When Two Learners Are Better Than One: Using Flashcards with a Partner Improves Metacognitive Accuracy

*Megan N. Imundo[1,2], Inez Zung[3], Mary C. Whatley[1, 4] and Steven C. Pan[5]

[1]Department of Psychology, University of California, Los Angeles
[2]Learning Engineering Institute, Arizona State University
[3]Department of Psychology, University of California, San Diego
[4]Department of Psychology, Western Carolina University
[5]Department of Psychology, National University of Singapore

Megan N. Imundo | mimundo@asu.edu | https://orcid.org/0000-003-4599-4777
Inez Zung | izung@ucsd.edu | https://orcid.org/0000-0002-0947-2309
Mary C. Whatley | mwhatley@email.wcu.edu | https://orcid.org/0000-0003-3609-5630
Steven C. Pan | scp@nus.edu.sg | https://orcid.org/0000-0001-9080-5651

Correspondence concerning this article should be addressed to Megan Imundo, Learning Engineering Institute, Arizona State University. Email address: mimundo@asu.edu. Mailing address: 121 Ira D. Payne Educational Hall, 1000 S Forest Mall, Tempe, AZ 85281.

Declaration

Ethics Approval

Ethical approval was granted by UCLA's Institutional Review Board (IRB# 11-002880).

Author Contributions

M.N.I., I.Z., and S.C.P. designed the experiments and developed the materials. M.N.I., I.Z., S.C.P., and M.C.W. collected experiment data. M.N.I. analyzed and interpreted the data with input from I.Z., S.C.P, and M.C.W. M.N.I drafted the manuscript and I.Z, S.C.P., and M.C.W. provided substantial manuscript edits. All authors approved the manuscript for submission.

Competing Interests

The authors have no conflicts of interest to declare.

Funding

No funding supported this research.

Data Availability Statement

De-identified research data are publicly available via Open Science Framework https://osf.io/hac38/?view_only=9ff44668b2184f1b8e412452f3a41640.

Acknowledgements

An earlier version of this work is included in the first author's dissertation. The authors would like to thank Elizabeth Ligon Bjork, Robert Bjork, Melissa Paquette-Smith, Idan Blank, and Alan Castel for their comments and suggestions on an earlier version of this manuscript.

Abstract

We investigated the benefits of two ways to use flashcards to perform retrieval practice: alone versus with a partner. In three experiments, undergraduate students learned word-definition pairs using flashcards alone (Individual condition) or with another student (Paired condition). Participants then made global judgments of learning (gJOLs; Experiments 1-3), and item-level judgments of learning (iJOLs; Experiment 3). Finally, participants took a cued-recall test after a 5-min delay (Experiments 1-3) and a 24-hour delay (Experiments 2-3). In Experiment 1, students in the Paired condition dropped flashcards less often than in the Individual condition (dropping was prohibited in Experiments 2-3). In addition, although final test performance tended to be similar across conditions, inaccurate gJOLs for the immediate test—inflated by ~20% relative to actual immediate test performance—were common in the Individual condition but not in the Paired condition in Experiments 1-2. In Experiment 3, we tested whether this difference in metacognitive calibration was due to the Paired condition requiring overt retrieval by instructing participants in the Individual condition to retrieve out loud. With this change, participants in the Individual and Paired conditions reported similarly accurate gJOLs and iJOLs. Taken together, these findings suggest that although performing retrieval practice with flashcards alone versus with a partner yields comparable amounts of learning, doing so with a partner can increase metacognitive accuracy, a benefit possibly driven by the facilitation of overt retrieval. Overall, these findings have implications for self-regulated learning and effective exam preparation.

Keywords: flashcards; self-regulated learning; metacognition; retrieval practice; test-enhanced learning

When Two Learners Are Better Than One: Using Flashcards with a Partner Improves Metacognitive Accuracy

Learning scientists often recommend that students use flashcards to prepare for exams (e.g., Smith & Weinstein, 2016). This suggestion is based on the premise that flashcards facilitate *retrieval practice* (i.e., practice testing), which is a potent enhancer of long-term memory (i.e., the *testing effect*; [Author] & Rickard, 2018; Roediger & Butler, 2011; Rowland, 2014 offer comprehensive reviews). Indeed, an in-depth review of popular learning techniques ranked retrieval practice as among the most effective (Dunlosky et al., 2013). Large surveys indicate that most undergraduate students use flashcards to prepare for their classes and often engage in retrieval practice when doing so, with the most common purpose being to learn vocabulary (Wissman et al., 2012; [Authors], 2022a). Flashcards are commonly prepared by writing a key concept or term on one side and associated information (e.g., related concepts, definitions, etc.) on the reverse, thus making it convenient to quiz oneself or others.

Beyond its benefits for memory, retrieval practice can also aid learning in other, less obvious ways. One such benefit involves improving students' control of study behaviors (e.g., time per item, decision to stop studying) during self-regulated learning. According to prominent theories of metacognition (e.g., Nelson & Narens, 1990), such control is commonly based on students' monitoring of their own learning (e.g., judgments of learning, confidence in retrieved answers). If a student inaccurately monitors her learning and is overconfident, then she may stop studying prematurely and be left with poor mastery of to-be-learned information. Retrieval practice can prevent that overconfidence: Miller and Geraci (2014) found that a single retrieval practice opportunity, which usually provides learners with concrete evidence as to their mastery of the material (e.g., via retrieval success or failure), can lower inflated judgments of learning (also Tullis et al., 2013). Retrieval practice can also help students optimize their study activities: Soderstrom and Bjork (2014) found that students spend more time studying difficult materials, and learn them more effectively, after engaging in retrieval practice. These findings reinforce the value of retrieval practice as not just a memory enhancer, but also a way to improve metacognitive accuracy and study decisions. It should be noted, however, that such benefits have typically been demonstrated using methods that do not involve flashcards.

Optimizing Flashcard-Based Retrieval Practice

Although flashcards can facilitate retrieval practice, the conditions under which they are most effective remains to be fully established (Lin et al., 2018; [Authors], 2022b; Senzaki et al., 2017; [Authors], 2022a offer additional discussion), and there is evidence that students use flashcards ineffectively and remain susceptible to illusions of competence when doing so. For instance, students may choose to download premade flashcard sets, even though generating flashcards can facilitate learning ([Authors], 2022b). Students also often drop flashcards before their content is well-learned: Kornell and Bjork (2008) found that dropping is common after just one correct retrieval attempt, resulting in reduced learning relative to conditions wherein dropping is disallowed. Further, students prefer smaller flashcard stacks, thinking that they are more beneficial for learning (Wissman et al., 2012), when larger stacks enable learning to be better distributed out in time (i.e., the *spacing effect*; Kornell, 2009). Finally, one-third of students do not always check the accuracy of their responses when using flashcards (Wissman et al., 2012). This pattern is especially problematic when considering that students sometimes drop flashcards even before a single successful retrieval (possibly due to inadequately assessing the correctness of their responses; e.g., Kornell & Bjork, 2008, Experiment 3). Together, these findings reveal substantial room for improvement in students' use of flashcards.

One promising method for improving flashcard use involves doing so with a partner—that is, using flashcards in pairs as opposed to individually. There are several reasons why using flashcards in pairs may be beneficial. First, is the need for overt responses. Some studies of the testing effect find that overt responding more reliably produces a testing effect than covert responding (Jönsson et al., 2014; Kubik et al., 2020; Krumboltz & Weisman, 1962), whereas others *do* observe a testing effect following covert retrieval practice (Carpenter & Pashler, 2007), or observe no difference in the testing effect between overt and covert retrieval practice (Putnam & Roediger, 2013). There is not yet an established explanation as to why overt retrieval practice may at times lead to greater learning than covert retrieval practice. Jönsson and colleagues (2014) suggest that overt retrieval might elicit greater processing as it requires both the attempt to retrieve the target content and the overt expression of that content. Tauber and colleagues (2018) suggest that overt retrieval practice establishes accountability for the learner. This accountability may encourage more complete retrieval attempts, especially when material is complex enough (i.e., more than single words) to allow for exhaustive retrieval. Thus, overt retrieval may discourage learners from "cheating themselves" by not fully articulating a response to a given question or cue, with retrieval attempts more potent and more informative for metacognitive judgments as a result.

Second, the presence of others may affect learners' emotional states positively. Supporting evidence comes from students' self-reports which indicate that studying with others increases motivation to learn, is more enjoyable, and improves learning relative to studying individually (McCabe & Lummis, 2018; Wissman & Rawson, 2016). One review of the collaborative testing literature, for example, indicates that testing with peers might reduce test anxiety (LoGuidice et al., 2015). Evidence suggests that experiencing positive emotions can contribute to learning outcomes (Holzer et al., 2021; Pekrun et al., 2002). Affect also has implications for a learner's metacognitive experiences (Efklides, 2006). The Metacognitive and Affective Model of Self-Regulated Learning (Efklides et al., 2018; also Hayat et al., 2020), for example, suggests that metacognition includes an affective component and metacognition both affects and is affected by positive and negative emotions. Further, the presence of others may increase motivation during learning (i.e., social facilitation); however, if students fear evaluation from their partner, then their learning may suffer (Geen, 1983).

Third, learners may seek feedback from their partner rather than assessing the validity of their response via a sense of fluency, thus reducing susceptibility to illusions of competence. Additionally, a partner might offer explanations and correct errors, further facilitating learning (Johnson et al., 1998; LoGuidice et al., 2015). All of these reasons suggest that using flashcards with a partner—which has yet to be extensively investigated—may be beneficial.

The Present Study

We investigated the hypothesis that flashcard-based retrieval practice with a partner is better for learning than individual flashcard-based retrieval practice. Additionally, in an exploratory manner, we examined if flashcard-based retrieval practice leads participants to report more accurate post-learning metacognitive judgments when it is implemented with a partner as opposed to implemented individually. In a similarly exploratory approach, we also examined potential differences between individual and paired flashcard learning in terms of the mechanics of flashcard use (e.g., cycles through the flashcard set), associated study decisions (e.g., dropping cards), and affective states.

Across three experiments, undergraduate students learned word-definition pairs using flashcards alone (the Individual condition) or with another student (the Paired condition), answered relevant survey questions, and then completed a final test. In Experiment 1, dropping of flashcards was allowed whereas in Experiments 2-3 it was prohibited. Additionally, whereas in Experiment 1 both Individual and Paired learners engaged in cycles of study and retrieval practice, in Experiments 2-3, all learners completed an initial study period such that Individual learners then only engaged in retrieval practice. We believe that this approach is more aligned with students' own behaviors when using flashcards in daily life ([Authors], 2022a). Finally, in Experiment 3 participants in the Individual condition were instructed to overtly retrieve (i.e., talk out loud) during the flashcard phase. Importantly, in all experiments and across conditions, we controlled for total learning time, used the same flashcards and learning environments, and gave similar instructions.

Experiment 1

In the first experiment, learners had 20 min each to study a set of vocabulary-definition pairs and to perform retrieval practice on those pairs. They were assigned to do so by themselves or with a partner. In the case of Individual learners, such learning involved 20 min of studying followed by 20 min of retrieval practice. For Paired learners, the logistics were somewhat more complex: One partner served as the "tester" and the other partner as the "testee" before the roles reversed. Hence, in the Paired condition, one partner engaged in 20 min of practice testing from the outset, whereas the other partner did so after those 20 min had elapsed.

Method

The study was preregistered at: https://osf.io/mqunz/?view_only=bdb8d5cce52c43a6ba400a58a70749f5

Participants

One hundred and fifty-two undergraduate students (Individual condition, *n* = 64; Paired condition, *n* = 88) from the participant pool at a large public research university participated in exchange for course credit. Data from two additional participants were excluded because they experienced technical malfunctions. The target sample size, 150, was determined using an *a priori* power analysis conducted in G*Power (Faul et al., 2007) in which at least 32 participants per group is needed to detect a medium effect size (Cohen's *f* = 0.25) in a between-participants design at 80% power and with a standard .05 error probability. To reach that target, data collection occurred continuously for eight weeks and concluded only with the scheduled close of the participant pool recruitment period.

Design

The experiment employed a 2 x 2 between-subjects factorial design with factors of Condition (Individual versus Paired) and First Learning Activity (Study First versus Test First; detailed later in this manuscript). Participants (a) learned individually or in pairs and (b) studied or tested first before switching learning activities.

Materials

The materials included 40 word-definition pairs, each consisting of a Graduate Record Examination (GRE) vocabulary word and its definition (e.g., *monolithic*: *made of only one stone*). The words were drawn from *The Economist*'s "Most Difficult GRE Words" list for 2020, whereas the definitions were drawn from Dictionary.com. The words and their definitions were 4-10 letters and 5-10 words in length, respectively; the words had a Kucera-Francis frequency of 1-3. In the case of multiple definitions, the first definition was used, and if that definition contained the GRE word, the second definition was used. All stimuli are listed in the Appendix.

Each word-definition pair was printed on a 4 x 6 in. white index flashcard. For the *standard* flashcard set, which was designed for retrieval practice, each card displayed a GRE word on the front and the word and its definition on the back. For the *study-only* flashcard set, which was designed for studying, each card displayed a GRE word and its definition on the front and the back was blank. All text was printed in Times New Roman size 24 font, with the GRE words bolded. There were 40 cards per flashcard set, with one card per word-definition pair.

Global Judgments of Learning (gJOL). Participants were asked to predict their performance on the immediate test: "If, in a few minutes, you were shown the definitions you just studied, for what percentage (%) of these definitions are you confident you could remember the corresponding word?" The gJOL was open response from 0% to 100%.

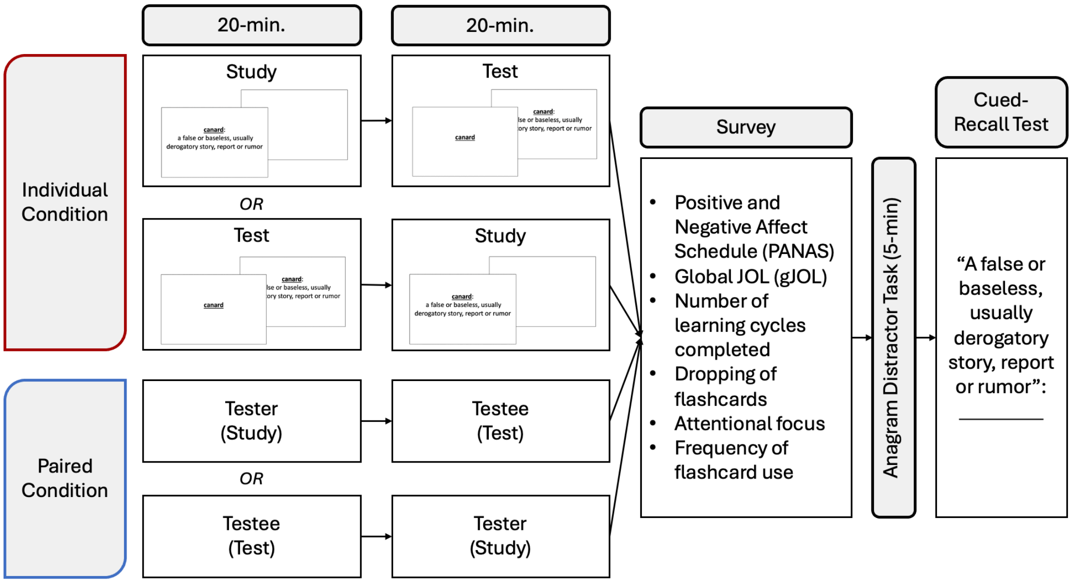
20-min.
20-min.
Individual Condition
Study
canard:
a false or baseless, usually derogatory story, report or rumor
Test
canard
OR
Test
Study
Paired Condition
Tester (Study)
Testee (Test)
OR
Testee (Test)
Tester (Study)
Survey
• Positive and Negative Affect Schedule (PANAS)
• Global JOL (gJOL)
• Number of learning cycles completed
• Dropping of flashcards
• Attentional focus
• Frequency of flashcard use
Anagram Distractor Task (5-min)
Cued-Recall Test
"A false or baseless, usually derogatory story, report or rumor":

Figure 1. Procedure used in Experiment 1.

Procedure

The experiment was run in 2-hr timeslots involving up to four participants each and using three nearly-identical laboratory testing rooms (see Figure 1). All participants were told that they would be learning vocabulary words using flashcards, and all flashcards were randomly shuffled prior to each timeslot. Each Individual learner completed the experiment in a separate testing room, whereas the two Paired learners per timeslot did so in a shared testing room.

The experiment consisted of four phases. All participants first completed a flashcard phase in which they learned and practiced challenging vocabulary-definition pairs. Then they completed a series of survey questions—which included providing a global judgment of learning (gJOL)—, a distractor task, and a final cued-recall test.

Random Assignment and Counterbalancing. Within each timeslot, two participants were randomly assigned to the Paired condition and up to two participants were randomly assigned to the Individual condition. When fewer than four participants signed up for a timeslot, two were assigned to the Paired condition (if possible) and any others were randomly assigned to the Individual condition. The decision to prioritize filling the Paired condition occurred prior to data collection and stemmed from the inherent challenge of bringing two participants together in one timeslot to run that condition (it also maintained random assignment and was consistently applied by all experimenters, thus reducing potential bias). A moderate imbalance in sample size per condition resulted.

Given that using flashcards in pairs entails one person being tested at a time and the other person viewing (i.e., studying) the answers while administering the tests, participants' engagement in studying or testing from the outset of the experiment (before switching activities, which resembles using flashcards across separate study and test phases) was counterbalanced. Thus, task order (i.e., First Learning Activity) was equated across both conditions.

Flashcard Phase

Individual condition. The experimenter seated each participant in a testing room, distributed the study-only or standard flashcard set and, depending on the given set, instructed them to learn the words via studying (i.e., reading) or testing (i.e., retrieval practice). Participants were permitted to cycle through the set as many times as desired and in any order for 20 min. Skipping or dropping flashcards was allowed but not specifically discussed. Afterwards, the flashcard set was replaced (i.e., the standard set was switched for the study-only set, or vice versa) and participants were instructed to use the new set for another 20 min. Hence, equal amounts of time were spent engaged in studying and testing.

Paired condition. Participants were seated face-to-face at a small table on which the standard flashcard set was placed. The experimenter demonstrated how the flashcards were to be used. One participant (the "tester") was to hold up each flashcard with the word-only side facing the other participant (the "testee") and read the word and definition silently as the "testee" attempted to verbally provide a definition. After the "testee" indicated that they had finished their attempt, the "tester" was to reverse the card to reveal the definition. Participants proceeded accordingly for 20 min, during which they were permitted to cycle through the set as many times as desired and in any order. As in the Individual condition, dropping was allowed by either the tester or testee but was not explicitly discussed. Verbal feedback was disallowed to minimize off-task conversations and to ensure that participants in the Paired condition did not have an unfair advantage over participants in the Individual condition due to receiving personalized or elaborative feedback. After 20 min, the experimenter directed participants to switch roles and continue for another 20 min. Thus, equal amounts of time were spent engaged in studying (as the "tester") and testing (as the "testee").

Survey and Distractor Task. After the flashcard phase, participants used desktop computers to (a) answer demographic questions, (b) complete the Positive and Negative Affect Schedule—Short Form (PANAS; Watson et al., 1988), (c) provide a global Judgment of Learning (gJOL), (d) report their level of attentional focus from 0-100%, and (e) answer questions regarding their activities during the flashcard phase and their own flashcard use in everyday learning sessions. The exact wording for each of these items (and the survey items used in the subsequent experiments) is available at https://osf.io/hac38/?view_only=9ff44668b2184f1b8e412452f3a41640. Participants then completed a 5-min distractor task during which they solved anagrams.

Final Cued-Recall Test. During the final cued-recall test, each of the 40 definitions were presented individually and in a random order for 60 s. Participants attempted to type the matching GRE word (similar to [Author] & Rickard, 2017). The experiment concluded afterwards.

Results

Data are available at https://osf.io/hac38/?view_only=9ff44668b2184f1b8e412452f3a41640.

All analyses were conducted using independent samples t-tests with equal variances assumed unless otherwise noted. In all analyses, α was set at .05. The sample sizes per analysis differed slightly in some cases as some participants declined to answer all questions. In a parallel set of analyses not reported here, the effect of First Learning Activity—that is, whether a participant had engaged in studying prior to testing, or vice versa—was not significant on any aspect of measured behavior during the learning or final test phases. Those patterns were unsurprising given that such effects were potentially eclipsed by subsequent cycles of testing and studying. Consequently, all analyses reported here involve data collapsed across First Learning Activity. Parallel analyses that do not do so are included in the supplemental online materials https://osf.io/hac38/?view_only=9ff44668b2184f1b8e412452f3a41640.

Flashcard Phase

Number of Learning Cycles. Participants indicated the number of learning cycles (i.e., practicing through the entire flashcard set) they completed per 20-min period. These responses were summed for a total number of cycles in the entire flashcard phase; if participants indicated an incomplete cycle, then 0.50 was added (this method, albeit somewhat imprecise, was consistently applied across conditions). Individual learners typically completed one more learning cycle ($M = 5.36$, $SD = 1.64$) across the entire flashcard phase than did Paired learners ($M = 4.32$, $SD = 1.44$). This difference was significant, $t\,(150) = 4.16$, $p < .001$, $d = 0.68$, 95% CI [0.55, 1.54].

Dropping of Flashcards. Participants reported whether they had dropped flashcards from study, and if so, why they chose to do so. These data were coded by two independent raters blind to condition (with interrater reliabilities of Cohen's κ = .99 and .85 for if they dropped and why, respectively). A Chi-square test revealed that significantly more Individual learners (53%) dropped flashcards from study than Paired learners (5%), χ^2 (2) = 44.43, $p < .001$. Fifty-eight percent of all participants who dropped a flashcard from study did so because they believed that they had learned the word-definition pair, 32% did so because they deemed the pair too difficult to learn, and 11% did so for other reasons. As only four Paired learners dropped flashcards, formal comparisons of reasons for dropping between conditions were not possible. Those four participants, however, all dropped cards because they deemed materials too difficult to learn, whereas only 24% of Individual learners dropped flashcards for that reason (most did so based on sufficient learning).

Final Cued-Recall Test

Overall Performance. Given the difficulty of the GRE words, we used an accuracy threshold wherein final test responses had to match the actual spelling by ≥ 75% to be counted as correct. Corresponding analyses under strict scoring (i.e., perfect spelling) yielded the same patterns (available in online supplemental materials). Contrary to our hypothesis that Paired flashcard learning would yield higher test performance than Individual flashcard learning, final test performance was not significantly different between the Individual and Paired conditions, t (150) = 1.29, p = .20, d = 0.21, 95% CI [-.03, .13]. This result indicates that recall of the GRE words was no different shortly after individual or paired flashcard learning (Table 1 presents the descriptive statistics for each condition).

Table 1

Cued-Recall Test Performance in Experiments 1-3

Condition	Experiment 1[1]		Experiment 2				Experiment 3			
	Immediate Test		Immediate Test		Delayed Test		Immediate Test		Delayed Test	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Individual	.48	.24	.49	.28	.40	.28	.42	.26	.36	.23
Paired	.43	.23	.44	.25	.35	.24	.37	.24	.34	.22

[1] Only an immediate cued-recall test was administered in Experiment 1

Metacognitive Judgments

Correlations with Final Test Performance. To examine whether there was a significant relationship between participants' own assessment of their learning and their actual test score, we conducted a series of exploratory bivariate correlations relating gJOL and final test performance for both conditions (see Figure 2). Individual learners demonstrated moderate-to-large correlations between their gJOL and final test performance when learning individually, *r* (61) = .59, *p* < .001, as did Paired learners, *r* (86) = .60, *p* < .001. These correlations suggest that participants engaged in appropriate metacognitive monitoring, with participants who reported greater gJOLs tending to score higher on the cued-recall test afterward. Although the magnitude of the relationships between gJOL and test performance was similar between the Paired and Individual conditions, Figure 2 clearly shows that the intercepts of the regression lines between the two conditions (computed by regressing test performance onto gJOL data) differ, prompting further analyses of participants' metacognitive calibration.

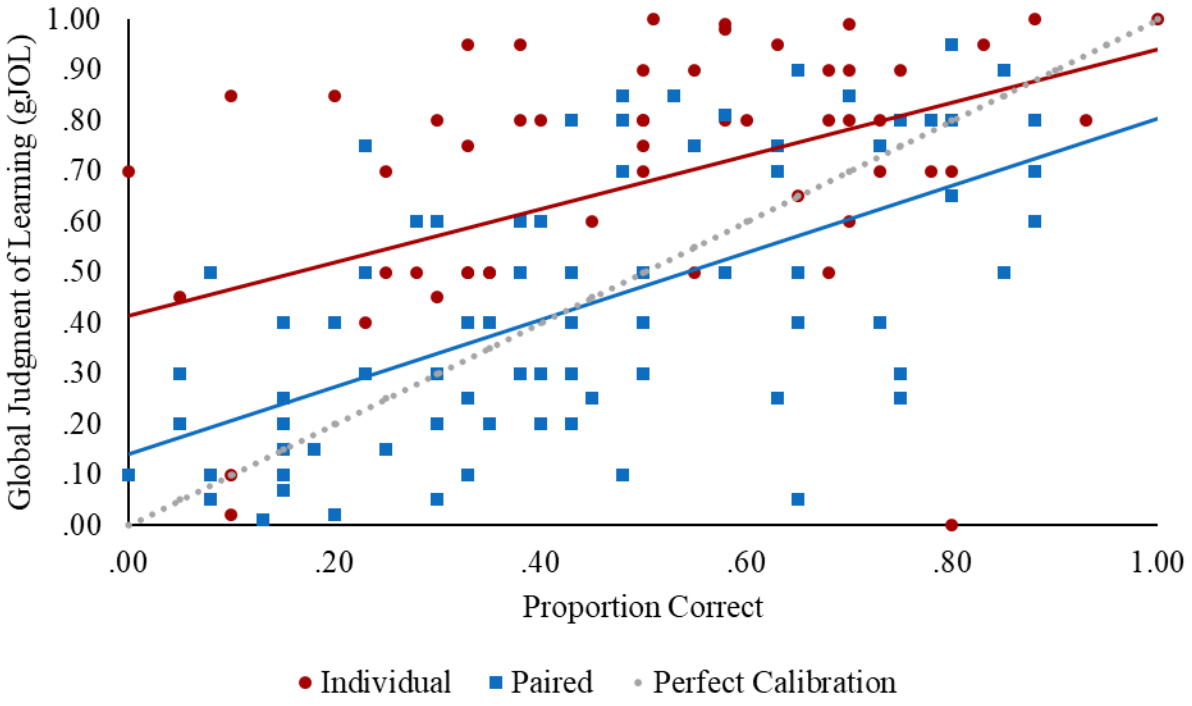

Figure 2. Metacognitive calibration demonstrated by those in the Individual flashcard learning and the Paired flashcard learning conditions. Each panel displays the correlation between final test performance and global judgments of learning (gJOLs). A dotted line represents the hypothetical case of perfect calibration between gJOLs and test scores; crucially, participants in the Individual condition tended to substantially overestimate their final test performance.

Metacognitive Calibration. Metacognitive calibration is a form of absolute metacognitive accuracy; i.e., the extent to which a learner can accurately estimate their learning. Here, metacognitive calibration was measured as the difference in size between their predicted test performance and their actual test performance. We computed metacognitive calibration by subtracting participants' actual test performance from their gJOLs, with positive scores indicating overconfidence and negative scores indicating underconfidence. Unlike the previous analyses, metacognitive calibration provides evidence for the direction of participants' judgment errors (e.g., if one condition tends to exhibit overestimation and the other condition tends to exhibit underestimation, then their average calibration will differ even if their correlation coefficients are similar). Thus, gJOL-test performance correlations and metacognitive calibration scores provide complementary, but distinct, information about learners' metacognitive judgments.

An exploratory independent samples t-test compared Individual and Paired learners' metacognitive calibration scores (Figure 2). Individual learners were overconfident (*M* = .20, *SD* = .22), whereas Paired learners were relatively accurate (*M* = .00, *SD* = .22), *t* (149) = 5.66, *p* < .001, 95% CI [.13, .28].

Positive and Negative Affect

We conducted separate analyses for the positive affect and negative affect subscales of the PANAS. Participants reported comparable positive affect in the Individual (*M* = 26.05, *SD* = 7.54) and Paired (*M* = 26.28, *SD* = 8.12) conditions, *t* (150) = -0.18, *p* = .86, *d* = 0.03, 95% CI [-2.80, 2.32]. Those who studied in pairs, however, reported significantly higher negative affect (*M* = 16.93, *SD* = 6.54) than those who studied individually (*M* = 14.20, *SD* = 3.52), *t* (150) = -3.03, *p* = .003, *d* = 0.50, 95% CI [-4.51, -0.95].

Attentional Focus

Self-reported percentage time focused during the experimental tasks did not significantly differ between the Individual (*M* = 78.4%, *SD* = 16.5%) and Paired (*M* = 80.8%, *SD* = 19.1%) conditions, *t* (150) = -0.78, *p* = .44, *d* = 0.13, 95% CI [-8.18, 3.55].

Experiment 1 Discussion

With respect to effects on memory, the results of Experiment 1 suggest that collaborative and individual practice of difficult vocabulary-definition pairs using flashcards yield comparable test performance after a 5-min delay. These results were contrary to our predictions. It is, however, possible that the delay between the learning and test phases was not long enough to observe the benefits of collaborative practice. In line with the framework of desirable difficulties (Bjork, 1994), the benefits of more challenging but potentially beneficial learning activities are often observed at a delay (e.g., Roediger & Karpicke, 2006). It is thus possible that the positive effects of more effortful or complete retrieval encouraged by Paired flashcard practice testing may emerge on a delayed test.

There were, however, some benefits of collaborative practice that may be particularly meaningful for learners engaging in self-regulated study. Paired learners were far less likely to drop cards from study than Individual learners. Moreover, prediction errors of test performance from Paired learners did not exhibit a systematic bias whereas Individual learners on average overestimated their learning by approximately 20%. Possibly, these two results are related: If Paired learners were more metacognitively accurate during the flashcard phase of the study than Individual learners, they may have been less likely to prematurely drop cards from study. Vice versa, if Paired learners were less likely to drop cards from study than Individual learners for other reasons (perhaps because their partner was holding the flashcard deck, adding friction to the drop decision, or because they were instructed to limit discussion with their partner during the flashcard learning phase), their metacognitive judgments may have benefited from relatively equal time spent on each vocabulary term. In our view, it is crucial to ascertain whether the metacognitive calibration benefit in the Paired condition is merely a result of lower rates of dropping flashcards, which we address in the second experiment.

Finally, the effect of First Learning Activity (i.e., whether a participant had engaged in studying prior to testing, or vice versa) did not significantly impact any aspect of behavior during the learning or final test phases, possibly because any such effects were eclipsed by subsequent cycles of testing and studying. From an ecological validity standpoint, requiring that students first study and then test themselves (or vice versa) seems at odds with the common view of flashcards as a retrieval practice tool. Additionally, the effects of collaboration on learning often have been examined within the context of testing on previously studied content, and are therefore often compared to individual testing (e.g., Barber et al., 2010; Gilley & Clarkston, 2014; [Author], 2023). It may therefore be more appropriate to compare the effects of paired flashcard practice to the effects of individual retrieval practice with flashcards.

Experiment 2

Experiment 2 again compared the effects of individual versus paired flashcard use on learning. To examine if there might be a benefit of paired practice over individual practice for long-term learning, a 24-hr delayed test was added. We chose a 24-hr delay because delays of one day or longer tend to yield stronger testing effects than delays occurring within the same day (Rowland, 2014). To rule out the possibility that Paired learners are more metacognitively accurate simply due to lower rates of dropping flashcards from study, dropping flashcards from study was explicitly prohibited in Experiment 2. Additionally, to increase participants' ease in interacting with one another in the Paired condition, a brief icebreaker activity prior to the flashcard portion was incorporated. Finally, as the effect of First Learning Activity (i.e., whether a participant had engaged in studying prior to testing, or vice versa) did not significantly impact any aspect of behavior during the flashcard phase or final test performance, First Learning Activity was removed as a factor and a period of initial study of the vocabulary-definition pairs prior to the flashcard phase was added.

Method

Experiment 2 was not preregistered.

Participants

One hundred and forty-one participants were included in this study (Individual: $n = 78$, Paired: $n = 63$). An additional thirty participants were recruited for this study but were excluded due to technical issues or experimenter error ($n = 4$), for failing to follow instructions ($n = 11$;

e.g., did not practice test the entire time), or for reporting that they dropped flashcards from study during that phase (*n* = 15).

Of the 141 participants in the final sample, all reported an immediate gJOL and 84 (59.6%) offered a delayed gJOL. Six (4.3%) participants did not report a delayed JOL because they did not complete the delayed test portion of the study. An additional 50 participants (35.5%) took the delayed test but chose not to offer a gJOL (in accordance with our IRB protocol, participants were not required to answer every question)[2]. Finally, one participant (0.7%) mistakenly reported that they were participating in Session 1 (rather than Session 2) of the study when inputting their information into the delayed test link such that the page prompting participants for a gJOL did not appear.

Design

Experiment 2 employed a 2 x 2 mixed factorial design with Condition (Individual or Paired) as the between-subjects factor and Test Delay (5-min or 24-hr) as the within-subjects factor. The 40 word-definition pairs used in this study were divided into two sets of 20 pairs (i.e., Set A and Set B): One set was used for the immediate test and one set was used for the 24-hr delayed test, counterbalanced across participants by time slot. Although First Learning Activity was not manipulated for the Individual condition in this experiment and was not included in any subsequent statistical models, the nature of the Paired condition required that one member of the pair act as the tester first and one member of the pair act as the testee first.

Materials

The materials used in Experiment 2 were identical to the materials used in Experiment 1 except that only the standard flashcard set was used. Given a change in the software used to run the final test portion of the study (more details below), the cued-recall test was scored by two independent raters. Interrater reliability for all cued-recall test items was adequate (Cohen's *κ's* = .84 – 1.00). All disagreements were resolved by a third rater. Two gJOLs were used in this experiment, each open response from 0% to 100%. The first gJOL was to predict performance on the immediate test, "If, in a few minutes, you were shown the definitions you just studied, for what percentage (%) of these definitions are you confident you could remember the corresponding word?" and the second gJOL was to predict performance on the delayed test (and was administered immediately prior to the delayed test): "If, in a few minutes, you were shown the definitions you studied in Part 1 (yesterday using flashcards in the psychology lab), for what percentage (%) of these definitions are you confident you could remember the corresponding word?"

[2] It is not clear why so many participants chose not to report a delayed gJOL. It is possible that, as participants were not told that the delayed portion of the study would include a test, they were surprised by the prompt for a gJOL and were unsure how to respond.

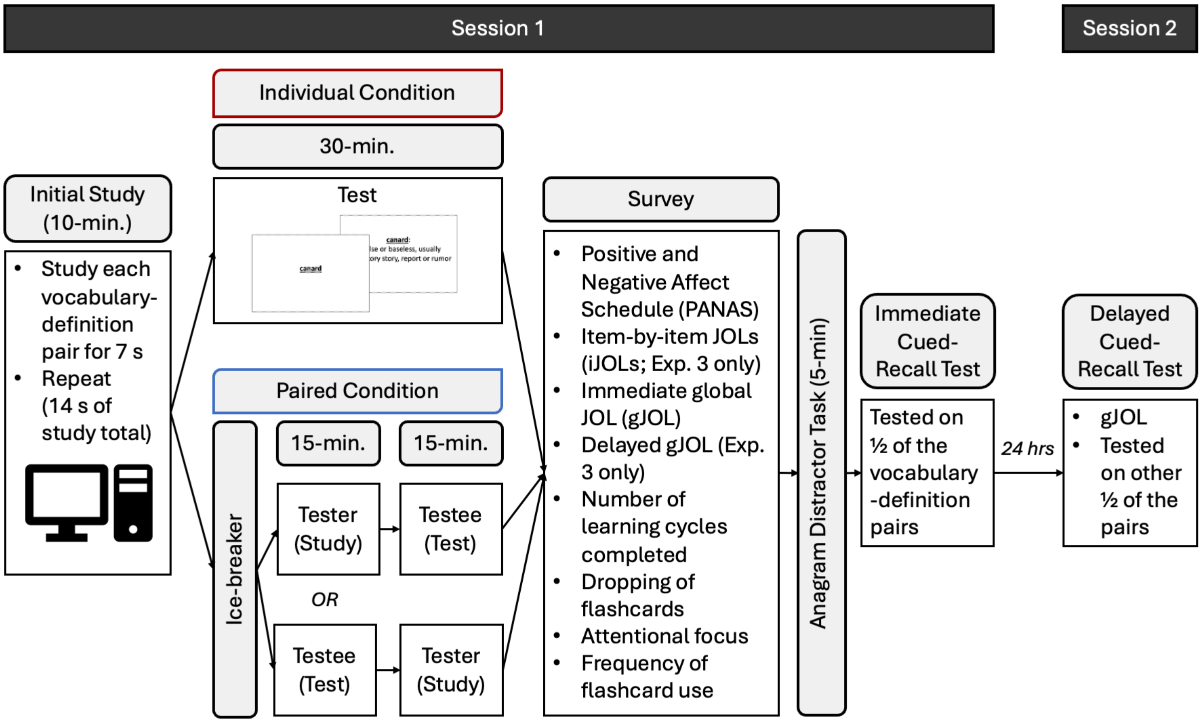
Session 1
Session 2
Initial Study (10-min.)
Study each vocabulary-definition pair for 7 s
Repeat (14 s of study total)
Individual Condition
30-min.
Test
canard
canard:
lse or baseless, usually
tory story, report or rumor
Paired Condition
Ice-breaker
15-min.
15-min.
Tester (Study)
Testee (Test)
OR
Testee (Test)
Tester (Study)
Survey
Positive and Negative Affect Schedule (PANAS)
Item-by-item JOLs (iJOLs; Exp. 3 only)
Immediate global JOL (gJOL)
Delayed gJOL (Exp. 3 only)
Number of learning cycles completed
Dropping of flashcards
Attentional focus
Frequency of flashcard use
Anagram Distractor Task (5-min)
Immediate Cued-Recall Test
Tested on ½ of the vocabulary -definition pairs
24 hrs
Delayed Cued-Recall Test
gJOL
Tested on other ½ of the pairs

Figure 3. Procedure of Experiments 2 and 3.

Procedure

Aside from the following changes listed below, the procedure of Experiment 2 was the same as Experiment 1 (see Figure 3).

The experiment was run in two sessions spaced 24 hrs apart. Aside from the flashcard portion, all phases of the study were run using Qualtrics (https://www.qualtrics.com/). The first session was run in 90-min timeslots involving up to six participants each and using four nearly-identical laboratory testing rooms. The session began with an initial study phase conducted individually on a desktop computer. During the initial study phase, participants studied each vocabulary-definition pair for seven seconds one-at-a-time in a random order. They did this twice, studying each vocabulary-definition pair for a total of 14 s, for an overall study time of approximately 10 min.

Flashcard Phase. Given that learners received approximately 10 min of initial study, the flashcard portion of the study was shortened to two 15-min periods (such that total time spent learning the materials remained approximately 40 min). During the flashcard phase, all participants solely used the standard flashcard set.

Individual condition. Participants were instructed to test themselves during the entirety of the flashcard phase. They were told that the experimenter would check in on them after 15 min. Dropping of flashcards was prohibited.

Paired condition. Given the elevated negative affect reported by Paired learners in

Experiment 1, two changes were made to make learners feel more comfortable during the study and to allow for behaviors that students might engage in when collaboratively practice testing in daily life. First, between the initial study phase and the flashcard phase, Paired learners were given two min to complete an icebreaker activity. During this activity, participants were encouraged to introduce themselves to their partner and to converse with them to find one thing that they had in common (e.g., favorite color). Second, although explanations and clarifications were still disallowed during the flashcard phase to avoid an unfair benefit to the Paired condition, participants were told that they could provide brief comments (e.g., good job).

Survey and Distractor Task. As dropping flashcards from study was explicitly prohibited, participants were asked whether they dropped flashcards from study only as a compliance check; the question about why they dropped flashcards from study was removed.

Final Cued-Recall Test

Immediate (5-min). Twenty definitions were presented. Prior to completing the test, participants reported a gJOL (as they did in Experiment 1).

Delayed (24-hr). The morning after Session 1, participants were emailed the test link and were told that they had until 11:59pm that day to complete the test on their own laptop or desktop computer in a quiet, distraction-free place. Prior to completing the test, participants again reported a gJOL.

Results

Flashcard Phase

Number of Practice Cycles. Participants indicated the number of practice cycles (i.e., practicing through the entire flashcard set) they completed per 15-min period of the flashcard phase. These two numbers were again summed to compute a total number of practice cycles. Unlike in Experiment 1, Individual learners ($M = 4.29$, $SD = 1.75$) and Paired learners ($M = 3.95$, $SD = 1.45$) completed about the same number of practice cycles through the flashcard deck, t (139) = 1.22, $p = .22$, $d = .21$, 95% CI [-.21, .88].

Final Cued-Recall Test

Overall Performance. To examine the effect of individual versus paired flashcard practice on learning, a 2 x 2 ANOVA was conducted with Condition (Individual or Paired) as the between-subjects factor, Test Delay (5-min or 24-hr) as the within-subjects factor, and test performance as the dependent variable. Six participants did not complete the delayed test[3] and were therefore excluded from this analysis, leaving 75 Individual and 60 Paired learners in the analysis.

Immediate test scores were higher than delayed test scores, suggesting that forgetting occurred during the 24-hr delay, F (1, 133) = 41.80, $p < .001$, $\eta_p^2 = .24$. Replicating the result of Experiment 1, Paired and Individual learners overall demonstrated similar test performance, F

[3] Additionally, seven participants completed the delayed test late (but within 48-hrs of the first session of the study). A parallel analysis available in the online supplemental materials indicated that excluding these participants does not change the pattern of results.

(1, 133) = 1.44, p = .23, η_p^2 = .01[4]. The nonsignificant Condition x Test Delay interaction suggests that this similarity did not change between the immediate test and the delayed test, F (1, 133) = 0.003, p = .96, η_p^2 < .001. Performance on both the immediate and delayed tests, however, were numerically lower in the Paired condition (as was the case in Experiment 1), which suggests that there may be a modest reduction in the efficacy of learning (or the rate of learning) that occurs when using flashcards in pairs versus individually.

Equivalence Test. To examine if the null effect obtained in Experiment 2 was equivalent with the null effect obtained in Experiment 1, we conducted a two one-sided tests (TOST) procedure to test for equivalence (Lakens et al., 2018) using the TOSTER package in R (Caldwell, 2022; Lakens, 2017). First, we set the smallest effect size of interest. Based on the group sizes from Experiment 1 with α = .05, we used G*Power (Faul et al., 2007) to determine that the smallest effect size Experiment 1 had 80% power to detect was d = 0.47. For that reason, we set the lower equivalence bound to d = -0.47 and the upper equivalence bound to d = 0.47. We then used the data obtained in Experiment 2 to run two Welch's one-sided t-tests. The test for the upper bound was significant, t (137.57) = -1.67, p = .048, as was the test for the lower bound, t (137.57) = 3.91, p < .001. These significant t-tests indicate that we can reject the null hypothesis that the true effect was smaller than d = -0.47 or larger than d = 0.47; i.e., that the effect size falls within the equivalence range. Thus, we can conclude the null effect obtained in Experiment 2 is equivalent to the null effect obtained in Experiment 1.

Metacognitive Judgments

Correlations with Final Test Performance. To examine whether there was a significant relationship between participants' own assessments of their learning and their actual test score, a series of bivariate correlations related gJOL and final test performance for both conditions and for both the Immediate and Delayed tests (see Figure 4).

As in Experiment 1, for the immediate test, participants demonstrated moderate-to-large correlations between their gJOL and their actual final test performance after learning individually, r (76) = .51, p < .001, and after learning with a partner, r (61) = .57, p < .001. These correlations were somewhat reduced when examining the relationship between delayed gJOLs and performance on the delayed test, Individual: r (51) = .43, p = .001; Paired: r (29) = .36, p = .047.

Metacognitive Calibration. To include the maximum number of participants in the analysis of metacognitive calibration at immediate test, participants' metacognitive calibration was analyzed using separate independent samples t-tests for the Immediate and Delayed tests.

Immediate Test. Again replicating the results of Experiment 1, Individual learners (M = .18, SD = .27) were more overconfident than Paired learners (M = .08, SD = .26), t (139) = 2.14, p = .034, d = 0.36, 95% CI [.007, .19].

Delayed Test. In contrast to the results for the immediate test, both Individual learners (M = -.05, SD = .27) and Paired learners (M = -.02, SD = .26) were well-calibrated, if slightly underconfident, t (82) = -0.39, p = .70, d = -.09, 95% CI [-.14, .10].

[4] An independent samples t-test examining the effect of Condition at immediate test only (n = 141) obtained the same result.

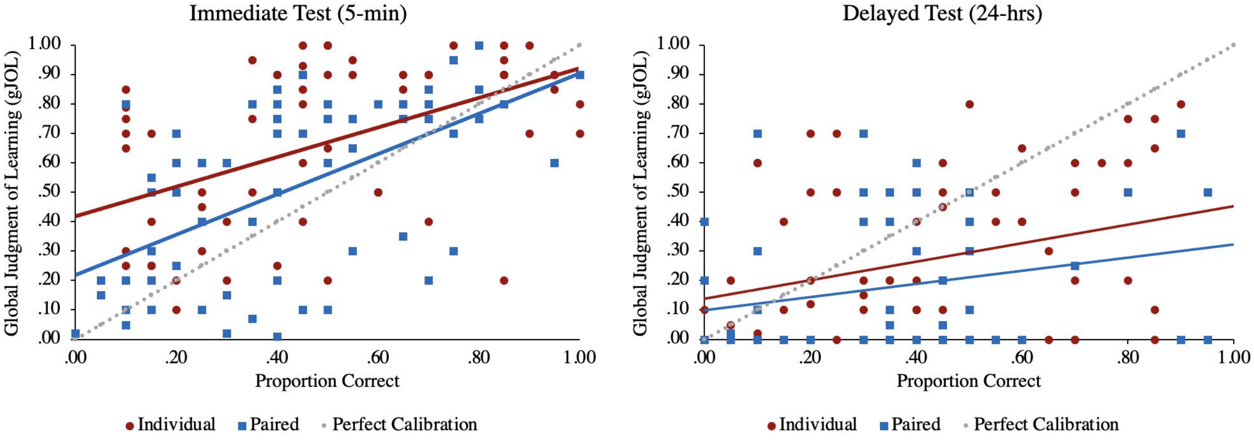

Figure 4. *Metacognitive calibration for the Immediate test (left panel) and the Delayed test (right panel). Each panel displays the correlation between test performance and global judgments of learning (gJOLs). The red and blue lines represent least squares regression fits to Individual and Paired data, respectively. A dotted line represents the hypothetical case of perfect calibration between gJOLs and test scores; again, participants in the Individual condition tended to substantially overestimate their future cued-recall test performance for the immediate test but this tendency did not extend to the delayed test.*

Positive and Negative Affect

As in Experiment 1, there was no difference in self-reported positive affect by Individual learners (*M* = 26.95, *SD* = 8.11) and Paired learners (*M* = 27.73, *SD* = 8.00), *t* (139) = -0.57, *p* = .57, *d* = -0.10, 95% CI [-3.48, 1.92]. However, in contrast to Experiment 1, self-reported negative affect also did not differ between Individual learners (*M* = 14.73, *SD* = 4.41) and Paired learners (*M* = 15.54, *SD* = 3.99), *t* (130) = -1.13, *p* = .26, *d* = -0.19, 95% CI [-2.22, 6.61]. It is possible that the inclusion of the icebreaker activity and the eased restrictions on verbal exchanges led to less negative affect for the Paired condition in Experiment 2.

Attentional Focus

As in Experiment 1, self-reported percentage time focused during the experimental tasks did not significantly differ between the Individual (*M* = 86.3%, *SD* = 14.6%) and Paired (*M* = 87.8%, *SD* = 12.8%) learning conditions, *t* (139) = -0.61, *p* = .54, *d* = -0.10, 95% CI [-6.05, 3.20].

Experiment 2 Discussion

Experiment 2 replicated and extended the two primary findings of Experiment 1. First, immediate cued-recall test performance was similar across those who used flashcards individually and those who used flashcards collaboratively (combining Experiment 1 and 2 data together shows this result is highly similar across the two experiments; analysis available in the online supplemental materials). This similarity was then maintained for the delayed (24-hr) test. This result does not align with the suggestion that paired flashcard learning might encourage more effortful retrieval and thus act as a "desirable difficulty" with its benefits emerging after a delay (as is sometimes the case when comparing the effects of more effortful versus less effortful learning strategies, e.g., testing versus restudy, Roediger & Karpicke, 2006). Second, participants in the Individual condition again demonstrated overconfidence in their learning for the immediate test whereas participants in the Paired condition again demonstrated relatively accurate metacognitive judgments. This overconfidence occurred even after dropping of flashcards was prohibited in Experiment 2, suggesting that the miscalibration observed in the Individual condition cannot simply be attributed to a lack of exposure to dropped items. At a delay, however, participants' metacognitive judgments were similarly well-calibrated. This improved metacognitive calibration at a delay is in line with prior work demonstrating that delayed JOLs tend to be more accurate than JOLs made immediately after learning (e.g., Nelson & Dunlosky, 1991).

In Experiment 2, we did not find evidence that learner affect, number of learning cycles, or level of focus differed between the two conditions. Consequently, in Experiment 3 we sought to identify a potential explanation for the metacognitive benefits of paired flashcard learning. Specifically, we investigated whether the use of overt retrieval in paired flashcard learning might drive its benefit. Perhaps requiring participants to overtly retrieve—regardless of whether another person is present or not—offers the learner more concrete evidence of their learning, informing more accurate metacognitive judgments.

Experiment 3

In Experiment 3 we examined whether learners were more metacognitively accurate following paired flashcard learning as compared to individual flashcard learning because learners in the Paired condition were required to overtly retrieve. Overt retrieval, in contrast to covert retrieval, might establish natural accountability for one's responses during retrieval practice that could enhance the benefits of testing. For example, Sumeracki and Castillo (2022) observed a testing effect for students in a classroom that overtly retrieved during practice testing but not for students that covertly retrieved. However, overt and covert retrieval practice both produced a testing effect when students were first informed that one of them would be called on randomly by the teacher. Beyond learning, this accountability may extend benefits to metacognitive judgments by encouraging greater completeness of one's retrieved answers and thus enhancing the quality of evidence available when making these judgments (Tauber et al., 2018). In the present experiment, we instructed participants in the Individual condition to retrieve out loud during the flashcard phase. If overt retrieval was responsible for the benefits of Paired flashcard learning, then we would expect to observe no difference in metacognitive calibration between the two conditions.

Additionally, participants in Experiment 3 reported both global JOLs, as they did in Experiments 1 and 2, and item-level JOLs (iJOLs); i.e., participants predicted both their overall performance and their likelihood of retrieving each vocabulary-definition pair. Item-level JOLs are commonly used in studies of metacognition to measure individuals' relative accuracy (*metacognitive resolution*; i.e., their ability to discriminate between information that will or will not be remembered; Rhodes, 2016; Vuorre & Metcalfe, 2022). Until this point, we had assessed participants' absolute accuracy (i.e., *metacognitive calibration*), measuring the difference between participants' average/overall metacognitive judgments and their actual learning outcomes. Resolution and calibration reflect different dimensions of metacognition (Rhodes, 2016) and can thus at times offer divergent results that offer insight into metacognitive processes; for example, in studies of age-related differences in metacognition (Siegel & Castel, 2019). Further, the rate of dropping flashcards from study in the Individual condition in Experiment 1 suggests that learners at least sometimes engage in spontaneous item-level judgments during flashcard learning. Thus, we incorporated both types of metacognitive judgments in Experiment 3.

Method

Experiment 3 was preregistered at https://aspredicted.org/4D9_DFP.

Participants

Four-hundred and five participants were included in this study (Individual: n = 187, Paired: n = 218). Eighty participants were recruited from the Psychology subject pool at the same large public research university as Experiments 1 and 2, and 325 participants were recruited from the Psychology subject pool at a similar large public research university in the same region. An additional 120 participants were recruited for this study but were excluded based on our preregistered criteria: dropping flashcards (n = 37), using their phone to complete the study (n = 8), failure to follow instructions (e.g., reporting that they did not retrieve during the flashcard portion of the study) (n = 67), experimenter error (n = 7), and technical issues (n = 1). We collected greater than our preregistered number of participants in this study because of an error in the set-up of the study that was not identified until midway through data collection. A subset of participants erroneously received the same set of vocabulary words to test on during both the immediate and delayed tests, rendering their delayed test scores unusable. To ensure that we were adequately powered for all our planned analyses, we collected additional data until we met our preregistered sample size for participants with usable delayed test scores (n = 210).

Design

Experiment 3 employed a 2 x 2 mixed factorial design with Condition (Individual or Paired) as the between-subjects factor and Test Delay (5-min or 24-hr) as the within-subjects factor.

Materials

The materials used in Experiment 3 were identical to the materials used in Experiment 2. A subset of the cued-recall test responses (n = 1320 responses) was scored by two independent raters. Interrater reliability for all cued-recall test items was adequate (Cohen's κ*'s* = .68 – 1.00). All disagreements were resolved by discussion and the remaining cued-recall test responses were scored by a single rater.

Procedure

Aside from the following changes listed below, the procedure of Experiment 3 was the same as in Experiment 2. All participants completed the study in nearly-identical laboratory testing rooms. Participants from one university completed the study on desktop computers and participants from the other university completed the study on their own laptop computers.

Flashcard Phase. Participants in the Individual condition were instructed to test themselves out loud during the entirety of the flashcard phase. To ensure compliance, an audio monitor was placed in the center of the laboratory testing room and the receiver was placed in a separate area with the experimenter. Participants were informed that the audio monitor only transmitted sound to the experimenter (i.e., it did not record their audio). If the experimenter noted that the participant had stopped testing themselves aloud for more than two min they checked in on the participant and reminded them to test themselves aloud.

Survey. There were two additions made to the survey that was administered after the flashcard learning phase in Session 1. The first was an additional global JOL (i.e., Session 1 gJOL: Delayed Test) that queried participants about how well they believed they would do on a delayed test: "If tomorrow you were shown the definitions you just studied, what percentage (%) of these definitions are you confident you could remember the corresponding word?" This gJOL was an exploratory item to investigate if participants in each condition might differ in their tendency to predict their long-term learning, and to ensure alignment between the gJOLs and the item-by-item JOLs that participants also gave (described below). For clarity, we now refer to the gJOL for the delayed test administered in this and the previous experiment as "Session 2 gJOL: Delayed Test).

The second addition was the inclusion of item-level JOLs (iJOLs). These iJOLs were placed at the beginning of the survey. Participants were shown the vocabulary-definition pairs one-at-a-time in a random order and asked to rate the likelihood that they would be able to type the correct vocabulary word if shown only the definition from 0% (will not be able to) to 100% (certainly will be able to). For each vocabulary-definition pair, they gave two ratings using a slider scale: one for if shown the definition "in a few minutes" and one for if shown the definition "tomorrow." Participants were instructed to report their initial judgment upon seeing the vocabulary-definition pair. If they did not report their iJOLs for a given pair within 10 s, a message appeared on the screen encouraging them to respond.

Results[5]

Flashcard Phase

Number of Practice Cycles. Unlike Experiment 2 (but like Experiment 1), Individual learners (*M* = 4.55, *SD* = 1.82) completed about one more practice cycle through the flashcard set than Paired learners (*M* = 3.82, *SD* = 1.38), *t* (401) = 4.53, *p* < .001, *d* = 0.45, 95% CI [0.41, 1.04].

Final Cued-Recall Test

Overall Performance

We analyzed final cued-recall test performance using a 2 (Condition: Individual or Paired) x 2 (Test Delay: Immediate or Delayed) mixed ANOVA, with Condition as a between-subjects factor and Test Delay as the within-subjects factor. This analysis only included participants who had both usable immediate and delayed test scores. Unlike in the previous

[5] The number of participants included in each analysis varies slightly because of missing data or because a participant chose not to answer every question.

experiments, the Test Delay x Condition interaction was significant, F (1, 208) = 5.11, p = .02, η_p^2 = .02. The pattern of the interaction suggested that the Individual condition's rate of forgetting was greater than the Paired condition's. Since the interaction was significant, follow-up independent samples t-tests were conducted. The effect of Condition at the immediate test was significant, t (208) = 2.20, p = .03, d = 0.31, 95% CI [.007, .14]. The Individual condition scored significantly higher (M = .43, SD = .25) than the Paired condition (M = .36, SD = .22) at the immediate test. The effect of Condition at the delayed test was nonsignificant, t (208) = 0.46, p = .64, d = 0.07, 95% CI [-.05, .08]. The Individual (M = .36, SD = .23) and Paired (M = .34, SD = .22) conditions scored similarly on the delayed test[6].

Immediate Test Only

Given that a number of participants only had usable test data for the immediate test, we ran an additional independent samples t-test comparing immediate test scores for all Individual and Paired participants who had a usable immediate test score (i.e., regardless of whether they had usable delayed test data). In this analysis, which included an additional 193 participants, the difference between the Individual (M = .42, SD = .26) and Paired (M = .37, SD = .24) conditions' immediate test scores was nonsignificant, t (401) = 1.69, p = .09, d = 0.17, 95% CI [-.007, .09], although numerically higher for the Individual condition.

Metacognitive Judgments

Correlations with Final Test Performance

Individual learners demonstrated moderate correlations between their gJOL and immediate final test performance when learning individually, r (184) = .51, p < .001, as when learning in pairs, r (213) = .46, p < .001 (see Figure 4). The strength of these correlations was maintained when examining the relationship between Session 2 gJOL: Delayed Test judgments and performance on the delayed test for participants in the Individual condition, r (88) = .59, p < .001, but was somewhat reduced for participants in the Paired condition, r (118) = .32, p < .001.

New to Experiment 3 was a gJOL in Session 1 asking participants to predict their test performance if given a test on the vocabulary-definition pairs the next day (i.e., Session 1 gJOL: Delayed Test). For this gJOL, participants' judgments were less strongly related to actual test performance than participants' immediate test gJOLs: Individual: r (88) = .36, p < .001, Paired: r (118) = .35, p < .001. Possible explanations for these differences in the magnitude of the gJOL-test performance correlations will be explored in the following sections.

Metacognitive Calibration

Metacognitive calibration was again calculated by subtracting participants' actual test performance from their predicted performance (i.e., their gJOL). To include the maximum number of participants in the analysis of metacognitive calibration at immediate test, participants' metacognitive calibration was analyzed using separate independent samples t-tests for the Immediate and Delayed tests. A Bonferroni correction for multiple comparisons was used such that the standard for significance was p < .017.

[6] Fifteen participants completed the delayed test late (but within 48-hrs of the first session of the study). A parallel analysis indicated that excluding these participants did not change the pattern of results (see online supplemental materials located at https://osf.io/hac38/?view_only=9ff44668b2184f1b8e412452f3a41640).

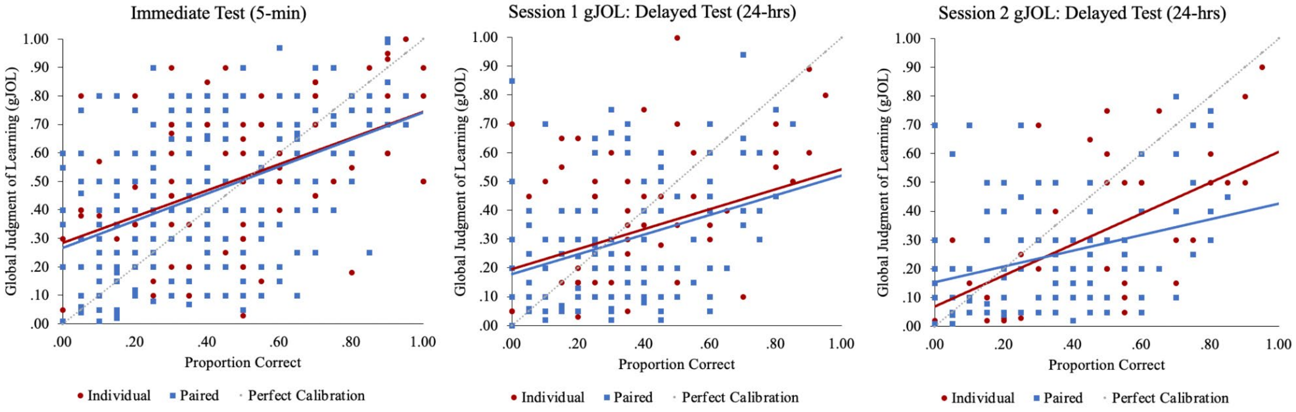

Figure 5. *Metacognitive calibration for the Immediate test and immediate global JOL (gJOL) (left panel), the Delayed test and Delayed test gJOL administered after the flashcard learning phase in Session 1 (middle panel), and the Delayed test and Delayed gJOL administered in Session 2 immediately prior to the Delayed test (right panel). Each panel displays the correlation between test performance and gJOLs. The red and blue lines represent least squares regression fits to Individual and Paired data, respectively. The dotted lines represent the hypothetical case of perfect calibration between gJOLs and test scores.*

Immediate Test. Unlike in Experiments 1 and 2, Individual learners (*M* = .06, *SD* = .25) and Paired learners (*M* = .07, *SD* = .25), had calibration scores close to 0 (i.e., they were relatively accurate, if slightly overconfident) and these scores did not significantly differ from one another, *t* (399) = -0.41, *p* = .69, *d* = -0.04, 95% CI [-.06, .04].

Session 2 gJOL: Delayed Test. Similar to the results of Experiments 1 and 2 (and aligned with participants' calibration at immediate test), both Individual learners (*M* = -.10, *SD* = .20) and Paired learners (*M* = -.10, *SD* = .24) were well-calibrated, if slightly underconfident, *t* (208) = -0.02, *p* = .99, *d* = -.002, 95% CI [-.06, .06].

Session 1 gJOL: Delayed Test. Like the other calibration scores obtained in Experiment 3, participants in the Individual (*M* = -.04, *SD* = .26) and Paired (*M* = -.05, *SD* = .26) conditions were both slightly underconfident, *t* (208) = 0.26, *p* = .80, *d* = 0.04, 95% CI [-.06, .08].

We further compared participants' calibration scores between the two delayed test gJOLs using an ANOVA with Delayed Test gJOL Timing (Session 1 or Session 2) as the within-subjects factor and Condition (Individual or Paired) as the between-subjects factor. The interaction between these factors was nonsignificant, *F* (1, 208) = 0.19, *p* = .67, η_p^2 = .001, as was the main effect of Condition, *F* (1, 208) = 0.02, *p* = .89, η_p^2 < .001. The main effect of Delayed Test gJOL Timing, however, was significant, *F* (1, 208) = 21.87, *p* < .001, η_p^2 = .10. Session 1 Delayed Test gJOLs (*M* = -.04, *SD* = .26) were significantly more accurate than Session 2 Delayed Test gJOLs (*M* = -.10, *SD* = .23).

Metacognitive Resolution

Participants' metacognitive resolution was determined by associating participants' iJOLs with their performance on that item (i.e., whether they got the item correct or incorrect on the cued-recall test) to compute a Kruskal-Goodman gamma correlation (Nelson, 1984). Gamma correlations offer insight into participants' ability to discriminate between vocabulary-definition pairs that ultimately were remembered and those which were not. Four participants had their data excluded from the analyses because they were missing five or more iJOLs (this exclusion criterion was preregistered) and 17 additional participants were excluded from the analyses because of a lack of variation in test scores (i.e., they scored 0% or 100%).

A 2 (Condition: Individual or Paired) x 2 (Test Delay: Immediate or Delayed) mixed ANOVA was conducted with Condition as the between-subjects factor, Test Delay as the within-subjects factor, and the participant's gamma correlation as the dependent variable. Overall, the mean gamma correlation for each condition at each test delay suggests that students' iJOLs were moderately positively associated with their actual test performance. The Condition x Test Delay interaction was nonsignificant, F (1, 187) = 0.73, p = .40, $\eta_p{}^2$ = .004. The main effect of Condition was also nonsignificant; participants in the Individual (M = .49, SD = .49) and Paired (M = .51, SD = .49) conditions had gamma correlations of similar magnitude, F (1, 187) = 0.16, p = .69, $\eta_p{}^2$ = .001.[7] There was, however, a significant main effect of Test Delay such that participants' gamma correlations were higher for the delayed test (M = .55, SD = .31) than for the immediate test (M = .46, SD = .34), F (1, 187) = 8.93, p = .003, $\eta_p{}^2$ = .05.

Positive and Negative Affect

As in Experiment 2, participants in the Individual (M = 25.44, SD = 7.89) and Paired (M = 26.28, SD = 8.05) conditions reported similar levels of positive affect, t (401) = -1.06, p = .29, d = -0.11, 95% CI [-2.41, 0.73]. Likewise, participants in the Individual (M = 15.84, SD = 5.35) and Paired (M = 16.03, SD = 5.77) conditions reported similar levels of negative affect, t (401) = -0.35, p = .73, d = -0.04, 95% CI [-1.29, 0.90].

Attentional Focus

As in Experiments 1 and 2, self-reported percentage time focused during the experimental tasks in Experiment 3 did not significantly differ between the Individual (M = 88.5%, SD = 14.0%) and Paired (M = 86.4%, SD = 13.1%) conditions, t (401) = 1.59, p = .11, d = 0.16, 95% CI [-0.51, 4.80].

Self-Reported Flashcard Use in Experiments 1-3

Table 2 summarizes data on participants' self-reported use of flashcards for exam preparation in daily life. Most students reported using flashcards at least sometimes when studying. When studying with friends, less than half of students reported using flashcards; even if they did use flashcards when studying with friends, they did so infrequently. Overall, students' self-reported flashcard practices suggest that, while they do commonly use flashcards when studying in daily life, they are far more likely to use flashcards when studying alone versus when studying with others.

[7] Examining the gamma correlations for all participants who had an immediate test score yielded the same result: t (379) = -0.33, p = .74, d = -0.03, 95% CI [-.09, .06].

Table 2. Frequency of Self-Reported Flashcard Use When Preparing for Exams

Frequency	When studying generally						When studying with a partner					
	Exp. 1		Exp. 2		Exp. 3		Exp. 1		Exp. 2		Exp. 3	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Never	19	12.5	18	12.8	55	13.6	25	16.4	36	35.5	108	26.8
Almost never	37	24.3	41	29.1	139	34.5	54	35.5	52	36.9	148	36.7
Sometimes	65	42.8	72	51.1	156	38.7	59	38.8	50	35.5	122	30.3
Almost every time	26	17.1	7	5.0	42	10.4	9	5.9	3	2.1	23	5.7
Every time	5	3.3	3	2.1	11	2.7	5	3.3	0	0.0	2	0.5
Total	152		141		403		152		141		403	

General Discussion

Across three experiments, using flashcards to learn with a partner did not yield greater learning compared to using flashcards alone; in fact, in Experiment 3, the Individual condition outperformed the Paired condition at the immediate test. Despite our expectation that collaboration might encourage more effortful retrieval and thus promote long-term learning, Individual and Paired flashcard use yielded learning that was not statistically different when assessed at a 24-hr delay in Experiments 2-3. Although performance did not differ significantly between the two learning conditions, we did observe two advantages of flashcard-based retrieval practice with a partner as opposed to individual retrieval practice. First, when dropping was neither explicitly allowed nor disallowed, Paired learners were far less likely to drop cards from study than Individual learners. Second, there was a striking metacognitive benefit to Paired learning observed in Experiments 1-2: Whereas Individual learners were often overconfident—overestimating learning by approximately 20% in both experiments—Paired learners exhibited more accurate judgments immediately after they had finished learning with flashcards. Instructing Individual learners to overtly retrieve in Experiment 3, however, mitigated this overconfidence. Together, these findings suggest that paired flashcard practice can offer metacognitive benefits that may be important for those using flashcards during self-regulated learning and offers evidence of a potential mechanism for these effects: the facilitation of overt retrieval.

Learning Efficiency of Individual versus Paired Flashcard Learning

In this set of experiments, we asked learners to self-report the number of learning cycles (i.e., the number of times they were able to get through all the cards in the flashcard set). We did so because collaborative learning activities can take longer than individual learning activities (Johnson & Johnson, 2009). Thus, although total time on task was maintained across conditions, we were interested in whether the number of learning cycles completed during the flashcard learning phase would differ between the two conditions. We found moderate evidence for paired flashcard learning being more inefficient (in terms of learning cycles completed) than individual flashcard learning. Individual learners in Experiments 1 and 3 on average completed approximately one more learning cycle than Paired learners, but learners in Experiment 2 completed a similar number of learning cycles regardless of condition (although, numerically, Individual learners completed more learning cycles than Paired learners). Reconciling with prior work, it is possible that the constraints on learners in the Paired condition (i.e., no elaborative explanations) led to a pace of study similar to the pace of study in the Individual condition, but that time spent in brief discussion or switching roles in the flashcard phase led to generally one fewer learning cycle completed by the Paired condition. Likewise, in Experiment 1, Individual learners' tendency to drop cards from study could have allowed them to complete more learning cycles than those in the Paired condition; indeed, Individual learners in Experiment 1 were the only group in the entire set of studies with an average number of cycles completed greater than five. It is also possible that learners struggled to accurately self-report the number of learning cycles. This challenge may have been more prominent for Paired learners who took on multiple roles during the learning phase (i.e., that of tester and testee) and may have given more attention to monitoring their partner than tracking their number of completed learning cycles.

Why is Paired Flashcard Learning Advantageous for Metacognition?

Our findings appear to stem from characteristics of using flashcards with a partner; in particular, its facilitation of overt responses during retrieval practice. Unlike their counterparts in the Individual condition (in the first two experiments), Paired learners had to clearly articulate a response before feedback was provided. In Experiment 3, when Individual learners were instructed to overtly retrieve, these learners were not susceptible to overconfidence and even outperformed Paired learners on the immediate test (possibly due to their ability to engage in retrieval practice for the full 30-min session whereas Paired learners only spent half that time retrieving and the rest being the "tester" for their partner). Broadly, overt retrieval possibly resulted in more effortful retrieval processes (Pyc & Rawson, 2009 offer a discussion about the benefits of effortful retrieval) which were not shortchanged by any peeking at the answers or half-hearted attempts at retrieval.

This suggestion is corroborated by evidence from Tauber and colleagues (2018). In two experiments, participants studied term-definition pairs from Psychology (e.g., confirmation bias). In the first experiment, participants in the retrieval practice conditions were then shown each term and either covertly or overtly (i.e., by typing) retrieved its definition and provided a judgment of knowing (i.e., a judgment of how well they knew the definition to the term). Despite the covert retrieval group scoring lower on the final recall test than the overt retrieval group, they reported significantly higher judgments of knowing during retrieval practice—a similar overconfidence to that observed in the present work using predictions of future test performance. In the second experiment, an additional "enhanced" covert retrieval group was given instructions on how to practice covert retrieval, and learners in the retrieval practice conditions also judged the completeness of their retrieval during retrieval practice. In contrast to evidence from their test performance and judgments of knowing, learners in the enhanced covert retrieval group reported retrieving more of the term definitions during retrieval practice than the overt retrieval condition.

These results suggest that overt retrieval practice may encourage exhaustive retrieval and offer better evidence of one's level of learning.

Although the facilitation of overt retrieval emerged as a key contributor to the benefits of paired flashcard learning, there are other features of paired flashcard learning that may also benefit the accuracy of metacognitive judgments. Paired learners, for example, received feedback only after a complete retrieval attempt and feedback was consistently provided. This consistent feedback from their partner obviated any issues with insufficient checking of answers (Wissman et al., 2016). Inconsistently seeking out feedback may have increased Individual learners' reliance on less diagnostic cues (e.g., ease of retrieved responses; Benjamin et al., 1998), yielding overconfidence. A further consideration involves the increased dropping of flashcards in the Individual condition when dropping was not explicitly prohibited. Such dropping commonly occurred because a given vocabulary-definition pair had been deemed sufficiently learned (which aligns with accounts of study-time allocation such as the Region of Proximal Learning Model; e.g., Metcalfe & Kornell, 2005) and likely deprived learners of robust evidence of their mastery of the vocabulary-definition pairs. Consequently, Individual learners in Experiment 1 based their global judgment of learning on impoverished information relative to Paired learners.

It should be noted that this poor metacognitive calibration in the Individual condition appeared to resolve at a 24-hr delay. In line with other work highlighting that delayed JOLs tend to be more accurate than immediate JOLs (e.g., Nelson & Dunlosky, 1991), it is possible that Individual learners were less susceptible to certain metacognitive illusions (e.g., the stability bias; Kornell & Bjork, 2009) after the passage of time. Additionally, the experience of taking the immediate test in Session 1 of Experiments 2-3 may have offered participants insight into their learning which informed their delayed JOL, and that this information was particularly useful for Individual learners. We investigated this possibility by asking learners in Experiment 3 to report a gJOL for the delayed test in Session 1 (i.e., predict their performance if tested tomorrow) and compared that to their gJOL for the delayed test administered in Session 2 immediately prior to the test. Offering evidence contrary to the possibility that participants had used their immediate test experience to inform their delayed metacognitive judgments, the gJOLs completed right before the delayed test were less well-calibrated (i.e., significantly more underconfident) than the ones completed in Session 1, and this difference was similar between the Individual and Paired conditions, which is overall consistent with the *underconfidence with practice* effect (Koriat et al., 2002).

Potential Effects of Collaborative Learning on Affective and Motivational States

Given classroom evidence that learning with others improves motivation and enjoyment (e.g., McCabe & Lummis, 2018), we were surprised to observe greater negative affect in the Paired condition in Experiment 1. One possible explanation is that being quizzed by a stranger increased anxiety or embarrassment. Although logistical and privacy constraints necessitated random assignment of strangers in the Paired condition, students typically know their study partners in more authentic learning environments (although students sometimes work with strangers in large classes or in assigned groups). This explanation is supported by the lack of evidence for elevated negative affect in Paired learners in Experiments 2 and 3, which incorporated a brief icebreaker activity to facilitate participants getting to know each other (if only superficially) and eased restrictions on verbal communication during the flashcard phase. Although students would likely work with those they know if engaging in paired flashcard learning in everyday life, these findings suggest that implementation of paired flashcard learning in a structured setting (e.g., as a classroom activity) should consider methods to increase students' comfort, particularly if students are asked to work with someone that they do not know.

Limitations and Future Work

The lack of differences in final test performance may stem from several design decisions. Although participants controlled their pace of study and dropping of flashcards, they did not control when to terminate the learning session (as commonly occurs during self-regulated learning). Results may have differed if participants stopped learning once they believed that they had sufficiently mastered the material. The Paired condition may have also been negatively impacted by participants' unfamiliarity with one another and limits on verbal discussion. Learning is supported both by knowledge construction and knowledge consolidation (Roelle et al., 2023). Whereas retrieval practice is particularly beneficial for knowledge consolidation (Roelle et al., 2023), collaboration may support learning by facilitating explanations and elaborations that might be particularly beneficial for knowledge construction (Fiorella & Mayer, 2016). Collaborative learning, for example, is often examined within the context of open-ended tasks which provide ample opportunity for knowledge construction (e.g., Zhu, 2012). It is possible that the carefully controlled procedure and setting of these three experiments impeded exchanges which might spontaneously occur in a real-world collaborative context and support knowledge-building and thus limited the benefits of collaborative flashcard use in the present work. It is further possible that the use of less-complex materials (vocabulary-definition pairs) did not promote the use of these potentially beneficial behaviors to the extent that using more complex materials (e.g., text passages) would have—although using vocabulary as the to-be-learned content aligns with students' self-reported flashcard practices (Authors, 2022a). To address some of these possibilities, future work might employ a "think-aloud" procedure (e.g., Nokes-Malach et al., 2012), or may recruit friends that tend to study together in more naturalistic settings (e.g., study groups).

Practical Implications

Across three experiments, we find that using flashcards in pairs results in more accurate judgments of learning than using flashcards alone, but that this advantage is not present when participants working alone are instructed to retrieve out loud. This result has practical implications for self-regulated learning and effective exam preparation. It suggests that paired flashcard learning can encourage learners to engage in overt retrieval of content during practice testing. Although learners could retrieve out loud by themselves, studies of flashcard learning suggest that the *de facto* procedure when studying with flashcards is to do so with covert retrieval ([Authors], 2022b); in this set of studies, participants did not show benefits of individual flashcard practice in Experiments 1 and 2 when they were not instructed to overtly retrieve during study and monitored to ensure adherence. Further, learners may be more consistent and comfortable retrieving out loud in a social setting than by themselves in, for example, a library or a dorm common area. Thus, when considering the fact that undergraduate students more often use flashcards when studying alone than with a friend (which implies that flashcards are commonly regarded as a solitary tool), it appears that many students are overlooking a potentially more beneficial method of using flashcards—that is, with a partner

References

Barber, S. J., Rajaram, S., & Aron, A. (2010). When two is too many: Collaborative encoding impairs memory. *Memory & Cognition*, *38*, 255-264.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55-68.

Caldwell, A. R. (2022). Exploring equivalence testing with the updated TOSTER R package. *PsyArXiv*.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474-478.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, *1*(1), 3-14.

Efklides, A., Schwartz, B. L., & Brown, V. (2018). Motivation and affect in self-regulated learning. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance (2nd ed.)* (pp. 64-82). Routledge.

Geen, R. G. (1983). Evaluation apprehension and the social facilitation/inhibition of learning. *Motivation and Emotion*, *7*(2), 203-212.

Gilley, B. H., & Clarkston, B. (2014). Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching*, *43*(3), 83-91.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*, 717-741.

Hayat, A. A., Shateri, K., Amini, M., & Shokrpour, N. (2020). Relationships between academic self-efficacy, learning-related emotions, and metacognitive learning strategies with academic performance in medical students: A structural equation model. *BMC Medical Education*, *20*(76).

Holzer, J., Korlat, S., Haider, C., Mayerhofer, M., Pelikan, E., Schober, B., Spiel, C., Toumazi, T., Salmela-Aro, K., Käser, U., Schultze-Krumbholz, A., Wachs, S., Dabas, M., Verma, S., Iliev, D., Andonovska-Trajkovska, D., Plichta, P., Pyżalski, J., Walter, N., Michalek-Kwiecień, J., Lewandowska-Walter, A., Wright, M. F., & Lüftenegger, M. (2021). Adolescent well-being and learning in times of COVID-19—A multi-country study of basic psychological need satisfaction, learning behavior, and the mediating roles of positive emotion and intrinsic motivation. *PLoS One*.

Imundo, M. (2023). Testing together: Collaborative and individual practice testing can yield different patterns of learning following practice testing with varied test formats. Dissertation.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, *38*, 365–379.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). Cooperative learning returns to college what evidence is there that it works? *Change: The Magazine of Higher Learning*, *30*(4), 26-35.

Jönsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research*, *78*, 623-633.

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*(2), 147–162.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297-1317.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*(2), 125-136.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449-468.

Krumboltz, J. D., & Weisman, R. G. (1962). The effect of overt versus covert responding to programed instruction on immediate and delayed retention. *Journal of Educational Psychology*, *53*(2), 89-92.

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

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations , and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269.

Lin, C., McDaniel, M. A., & Miyatsu, T. (2018). Effects of flashcards on learning authentic materials. *Journal of Applied Research in Memory and Cognition*, *7*(4), 529-539.

LoGuidice, A. B., Pachai, A. A., Kim, J. A. (2015). Testing together: When do students learn more through collaborative tests? *Scholarship of Teaching and Learning in Psychology*, *1*(4), 377-389.

McCabe, J. A., & Lummis, S. N. (2018). Why and how do undergraduates study in groups?. *Scholarship of Teaching and Learning in Psychology*, *4*(1), 27-42.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463-477.

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition: An International Journal, 29,* 131-140.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *84*, 93-116.

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.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 26, pp. 125-173). Academic Press.

Nokes-Malach, T. J., Meade, M. L., & Morrow, D. G. (2012). The effect of expertise on collaborative problem solving. *Thinking & Reasoning*, *18*(1), 32-58.

Pan, S. C., Zung, I., Imundo, M., Zhang, X., and Qiu, Y. (2022b). User-generated digital flashcards yield better learning than premade flashcards. *Journal of Applied Research in Memory and Cognition*. Advance online publication.

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied*, *23*(3), 278-292.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710-756.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Positive emotions in education. In E. Frydenberg (Ed.) *Beyond coping: Meeting goals, visions, and challenges* (pp. 149-173). Oxford University Press.

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect?. *Memory and Cognition*, *41*, 36-48.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437-447.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory* (pp. 65-80). Oxford University Press.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roelle, J., Endres, T., Abel, R., Obergassel, N., Nückles, M., & Renkl, A. (2023). Happy together? On the relationship between research on retrieval practice and generative learning using the case of follow-up learning tasks. *Educational Psychology Review*, *35*(4), 102.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432-1463.

Senzaki, S., Hackathorn, J., Appleby, D. C., & Gurung, R. A. (2017). Reinventing flashcards to increase student learning. *Psychology Learning & Teaching*, *16*(3), 353-368.

Siegel, A. L. M., & Castel, A. D. (2019). Age-related differences in metacognition for memory capacity and selectivity. *Memory*, *27*(9), 1236-1249.

Smith, M., & Weinstein, Y. (2016, June). Learn how to study using…retrieval practice. *The Learning Scientists*. https://www.learningscientists.org/blog/2016/6/23-1

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

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science*, *8*(2), 282-293.

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition, 7*(1), 106-115.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*, 429- 442.

Vuorre, M., & Metcalfe, J. (2022). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*, *17*, 269-291.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Wissman, K. T., & Rawson, K. A. (2016). How do students implement collaborative testing in real-world contexts?. *Memory*, *24*(2), 223-239.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568-579.

Zhu, C. (2012). Student satisfaction, performance, and knowledge construction in online collaborative learning. *Journal of Educational Technology & Society*, *15*(1), 127-136.

Zung, I., Imundo, M., and Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, *30*(8), 923-941.