled items will contribute to the pattern of UWP. As Finn and Metcalfe (2007) noted, any variance in JOLs for recalled items in the multi-cycle procedure used to investigate JOLs is likely to worsen calibration by removing JOLs further from the ceiling level of recall performance. Thus, if participants incorporate the information about the number of previous successful retrieval attempts into their JOLs, then items previously recalled once only should exhibit worse calibration than items previously recalled twice,

exaggerating UWP. The aim of the present study was to test this hypothesis by comparing the way in which people assign scale and binary ratings to previously recalled items in the UWP paradigm.

Experiment 1

In Experiment 1, we employed a common UWP procedure involving three study-test cycles. This methodology allowed us to investigate cycle-3 0-100% JOLs for items previously recalled once and twice. Previous research has shown that even a single successful recall attempt makes future recall success extremely likely (e.g., Koriat et al., 2002); therefore recall performance for items previously recalled once and twice was expected to be comparable and near ceiling. The question remained whether scale JOLs would be influenced by the number of past recall successes and how any such influence would affect the UWP pattern.

Method

Participants. Twenty-seven students participated in this study for course credit.

Materials and procedure. Sixty pairs of unrelated words were created from a set of 120 English nouns of medium frequency, ranging from four to eight letters in length. The same set of pairs was used for all study-test cycles. All pairs were randomly ordered anew for each participant on each study and test phase.

During the study phase, participants were presented individually with all pairs and instructed to memorize them for a future test. Each pair was presented on a computer screen for 1.5 seconds. After the presentation of each pair, the target was replaced by a prompt instructing participants to judge the likelihood of recalling the target from that pair at test when presented with the cue only. Participants were allowed to type in any value between 0% and 100%.

At test, participants were presented with one cue at a time and asked to type in

the target that accompanied this cue during the study phase. If they could not recall the target, they were instructed to press "Continue" to skip to the next cue. To ascertain that people had access to the information about the number of successful recall attempts for each pair, an additional task was implemented at Test 3. When presented with the cue, participants were asked to recall not only the target, but also the number of successful recall attempts for that target on the preceding cycles. The options presented to participants were: 2 (on both Test 1 and Test 2), 1 (on either Test 1 or Test 2), and 0 (on neither Test 1 nor Test 2). This judgment was made for all pairs, independently of whether they were recalled at Test 3 or not.

Before subjecting the results to analysis, participants' recall scores were checked manually. Responses were scored as correct whenever the stem of the word typed in by a participant matched the stem of the target (e.g., *silent* was considered correct if *silence* was the target). Misspelled words (e.g., *slience*) were also counted as correct responses.

Results and Discussion

Insert Table 1 about here

The means for JOLs, recall, and resolution (A_g, a nonparametric measure of resolution calculated from confidence judgments; e.g., Pollack, Norman, & Galanter, 1964) are presented in Table 1.[2] Even though this study was not concerned with

[2] In this context, A_g is an estimate of the area under the metacognitive (type-2) receiver operating characteristic (ROC) curve (a plot of the metacognitive hit rate as a function of the metacognitive false alarm rate). For metacognitive ROCs, the hit rates versus false alarm rates are the proportion of correct versus incorrect answers assigned a given level of confidence or higher, respectively. A_g computed with the trapezoidal rule is a superior measure of resolution compared to the more common Goodman-Kruskal gamma co-efficient computed with concordant and discordant pairs. See Higham, Zawadzka, and Hanczakowski (in press) for more detail.

resolution of metacognitive judgments, we provide resolution analyses for comparison with results of other studies present in the literature. Resolution scores were influenced by cycle, as evidenced by a one-way repeated-measures Analysis of Variance (ANOVA), $F(2, 52) = 53.912$, $MSE = .006$, $p < .001$, $\eta_p^2 = .675$. Resolution increased from cycle 1 to cycle 2, $t(26) = 7.622$, $SE = .018$, $p < .001$, $d = 1.47$, and from cycle 2 to cycle 3, $t(26) = 3.680$, $SE = .021$, $p = .001$, $d = 0.71$.

A 2 (measure: JOL, recall) x 3 (cycle: 1, 2, 3) repeated-measures ANOVA revealed a significant main effect of cycle, $F(2, 52) = 75.206$, $MSE = .014$, $p < .001$, $\eta_p^2 = .743$, showing that, in general, both JOLs and recall performance increased from cycle to cycle. The interaction was also significant, $F(2, 52) = 62.918$, $MSE = .007$, $p < .001$, $\eta_p^2 = .708$. Whereas on cycle 1, participants' mean JOLs were higher than their recall performance, $t(26) = 5.069$, $SE = .042$, $p < .001$, $d = 0.98$, this pattern was reversed on cycles 2 and 3, $t(26) = 3.131$, $SE = .039$, $p = .004$, $d = 0.61$ and $t(26) = 2.647$, $SE = .033$, $p = .014$, $d = 0.48$, respectively, revealing the UWP pattern. The main effect of measure was not significant, $F < 1$.

Insert Figure 1 about here

Mean recall and mean JOLs for items recalled never, once, and twice, are presented in the top panel of Figure 1. Also shown in the figure are analogous means for assessed (rather than actual) recall performance. On average, at Test 3, participants were able to correctly recall the number of previous successful recall attempts in 90% of cases. The mean gamma correlation computed for each participant between assessed and actual number of successful recall attempts was .986, confirming that participants were highly accurate in their assessments.

The crucial analyses of calibration on cycle 3 were conducted on the data set based on participants' actual past recall performance.[3] For previously unrecalled items, participants revealed good calibration, $t(26) = 1.191$, $SEM = .050$, $p = .245$, $d = 0.23$, while for items previously recalled once and twice, 0-100% JOLs underestimated recall performance, $t(26) = 6.102$, $SEM = .034$, $p < .001$, $d = 1.17$, and $t(26) = 3.128$, $SEM = .031$, $p = .004$, $d = 0.60$, respectively. Crucially, the magnitude of the UWP effect was much greater for items previously recalled once (21%) than twice (10%), $t(26) = 6.351$, $SEM = .017$, $p < .001$, $d = 1.22$. As seen in the top panel of Figure 1, this difference in calibration was caused mostly by lower 0-100% JOLs assigned to items previously recalled once than twice. Overall, this pattern indicates that systematic variance in JOLs based on the number of times items were previously recalled determines the magnitude of the UWP effect.[4]

In Experiment 1, we replicated the common finding that 0-100% JOLs underestimate recall performance from cycle 2 onward - the UWP effect. We also demonstrated that the UWP pattern for previously recalled items can be found irrespective of the number of past recall successes. The finding that the UWP pattern can be found for once-recalled items has been previously reported in experiments consisting of two study-test cycles (e.g., England & Serra, 2012; Koriat et al., 2002) as well as on the second of three cycles in Experiments 1 and 4 of Hanczakowski et al. (2013). More importantly, we also demonstrated that including all items that were

[3] For completeness, the same analyses were performed on the data set based on participants' assessments of past recall performance. The pattern of results was identical, with good calibration for previously unrecalled items, and the UWP pattern for items previously recalled once and twice.

[4] The lack of UWP for previously unrecalled items found in this experiment, which is not consistent with the predictions of the MPT account, was unexpected, especially since the same difference was significant in Experiment 1 of Hanczakowski et al. (2013) which used the same number of cycles, presentation times, list length and instructions. We suspect that the difference between recall and JOLs for previously unrecalled items may be less robust in three-cycle procedures, as compared to more commonly used two-cycle procedures, as the pool of items not recalled on any of the preceding cycles is limited, making the measures less stable.

previously recalled at least once in the analysis masks an important source of variance in both JOLs and the pattern of UWP. By examining separately items previously recalled once versus twice in the three-cycle procedure, we demonstrated that the variance in JOLs for already recalled items is not random. Furthermore, we were able to show that this variance strongly contributes to the UWP pattern, which is exaggerated for items that were previously recalled once only.

The traditional way of interpreting the pattern of results obtained in Experiment 1 would be to argue that participants are underconfident in their assessments of future recall of previously successfully retrieved items, particularly so if these items were retrieved only once on the preceding cycles. In other words, scale JOLs would reflect a pattern of inaccurate likelihood assessments concerning future recall. However, a different perspective is offered by a recent study by Hanczakowski et al. (2013). In this study, it was argued that computing calibration scores from JOLs made on a 0-100% scale may not be sufficient to draw conclusions regarding psychological under- or overconfidence. Hanczakowski et al. argued that scale JOLs likely do not reflect participants' assessments of the *likelihood* of future recall, but rather their *confidence* in future recall.

The difference between these two approaches to JOLs can be described using the pattern from Experiment 1. Recall for previously recalled items on cycle 3 is uniformly high and very close to ceiling. Yet, at the same time, participants clearly differentiate between these items, which drives lower 0-100% JOLs for once- as compared to twice-recalled items. Under the traditional likelihood interpretation, such a pattern demonstrates that participants do not believe that nearly all of previously once recalled items will be retrieved again on cycle 3. Under the confidence interpretation offered by Hanczakowski et al. (2013), such a pattern is equally

consistent with participants believing that nearly all previously recalled items will be retrieved again on cycle 3. What drives 0-100% JOLs assigned to once-recalled items down under this interpretation is an explicit demand posed by the scale JOL task – namely, to differentiate between various items at study. Even though participants may be convinced that items that were already recalled a short while ago will be recalled again on an immediate test, such as those commonly used for investigating UWP, they nevertheless use the scale provided by the experimenter to express their varying level of confidence in future recall.

Hanczakowski et al. (2013) proposed that one way to differentiate between likelihood and confidence interpretations is to use binary decisions, such as decisions to bet (or refrain from betting) that an item will be recalled, to assess participants' perceived likelihood of future recall.[5] If participants truly underestimate future recall of a subset of items, then the proportion of bets for these items should be lower than their mean recall performance. If, on the other hand, the pattern of underconfidence in scales does not reflect participants' estimations of the probability of future recall but rather varying levels of confidence in future recall, then proportions of bets should match the subsequent proportion of recalled items.

In Experiment 2 we used the logic outlined above to again examine calibration on cycle 3 for items previously recalled once or twice. In this experiment, we employed the betting task proposed by Hanczakowski et al. (2013). The crucial

[5] Hanczakowski et al. (2013) demonstrated that the binary-betting task produces the same results as the binary ("yes/no") JOL task, without suffering from a potentially serious drawback that characterizes the latter task. In the binary JOL task, people may assign different subjective values to two types of incorrect answers: metacognitive misses (correctly recalling an item assigned a "no" JOL) and metacognitive false alarms (failing to recall an item assigned a "yes" JOL), introducing a source of potential bias to the measure. In the betting task, on the other hand, penalties and rewards for different types of answers are objectively defined by the experimenter and equated, minimizing the possibility of biased responding.

questions were whether the UWP pattern would disappear in the betting task, replicating Hanczakowski et al., and whether any differences in calibration would be observed between items previously recalled once or twice.

Experiment 2

Method

Participants. Twenty-two students participated in this study for course credit.

Materials and procedure. The materials and procedure were the same as in Experiment 1 with one exception: instead of providing scale JOLs, participants were given a binary-betting task. Good calibration on the binary-betting task would be obtained if the proportion of bets equaled the proportion of recalled items. For each pair in the betting task, participants were asked whether they would like to bet they would later recall the target from that pair when presented with the cue. They were instructed that for correct bets they would gain a point, whereas for incorrect bets they would lose a point. If they refrained from betting, no points would be gained or lost. Participants were not shown their point count during the experiment.

Results and Discussion

The means for the proportion of bets, recall performance, and resolution (*d'*) are presented in Table 1.[6] Again, resolution was influenced by cycle, $F(2, 42) = 20.939$, $MSE = .284$, $p < .001$, $\eta_p^2 = .499$, and increased from cycle 1 to cycle 2, $t(21) = 4.418$, $SEM = .167$, $p < .001$, $d = 0.74$, and from cycle 2 to cycle 3, $t(21) = 2.123$, $SEM = .126$, $p = .046$, $d = 0.27$.

[6] *d'* is computed from the metacognitive hit rate and false alarm rate. In this context, these rates are defined as the proportion of correct versus incorrect answers that are bet upon. Technically, it is the standardized difference between the means of the correct and incorrect response distributions and, hence, provides a measure of resolution. See Higham et al. (in press) for discussion.

A 2 (measure: proportion of bets, recall performance) x 3 (cycle: 1, 2, 3) repeated-measures ANOVA revealed a significant main effect of cycle, $F(2, 42) = 123.270$, $MSE = .010$, $p < .001$, $\eta_p^2 = .854$, indicating an increase from cycle to cycle both in the proportion of bets and recall performance. The interaction was also significant, $F(2, 42) = 22.888$, $MSE = .004$, $p < .001$, $\eta_p^2 = .522$. Whereas on cycle 1, participants bet on a greater proportion of items than they later recalled, $t(21) = 4.084$, $SEM = .033$, $p = .001$, $d = 0.87$, there was no difference between these measures on cycles 2 and 3, $t(21) = 1.239$, $SEM = .033$, $p = .23$, $d = 0.26$ and $t < 1$, respectively. The main effect of measure was not significant, $F(2, 42) = 1.153$, $MSE = .0122$, $p = .29$, $\eta_p^2 = .052$. These results replicate the findings of Hanczakowski et al. (2013), again demonstrating that participants can accurately track future recall performance with their betting decisions.

Participants' assessments of the number of successful recall attempts were as good as they were with scale-JOLs (see the bottom panel of Figure 1). At Test 3, in 91% of cases participants were able to correctly assess the number of previously successful recall attempts for each pair, producing gamma = .977.

Analyses of calibration performed separately on each level of past recall performance (never, once, and twice) revealed that recall performance did not differ from the proportion of bets for any kind of items, all ts < 1.[7] Of particular interest is the fact that there was no UWP whatsoever for items previously recalled once and twice when the betting task was used. As a direct consequence of this pattern, calibration in the present experiment, calculated as the difference between the proportion of bets and mean recall performance, was equally good for items

[7] The same pattern of results was obtained when the analyses were performed on the data set based on participants' assessments of their past recall performance.

previously recalled once and twice, $t < 1$. This contrasts starkly with the results of Experiment 1 where calibration was markedly worse for items previously recalled once versus twice.

In total, the results revealed that, on cycle 3, people were well calibrated in the betting task independent of past recall performance. This suggests that participants were able to predict the very high probability of successful recall for items previously recalled at least once, as well as the lack of difference in future recall performance for these items, and executed their betting decisions accordingly. In other words, we argue that the results of Experiment 2 confirm that participants are well aware that previously recalled items are very likely to be recalled on a subsequent immediate test. This, in turn, indicates that the UWP pattern observed in Experiment 1 for items previously recalled (which was exaggerated for once-recalled items), should not be interpreted in terms of improper likelihood assessments. The UWP pattern revealed by scale JOLs in Experiment 1 speaks to the fact that participants are more confident that items previously recalled twice rather than once will be again recalled. The results of Experiment 2 demonstrate that participants also believe that both types of items will be probably be successfully retrieved.

The discrepancy between scale JOLs and binary betting decisions found in this study is similar to the pattern observed in Hanczakowski et al. (2013). In their study, UWP was present for scale JOLs, but not for binary bets, and the same pattern was observed here. Experiment 2 further demonstrated that cycle-3 calibration of betting decisions is very good not only overall, but also for each type of items considered in the three-cycle procedure: items recalled never, once and twice. This experiment, however, does not specifically determine why binary bets differ from scale JOLs, as two differences between the binary-betting task (as well as the binary yes/no JOL task

used by Hanczakowski et al., 2013; see footnote 4) and the 0-100% JOL task can be noted.[8] First, the number of possible response options differs, with two options in the binary tasks and multiple options in the 0-100% JOL task. Second, the use of numerical labels differs, with the binary task using none and the scale JOL task using percentage values as labels. Experiment 3 was designed mainly to address this confound.

Experiment 3

The main purpose of Experiment 3 was to adjudicate whether the discrepant findings of Experiments 1 and 2 should be assigned to different number of options available in the assessment tasks or differential use of numerical labels in these tasks. To this aim, the present experiment used another assessment task: one in which participants were asked to make binary 0/100% JOLs; that is, choose either "0%" or "100%" for each studied pair (see also Higham, Zawadzka, & Hanczakowski, 2015, for this type of rating task used to assess retrospective confidence). This task preserves the binary format of responding from Experiment 2 while simultaneously using numerical values that were used in Experiment 1. If the limited number of response options was responsible for abolishing the UWP pattern in Experiment 2, we would expect the same perfect calibration for all item types in Experiment 3. If it was the lack of numerical values that eliminated the UWP pattern, we would expect the pattern of UWP observed in Experiment 1.

Another aim of Experiment 3 was to shed more light on the differences between items previously recalled once and twice. Recall that in Experiment 1, we found a difference in the magnitude of the UWP effect between these item types. This difference was driven by variation in the magnitude of 0-100% JOLs, as recall for

[8]We thank an anonymous reviewer for pointing this issue to us.

both types of previously recalled items was comparable and at ceiling. Why, then, would participants assign different ratings to items for which future recall was almost identical? We speculated that participants were aware of subtle differences between items recalled once and twice which were not reflected in the measure of immediate recall, but which nevertheless shaped participants confidence in future recall. To test for this possibility, in Experiment 3 we implemented a manipulation of delay in testing on cycle 3 by testing half of the items immediately (as in previous experiments), and the other half after completing another experiment, approximately 40 minutes later. Previous research suggests that differences in memory strength between different classes of items are more likely to be detected if the criterial test is more difficult (Kornell, Bjork, & Garcia, 2011). Testing memory for one half of the items after a delay should obviously increase test difficulty for these items. Thus, if items previously recalled twice are indeed more strongly encoded than those recalled only once, a difference in delayed recall should be found.

Method

Participants. Twenty-four students participated in this study for course credit.

Materials and procedure. The materials were the same as in Experiments 1 and 2. On the first two cycles, only the rating task differed from those used in the previous experiments. For each studied pair, participants were asked to provide a binary 0/100% JOL, that is, to choose either “0%” or “100%”. The instructions for the rating task were adapted from those used by Hanczakowski et al. (2013) in their yes/no JOL experiment, with “yes” and “no” substituted with “100%” and “0%”, respectively. On cycle 3, the study phase was the same as on the previous cycles. The test, however, unbeknown to participants, was split into two halves. The first half of the test, on which participants were tested on 30 of the studied word pairs, was

administered immediately after the study phase. After completing this half of the test, participants took part in another, unrelated experiment, which took between 35 and 45 minutes (responding in that experiment was self paced). The second half of the test was administered immediately after this "filler" experiment and consisted of the remaining 30 word pairs. The assignment of word pairs to test halves was counterbalanced.

In contrast to Experiments 1 and 2, on cycle 3 tests participants were not asked to estimate the number of preceding recall successes for each pair. Such assessments were deemed superfluous given extremely high correlations between assessed and actual recall performance documented in Experiments 1 and 2, and excluding these assessments allowed for shortening the procedure, which, together with the filler experiment, took about 1 hour 30 minutes to complete.

Results and Discussion

The means for 0/100% JOLs, recall performance, and resolution (*d'*) are presented in Table 1. First, we performed the same analyses as in Experiments 1 and 2 on the full set of immediate data.[9] Resolution again was affected by cycle, $F(2, 40) = 26.425$, $MSE = .239$, $p < .001$, $\eta_p^2 = .569$: it increased from cycle 1 to cycle 2, $t(23) = 5.076$, $SE = .132$, $p < .001$, $d = 1.04$, and from cycle 2 to cycle 3, $t(20) = 3.071$, $SE = .131$, $p = .006$, $d = 0.67$.[10] A 2 (measure: 0/100% JOL, recall performance) x 3 (cycle: 1, 2, 3) ANOVA revealed a significant main effect of cycle, $F(2, 46) = 93.170$, $MSE = .012$, $p < .001$, $\eta_p^2 = .802$, which was qualified by a significant interaction, $F(1, 23) = 44.168$, $MSE = .009$, $p < .001$, $\eta_p^2 = .658$. On cycle 1, 0/100% JOLs exceeded recall

[9] Note that this means that, for cycle 3, only the 0/100% JOL and recall data for 30 word pairs presented on the immediate test were subjected to these analyses.

[10] It was not possible to calculate cycle-3 *d'* for three participants, as they correctly recalled all items.

performance, $t(23) = 5.184$, $SE = .045$, $p < .001$, $d = 1.06$, while on the next two cycles recall performance exceeded 0/100% JOLs, $t(23) = 1.999$, $SE = .032$, $p = .058$, $d = 0.41$ for cycle 2 and $t(23) = 3.513$, $SE = .024$, $p = .002$, $d = 0.72$ for cycle 3, demonstrating the UWP pattern. The main effect of measure was not significant, $F(1, 23) = 1.124$, $MSE = .027$, $p = .30$, $\eta_p^2 = .047$. Overall, these results demonstrate that it was not merely the change of task format from scale to binary that eliminated the UWP pattern in Experiment 2.

Insert Figure 2 about here

The crucial results concern calibration for various classes of items on cycle 3. The results for recall and 0/100% JOLs conditionalized on the number of previous recall successes (see panel A of Figure 2) demonstrate a different locus of the overall UWP effect observed in this experiment, as compared to Experiment 1. This time, the presence of the UWP pattern was driven by a huge underestimation (28% to 49%) of future recall performance for previously unrecalled items, $t(23) = 3.793$, $SE = .056$, $p = .001$, $d = 0.78$. For items recalled at least once, on the other hand, 0/100% JOLs tracked recall performance extremely well, $t < 1$ for once-recalled items and $t = 1$ for twice-recalled items. Consequently, calibration was equally good for items previously recalled once and twice, $t < 1$. This lack of difference in calibration joins the results of Experiment 2 in showing that participants can accurately predict future recall performance for previously recalled items if given an appropriate task.

In addition to the analyses of immediate recall data described above, we analyzed data from the delayed recall test (see panel B of Figure 2). The purpose of delaying one half of the cycle-3 test was to more clearly demonstrate that items

previously recalled once and twice indeed differ in terms of retention. Our results were consistent with this hypothesis: items previously recalled twice were more often recalled on the delayed test than items previously recalled once, $t(22) = 2.973$, $SE = .024$, $p = .007$, $d = 0.62$. This strengthens the claim that, in the scale-JOL task such as that employed in Experiment 1, participants can detect true differences between items, even if these differences are not revealed in the measure of immediate recall performance.

The main interest of the present experiment lies again in the calibration pattern. The pattern of calibration on cycle 3 for previously recalled items indicates that the discrepancy between Experiments 1 and 2 is likely caused by the number of response options, not the use of the betting or percentage framing. In both Experiments 2 and 3, binary responding was required and no UWP for previously recalled items was observed, which contrasts with the scale JOL results of Experiment 1. This implies that it is the requirement of providing fine-grained distinctions in the 0-100% JOL task that is responsible for UWP for previously recalled items.

The present observation of a reliable difference in delayed recall for previously once and twice recalled items leaves clues as to why the pattern of UWP is generated. The difference in delayed recall shows that there is a real difference in memory information for items previously recalled once and twice. Given a scale JOL task participants may feel inclined to express this real difference by lowering JOLs for items with weaker memorial basis (previously recalled once). However, immediate recall, which provides a standard against which these JOLs are assessed in terms of calibration, shows only a negligible difference between items previously recalled once and twice (1% in Experiment 1, 4% in Experiment 2, and 3% in Experiment 3). The result of this is the pattern of exaggerated UWP for once-recalled items in the scale

JOL task, a pattern that should not be taken to imply that participants are somehow unaware that the likelihood of immediate retrieval of the already mastered items is actually extremely high.

In the present study we focused mostly on items previously recalled once or twice. We argue that it is number of responding options which is crucial for the calibration pattern observed for these items. This, however, is not to say that the betting vs. percentage framing does not matter. The results of Experiment 3 clearly show that framing has important consequences for the calibration pattern on cycle 3 for items that were not previously recalled. Specifically, in Experiment 2 these items show perfect calibration in the betting task, yet in Experiment 3 a massive UWP was shown for these items in the 0/100% JOL task.

Why, then, do the results obtained in two binary tasks differ so much? One possibility is that the binary betting task might minimize the probability of biased responding as compared to other binary tasks (see footnote 4). In short, the betting task utilizes an explicit pay-off matrix in which losses from different incorrect responses – not betting on the item subsequently recalled or betting on the item subsequently unrecalled – are equated. In the 0/100% JOL task, on the other hand, the pay-off matrix is only implicit and thus different errors may have different psychological value. It is therefore possible that the well-defined pay-off matrix of the betting task facilitates perfect calibration for all types of items observed in Experiment 2, whereas an implicit pay-off matrix of the 0/100% JOL task introduces bias by which participants err on the side of too many 0% judgments. Alternatively, it might be the 100% label used in the binary 0/100% task that is responsible for the UWP pattern for previously unrecalled items in Experiment 3. The highest value on the confidence scale (“100%”) may seem inappropriate for such items, something not

shared by the decision to bet. If so, participants may be more reluctant to assign a 100% judgment than to bet on previously recalled items, which would produce the UWP effect we observed.

In total, our results clearly show that the issue of labels used to elicit metacognitive judgments requires further empirical scrutiny. It is worth pointing out here that it is not the first case described in the metacognitive literature of a discrepancy between two rating types that share the format of responding. To note just two prominent examples, McCabe and Soderstrom (2011) demonstrated differences in resolution between JOLs and judgments of remembering and knowing (JORKs), both elicited on three-point scales, and Dougherty et al. (2005) found differences between retrospective confidence judgments and delayed JOLs elicited on percentage scales.

General Discussion

The present study investigated the intricacies of the UWP pattern in a multi-cycle learning procedure. With respect to a general pattern of calibration, in Experiment 1, which used the 0-100% scale, the UWP pattern was present on cycles 2 and 3. In Experiment 2, which employed the binary betting task, no UWP was found on any of the cycles. The results of Experiment 2 replicate the findings of Hanczakowski et al. (2013), who also demonstrated good calibration in the binary betting task in their Experiments 3 and 4. Finally, Experiment 3, which used binary 0/100% JOLs, in which the UWP pattern reemerged on cycles 2 and 3 for previously unrecalled items, demonstrated that the UWP pattern can be reliably obtained when only two response options are available to participants.

The main objective of this study was to investigate the influence of the number of past recall successes on calibration across the task formats. Experiment 1

demonstrated that calibration was impaired regardless of the number of past recall successes, but the magnitude of this impairment was far greater for items previously recalled once rather than twice, which was mostly due to lowered 0-100% JOLs for the former item class. Conversely, in Experiment 2, betting decisions tracked recall performance almost perfectly for all classes of items, suggesting that people are able to predict their future recall performance with very high accuracy if given an appropriate task. Also in Experiment 3, 0/100% JOLs assigned to previously recalled items tracked cycle-3 recall extremely well. The results of Experiments 2 and 3 demonstrate that the difference in the magnitude of the UWP effect for previously recalled items that we observed in the scale task cannot be attributed to an inability to make correct predictions concerning future recall performance for these items. In general, the results for previously recalled items reveal a dissociation between scale and binary judgments with respect to calibration.

Why, then, does this dissociation occur? We believe that the results of the present study are consistent with the claim of Hanczakowski et al. (2013) that participants' 0-100% JOLs are not assessments of recall probability, but rather represent confidence judgments. The difference between *probability* and *confidence* ratings has profound consequences for calibration research for which it is common practice to directly compare mean JOLs and mean recall. For this comparison to be meaningful for the assessments of realism of JOLs, there is an assumption that intervals on the scale on which ratings are made are comparable to the intervals on the underlying psychological dimension that the scale values are meant to index. However, confidence scales are likely only ordinal; that is, JOLs may simply represent a rank ordering of the recallability of items, not recall probabilities. For the latter, participants must ensure that the psychological distance between 70% and 80%

is the same as that between 20% and 30% (or any other pairs of values that differ by 10%), which seems unlikely (e.g., see Poulton, 1979). That being the case, direct comparisons between mean of scale JOLs and mean recall provides little to no information about the realism of people's judgments.

The present findings confirm and extend those of Hanczakowski et al. (2013). First, they replicate their main finding that binary-betting and 0-100% JOL tasks give rise to different results, even though in theory they are supposed to measure the same underlying construct: probability of future recall. Second, our findings directly demonstrate different properties of binary and scale ratings. Betting decisions turn out to be able to accurately track recall performance, independent of the number of past recall successes. Scale JOLs, on the other hand, are dependent on the number of past recall successes to a much greater extent than immediate recall performance is, as evidenced by the difference in the magnitude of the UWP effect between items previously recalled once and twice. In total, our results strengthen the conclusion that scale JOLs, as compared to betting decisions, do not measure subjective probability, but rather are ratings of confidence in future recall.

How might an account that considers 0-100% JOLs to be confidence ratings rather than probability judgments account for our results? According to the confidence account of scale JOLs, even though participants in Experiment 1 were aware that the *probability* of recalling a previously recalled item is similar for the two classes of items and close to 100%, they may have assumed that each additional successful recall attempt warrants an increase in *confidence* in future recall. Confidence ratings are not aimed at providing numerical assessments of probability. For this reason, they do not have to track closely recall performance. Participants may base their JOLs on cues that allow them to demonstrate that they are aware of actual

differences amongst the subsets of highly recallable items. That these differences really exist is well exemplified by the difference in delayed recall performance between items previously recalled once and twice that we demonstrated in Experiment 3.

Insert Figure 3 about here

It is worth noting that the results of this study do not necessarily suggest that participants access different kinds of information in the scale-JOL and binary tasks. In fact, it is perfectly viable that participants have the same information available in scales and binary tasks, but choose different cues for each of the tasks, which, in turn, produces the scale/binary dissociation (see also Higham et al., in press).[11] Consider the three panels of Figure 3 which present a single-dimension interpretation of the results of Experiments 1-3. Binary tasks, presented in the middle (binary betting) and bottom (0/100%) panels, require only a single decision criterion: evidence available for each item either falls below or above that criterion. As previously recalled items all have high levels of evidence for future recall, virtually all of them fall above that single criterion. This results in the insensitivity of binary responses to the once versus twice cue that the present study revealed (i.e., all of them are assigned positive responses). Scale ratings (top panel), on the other hand, require multiple decision criteria - one for each confidence value. This means that several of these criteria (80%, 90% and 100% in the example presented in Figure

[11] Note that this reasoning is consistent with the results of Experiment 4 of Hanczakowski et al. (2013). In this experiment, binary-betting decisions and scale-JOL ratings were elicited for each studied item in order to eliminate the possibility that participants access different information in binary and scales tasks; nevertheless, the binary/scales dissociation was still found on cycle 2. The authors interpreted this result as inconsistent with the different-bases view of binary and scale ratings.

3) are positioned at the high end of the evidence dimension, where recall performance is at ceiling. In this case, the recall probability cue is of little use from the perspective of a participant who wants to discriminate between these strongly encoded items. For this reason, high JOLs have to be driven by other cues, not all of which have to be related to the probability of immediate recall.

The present study looked at only one type of cue that shapes scale JOLs: the number of previous successful retrievals. An interesting issue for future research would be to examine how this cue interacts with other cues in the multi-cycle procedure, such as, for example, the fluency of retrieval (e.g., Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005). Although Serra and Dunlosky (2005) have shown that the overall contribution of retrieval fluency to the UWP effect in a two-cycle procedure is minimal, this cue might potentially play a role in procedures consisting of three or more cycles. As mentioned above, previously recalled items that differ in the number of past recall successes most likely have more commonalities than differences, which limits the pool of cues participants can utilize in order to distinguish between these items. In such a case, if items previously recalled twice were indeed more fluently retrieved than items recalled only once, it could contribute to the difference in JOLs between these items (see also Pyc & Rawson, 2012, for the influence of retrieval fluency on JOLs for successfully recalled items).

Finally, it has to be noted as well that the number of successful retrieval attempts can be predictive of future recall under certain circumstances. As the results of the delayed test in Experiment 3 demonstrate, the more times an item is recalled, the better the memory for that item after a delay (see also Vaughn & Rawson, 2011), and scale JOLs are known to be sensitive to this cue (Pyc & Rawson, 2012). Even though participants are aware that their task is to predict recall on an immediate test,

rendering long-term predictions of retention irrelevant for the task they face, the number of previous successful recalls may be incorporated into ratings to demonstrate that items previously recalled once and twice differ in terms of memorial evidence.

In any case, it can safely be assumed that the quality and/or quantity of evidence for future recall differs between items previously recalled once versus twice, and scale JOLs pick up on that difference even when the items do not differ in terms of probability of future recall on an immediate test. For researchers interested in people's insight into their own memory, scale JOLs will provide more detailed information than less sensitive betting decisions. This sensitivity of scale JOLs comes, however, at a cost: it makes the task unsuitable for querying assessments of subjective probability. Certain (but not all) binary decisions, such as binary bets or yes/no JOLs, on the other hand, fare much better as proxies to subjective probability. Researchers interested in calibration of judgments of future memory performance should therefore opt for using the binary betting task with equated and well-defined rewards and penalties to reduce the chance of obtaining biased estimates of future recall probability. The present results therefore underscore the importance of choosing an appropriate task for answering a research question.

References

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a mnemonic index. *Journal of Experimental Psychology: General, 127,* 55-68. doi: 10.1037/0096-3445.127.1.55

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33,* 1096-1115. doi: 10.3758/BF03193216

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. doi: 10.3758/s13423-012-0237-7

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36,* 813-821. Doi : 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, & Cognition, 33,* 238-244. doi: 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. doi: 10.1016/j.jml.2007.03.006

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, 429-444. doi: 10.1016/j.jml.2013.05.003

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (in press). *Internal mapping and its impact on measures of absolute and relative metacognitive accuracy*. Chapter

to appear in J. Dunlosky and S. K. Tauber (Eds.) *Oxford Handbook of Metamemory*.

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2015). *Dissociations between percent-scale and binary metacognitive judgments: The case for convergent measures.* Manuscript in preparation.

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. doi: 10.1037/0096-3445.126.4.349

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80-113. doi: 10.1037/a0022171

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643-656. doi: 10.1037/0096-3445.133.4.643

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. doi: 10.1016/j.jml.2005.01.001

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. doi: 10.1037/0096-3445.131.2.147

Kornell, N., Bjork, R. A., & Garcia, M. A. (2009). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*, 85-97. doi: 10.1016/j.jml.2011.04.002

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based prospective

metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKs). *Journal of Experimental Psychology: General, 140,* 605-621. doi: 10.1037/a0024014

McGillivray, S., & Castel. A. D. (2011). Betting on memory leads to metacognitive improvement by younger and older adults. *Psychology and Aging, 26,* 137-142. doi: 10.1037/a0022681

Pollack, I., Norman, D. & Galanter, E. (1964). An efficient nonparametric analysis of recognition memory. *Psychonomic Science, 1*, 327-328.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86,* 777-803. doi: 10.1037/0033-2909.86.4.777

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an ex- planation via anchoring. *Journal of Experimental Psychology: General, 134,* 124–128. doi: 10.1037/0096-3445.134.1.124

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

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *The Quarterly Journal of Experimental Psychology, 65,* 2231-2257. doi: 10.1080/17470218.2012.685081

Pyc, M. A., & Rawson, K. A. (2012). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition, 40,* 976-988. doi:10.3758/s13421-012-0200-x

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

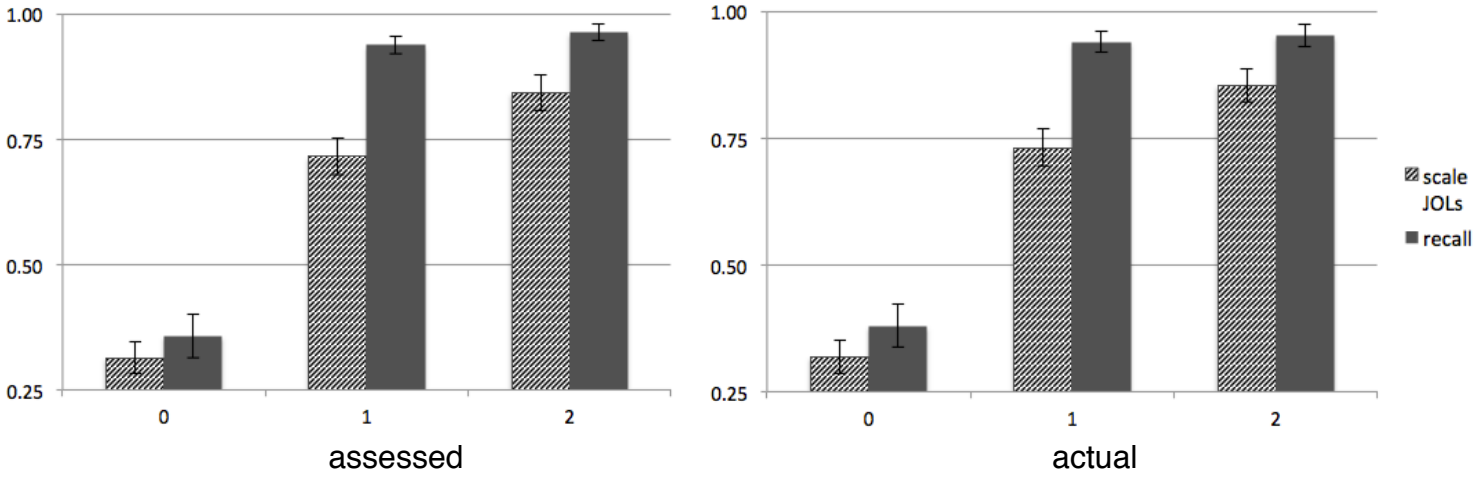
Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*, 1127-1131. doi:10.1177/0956797611417724

Table 1. *Means (SEs) for Recall Performance, Scale JOLs, Proportion of Bets, 0/100% JOLs and Resolution (A_g for Scale JOLs and d' for Bets and 0/100% JOLs) as a Function of Cycle in Experiments 1-3.*

	Measure					
Experiment and Cycle	Recall	Scale JOLs	Betting	0/100% JOLs	*A*	*d'*
Experiment 1						
cycle 1	.22 (.03)	.43 (.04)	-	-	.61 (.02)	-
cycle 2	.51 (.05)	.39 (.04)	-	-	.75 (.02)	-
cycle 3	.65 (.05)	.56 (.04)	-	-	.83 (.02)	-
Experiment 2						
cycle 1	.29 (.05)	-	.43 (.04)	-	-	1.10 (0.15)
cycle 2	.56 (.05)	-	.52 (.05)	-	-	1.84 (0.15)
cycle 3	.70 (.05)	-	.69 (.05)	-	-	2.11 (0.12)
Experiment 3						
cycle 1	.23 (.03)	-	-	.47 (.04)	-	0.76 (0.13)
cycle 2	.53 (.04)	-	-	.46 (.04)	-	1.43 (0.08)
cycle 3 immediate	.70 (.04)	-	-	.61 (.04)	-	1.81 (0.10)
cycle 3 delayed	.62 (.04)	-	-	.66 (.04)	-	2.08 (0.14)

Note: The row "cycle 3 immediate" in Experiment 3 combines 0/100% JOLs elicited during the immediate study phase 3 with recall data from the immediate test. The row "cycle 3 delayed" combines 0/100% JOLs elicited during the immediate study phase 3 with recall data from the delayed test.

Experiment 1



Experiment 2

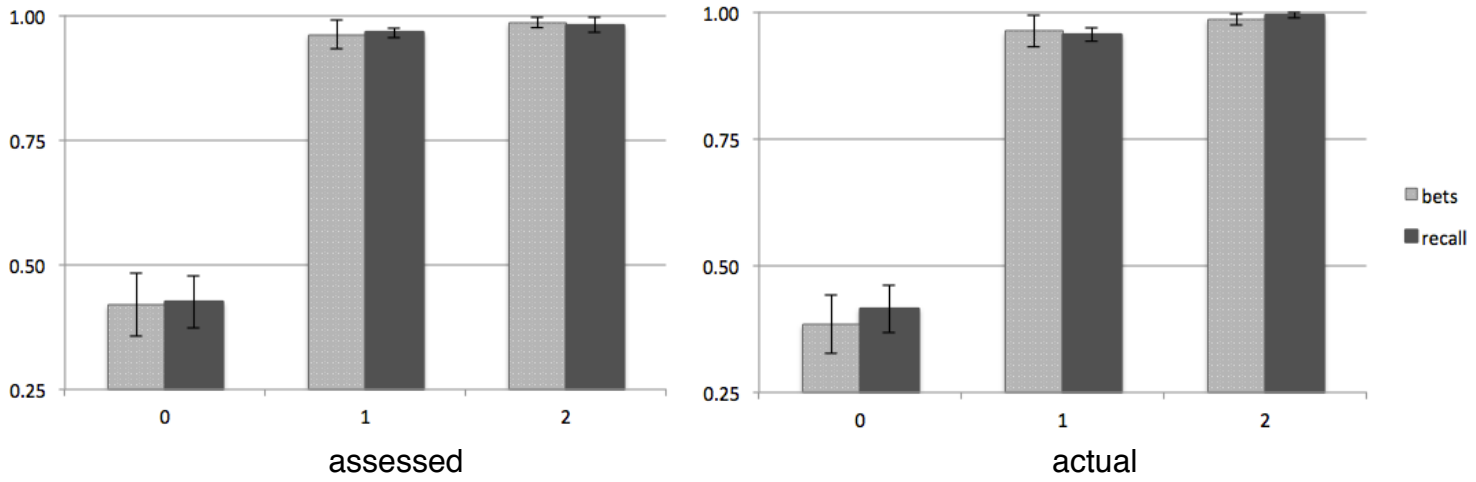


Figure 1. Mean 0-100% JOLs (Experiment 1), proportion of bets (Experiment 2) and recall performance (both experiments) on cycle 3 as a function of the assessed (left panel) and actual (right panel) number of previous recall successes on cycles 1 and 2. Error bars indicate standard error of the mean.

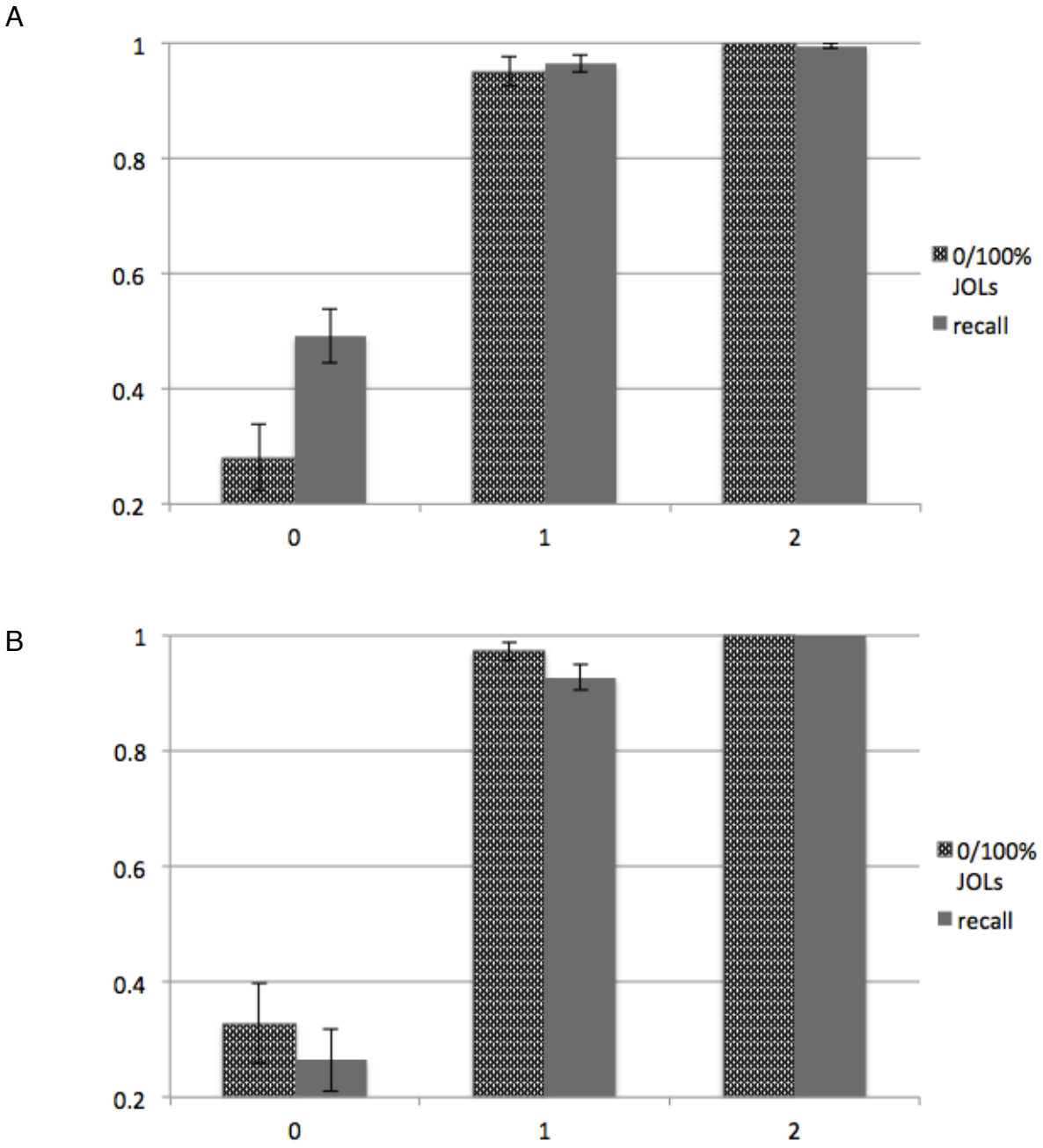


Figure 2. Mean 0/100% JOLs and recall performance on cycle 3 of Experiment 3. Panel A presents recall results from the immediate test. Panel B presents recall results from the delayed test. Error bars indicate standard error of the mean.

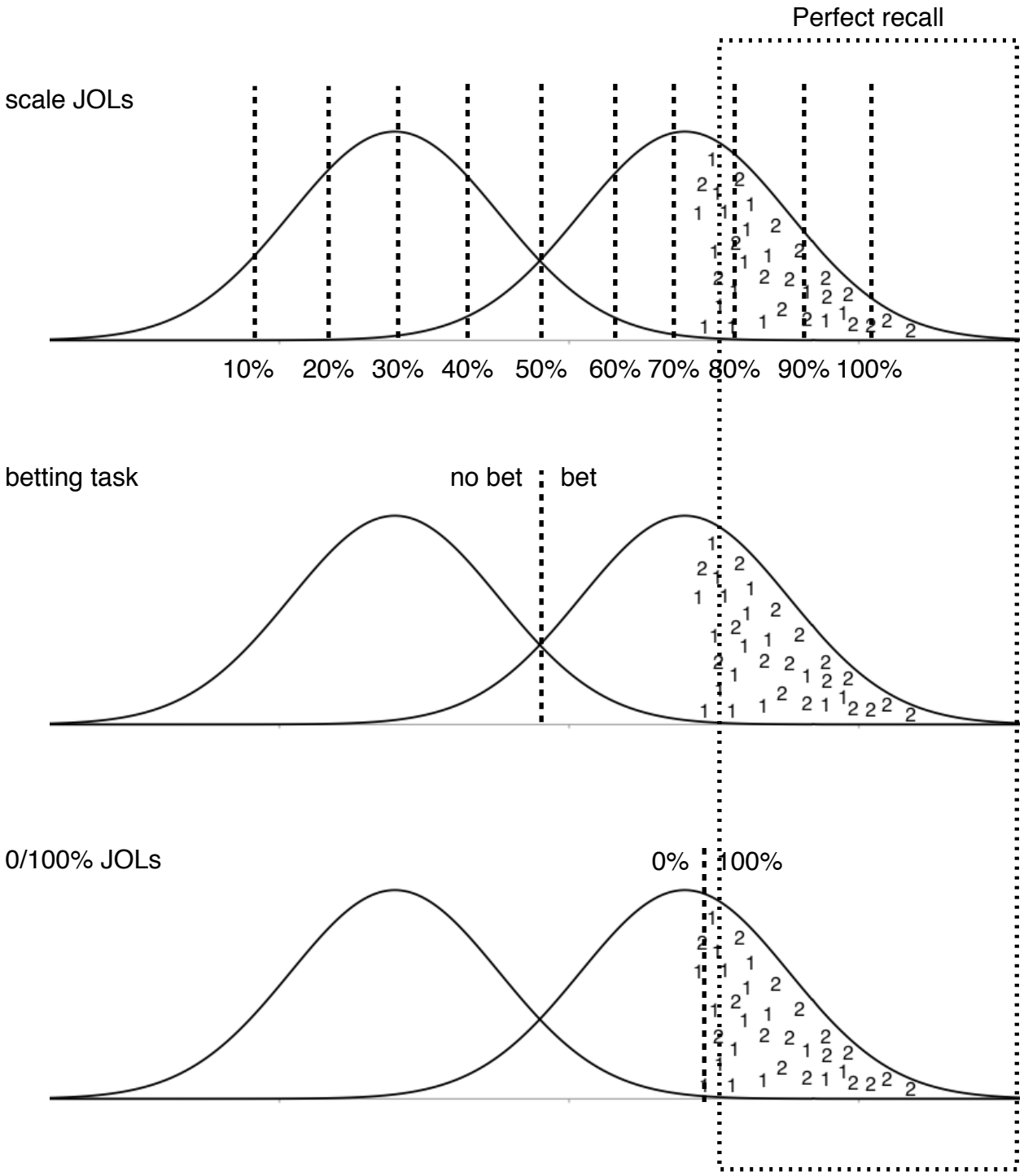


Figure 3. A simplified graphical depiction of results for previously recalled pairs in tasks used in Experiments 1-3. In each panel, two distributions are positioned on the evidence-for-future-recall dimension. The dimensions on the left represent unrecalled items, while the distributions on the right represent correctly recalled items. Numbers denote items previously recalled once (1) and twice (2). Dashed lines represent criteria: confidence criteria in increments of 10% in the case of Experiment 1, and binary criteria in Experiments 2 and 3. The dotted rectangular area encompasses items positioned in the upper part of the evidence dimension, for which recall performance is 100%.