Title

Testing Together: Collaborative and Individual Practice Testing Can Yield Different Patterns of Learning Following Practice Testing with Varied Test Formats

Author

Imundo, Megan Nicole

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Testing Together:

Collaborative and Individual Practice Testing Can Yield Different Patterns of Learning Following Practice Testing with Varied Test Formats

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Psychology

by

Megan Imundo

2023

ABSTRACT OF THE DISSERTATION

Testing Together:

Collaborative and Individual Practice Testing Can Yield Different Patterns of Learning Following Practice Testing with Varied Test Formats

by

Megan Imundo

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2023

Professor Elizabeth Ligon Bjork, Co-Chair

Professor Robert A. Bjork, Co-Chair

Considerable research attests that practice testing (sometimes referred to as *retrieval practice*) is a potent enhancer of memory (Bjork, 1975; Roediger & Karpicke, 2006; Rowland, 2014). Practice testing is thought to enhance learning because the act of retrieving information makes it more recallable in the future, possibly by strengthening or adding retrieval pathways to that information (Bjork, 1975) or by increasing the integration of recalled content with other information stored in long-term memory (Bjork & Bjork, 1992). A particularly useful feature of practice testing is its flexibility: Learners can engage in practice testing using a variety of unstructured (e.g., free-recall) and structured (e.g., multiple-choice) item formats. Much of the research on the efficacy of practice testing for learning, however, has centered on practice testing individually, despite evidence that working with others on other types of tasks (e.g., problem solving) can potentiate learning (Johnson & Johnson, 2009). Collaboration might also facilitate

processes, behaviors, and cognitions which could impact the efficacy of practice testing for learning, such as offering and receiving explanations (Lou et al., 2001), engaging in overt retrieval of information (Tauber et al., 2018), and correcting errors (Barber et al., 2010). I therefore explored if practice testing together (*collaborative practice testing*) would yield different patterns of learning than practice testing alone. I did so across three learning contexts—a large undergraduate STEM course, an online laboratory setting, and an in-person laboratory setting—and across three different structured test formats: multiple-choice, true-false, and flashcards. I elected to use structured practice test formats as they tend to more clearly guide learners' retrieval than unstructured test formats and therefore offer the possibility to more clearly interpret differential patterns in learning as evidence of differential patterns of retrieval during the practice testing event. Overall, I find evidence that collaborative practice testing can result in different patterns of learning than individual practice testing (Chapters 2 and 3) and learning-relevant outcomes (i.e., monitoring of one's learning; Chapter 4). Taken together, these findings suggest that learners may have much to gain by practice testing with others.

The dissertation of Megan Imundo is approved.

Idan Blank

Alan Dan Castel

Melissa Paquette-Smith

Elizabeth Ligon Bjork, Committee Co-Chair

Robert A. Bjork, Committee Co-Chair

University of California, Los Angeles

2023

For Mom, Dad, James, and Grandma and Grandpa Wahl

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 3

Chapter 4

ACKNOWLEDGEMENTS

Chapter 2 is a version of the following article:

Imundo, M. N., Paquette-Smith, M., Clark, C. M., & Bjork, E. L. (under review). The effects of collaborative practice testing on memory for course content in a college classroom.

To my advisors Dr. Elizabeth Ligon Bjork and Dr. Robert Bjork, thank you for trusting me with your ideas, and through all the twists and turns. Thank you also to my committee members Dr. Alan Castel, Dr. Idan Blank, and Dr. Melissa Paquette-Smith for your guidance and thoughtful questions. Melissa, your mentorship has meant the world to me. Thank you for all the writing sessions that may or may not have included actual writing but offered so much more.

I would also like to express endless gratitude to the co-authors of the research presented here: Dr. Courtney Clark, Dr. Melissa Paquette-Smith, Dr. Elizabeth Bjork, Dr. Robert Bjork, Jordan Brabec, Vaishali Denton, Dr. Steven Pan, and Inez Zung.

None of the work presented here could be possible without my undergraduate research assistants. Thank you for the time, effort, thoughtfulness, and dedication you put into these and so many other research projects. Everything I have done at UCLA has truly been a team effort and it has been a joy to discover knowledge together.

Teaching and research have been deeply intertwined throughout my career and so it is impossible for me to talk about one without the other. I want to especially thank Katie Dixie, K. Supriya, Chris Mott, Beth Goodhue, Peggy Davis, Laurel Westrup, and Nedda Mehdizadeh for challenging me to be the educator *I* want to be.

I also want to thank my friends and colleagues in the department, especially Mary, Emily, Victoria, Katie, Alex, Maggie, Danny, and, of course, my lab mates Saskia and Jordan.

My journey to this moment began long ago, far from UCLA, and there are so many people to honor for how they have shaped my life and my career.

First, I want to thank my mom and dad, Kimberly and Bob Imundo, for sacrificing so much for my education. Since I was a kid, you have both told me that I can be anything I want to be, and then you worked so hard every day to make sure that was true. Mom, it was inspiring to see you not just go back to school but go all the way to your master's degree. Whenever I doubted that I could do it, I thought of you and knew that I could. To my brother James, thank you for always picking up the phone. I also want to thank my grandfather, Carroll Wahl, who was one of the first to believe in me and my voice as a writer.

Second, I want to thank my Northwestern friends and colleagues who showed me how exciting and fulfilling conducting psychological science can be. Thank you to Dr. Renee Engeln for giving me my first chance at research and for having great taste in coffee shops. Thank you also to Dr. David Rapp for being an outstanding mentor, colleague, co-author, confidant, and supporter—I'm excited for all the research questions we have left to explore.

Speaking of David, I give all the love and gratitude to Rebecca Adler. In 2016, I asked you if it would be weird if I joined David's research lab—which you were already a part of—since we were already (a) sharing a room, (b) in the same sorority, and (c) taking a couple classes together. Luckily you gave the "all-clear" and we've been a dynamic duo ever since.

Finally, thank you to Josh for putting up with my grad student-isms, only somewhat judging my coffee preferences, and for encouraging me to set healthy boundaries with my laptop.

And, all my love to Sesame, the best writing buddy a girl could have.

VITA

2016-2018	Lab Manager Body and Media Lab Northwestern University Evanston, Illinois
2017-2018	Honors Thesis Student Reading Comprehension Lab Northwestern University Evanston, Illinois
2018	B.A. Psychology, Cognitive Science Northwestern University Evanston, Illinois
2019	M.A. Psychology University of California, Los Angeles
2019-2020 2020-2022 2022-2023	Teaching Assistant Teaching Associate Teaching Fellow University of California, Los Angeles
2021-2023	Instructional Research Consultant Community Instructional Transformation Initiative Center for the Advancement of Teaching University of California, Los Angeles
2022	C. Phil., Psychology University of California, Los Angeles
2023	Graduate Certificate in Writing Pedagogy University of California, Los Angeles
2023	Research Scientist Learning Engineering Institute Arizona State University Tempe, Arizona

SELECTED PUBLICATIONS

Clark, C. M., **Imundo, M. N.**, & Paquette-Smith, M. (2023). Identifying limitations in research activity. Encouraging retrieval practice. Society for the Teaching of Psychology (STP) Psychology Tools eBook. https://teachpsych.org/ebooks/psytoolbox

Engeln, R., & **Imundo, M. N.** (2020). I (don't) love my body: Counter-intuitive effects of body-affirming statements on college women. *Journal of Social & Clinical Psychology*, *39*(7), 617-639. https://doi.org/10.1521/jscp.2020.39.7.617

Engeln, R., Loach, R., **Imundo, M. N.**, & Zola, A. (2020). Compared to Facebook, Instagram use causes more appearance comparison and greater body dissatisfaction in college women. *Body Image*, *34*, 38-45. https://doi.org/j.bodyim.2020.04.007

Imundo, M. N., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2020). Where and how to learn: The distinct benefits of contextual variation, restudying, and retrieval practice for memory. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820968483

Imundo, M. N., & Rapp, D. N. (2021). Weight-of-evidence reporting can ameliorate the negative effects of falsely balanced texts. *Journal of Applied Research in Memory and Cognition*, *11*(2), 258-271. https://doi.org/10.1016/j.jarmac.2021.10.002

Pan, S. C., Zung, I., **Imundo, M. N.,** Zhang, X., & Qiu, Y. (2022). User-generated digital flashcards yield better learning than pre-made flashcards. *Journal of Applied Research in Memory and Cognition.* https://doi.org/10.1037/mac0000083

Paquette-Smith, M., **Imundo, M. N.**, & Clark, C. M. (2023). Encouraging retrieval practice. Society for the Teaching of Psychology (STP) Psychology Tools eBook. https://teachpsych.org/ebooks/psytoolbox

Rapp, D. N., **Imundo, M. N.**, & Adler, R. M. (2019). Do individual differences in conspiratorial and political leanings influence the use of inaccurate information? In P. Kendeou, D. H. Robinson, & M.T. McCrudden (Eds.), *Misinformation and Fake News in Education* (pp.103-122). Charlotte, NC: Information Age Publishing.

Salovich, N. A., **Imundo, M. N.**, & Rapp, D. N. (2023). Story stimuli for instantiating true and false beliefs about the world. *Behavior Research Instruments, 55,* 1907-1923. https://doi.org/10.3758/s13428-022-01904-6

Tien, I., **Imundo, M. N.**, & Bjork, E. L. (2023). Viewing oneself during synchronous online learning increases appearance anxiety and decreases memory for lecture content. *Applied Cognitive Psychology*, *37*(2), 443-451. https://doi.org/10.1002/acp.4048

Zung, I, **Imundo, M. N.**, & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, *30*(8), 923-941. https://doi.org/10.1080/09658211.2022.2058553

CHAPTER 1

General Introduction and Overview

A somewhat astounding body of work attests that *practice testing* (sometimes referred to as *retrieval practice*) is a powerful enhancer of memory (the testing effect; Bjork, 1975; Butler & Roediger, 2007; Imundo et al., 2021; Pan & Rickard, 2018; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014; Yue et al., 2015). Of currently known learning activities readily available to students, it is considered the most potent (Dunlosky et al., 2013). Therefore, facilitating effective practice testing is of considerable educational and practical interest.

Learners can engage in practice testing in a variety of ways. Some of these formats are fairly unstructured, such as free-recall (e.g., Roediger & Karpicke, 2006) or essay-based (e.g., Cocks, 1929) practice tests. Other formats contain far greater intrinsic structure, more clearly guiding the thought processes of learners. These types of practice test items can include multiple-choice questions (e.g., Little et al., 2012; Sparck et al., 2016), true-false propositions (e.g., Brabec et al., 2021; Uner et al., 2021), and even flashcards (Pan et al., 2022). Although engaging with each type of test format has the promise of promoting desirable student outcomes, each may also come with its own pitfalls. Multiple-choice questions, for example, may not facilitate learning of directly tested content to the same extent as other types of test formats (e.g., cued-recall; Foos & Fisher, 1988; McDaniel, Anderson et al., 2007; McDaniel, Roediger, & McDermott, 2007 offers a detailed review). Multiple-choice tests, however, can also facilitate learning of content conceptually related to target information (Little et al., 2012). True-false practice tests may enhance learning of directly tested and conceptually related content as well, but whether they do so is highly dependent on the validity of the true-false proposition (i.e.,

whether the correct response is true or false; Brabec et al., 2021). Finally, flashcard decks are a portable method of practice testing that is commonly used by students during self-regulated study sessions (Wissman et al., 2012; Zung et al., 2022). Students, however, may not always self-evaluate their learning (or even the correctness or completeness of their retrieval attempt) accurately and drop cards from study too quickly, leading to lower levels of learning than if students had not dropped cards from study (Kornell & Bjork, 2008).

How might learners use these forms of practice testing to benefit from the testing effect, and yet avoid some of the above potential pitfalls of these practice test formats? I offer practice testing with others–often referred to as *collaborative practice testing*–as an approach that may maximize the benefits (and potentially minimize the drawbacks) of some structured test formats. Below, I describe the potential limitations of (1) multiple-choice, (2) true-false, and (3) flashcard-based practice testing and why practice testing with others might facilitate beneficial processes which could overcome the potential drawbacks of these practice test formats.

Multiple-Choice Practice Testing Does Not Always Promote Effortful Retrieval

Despite being a commonly-used test format in educational contexts, multiple-choice tests have been criticized for facilitating less-beneficial recognition processes (Glover, 1989; Rowland, 2014 presents a meta-analytic review) over retrieval because they present learners with the correct answer. To answer the question correctly, therefore, it has been argued that the learner must only recognize the correct answer rather than recall it, thereby depriving learners of the opportunity to engage in beneficial generation (*the generation effect*; Slamecka & Graf, 1978) or retrieval processes (Little et al., 2012 provides a detailed discussion).

Based on the retrieval effort hypothesis (Pyc & Rawson, 2009), more effortful retrieval is thought to potentiate learning to a greater extent than less difficult retrieval. In Experiment 1,

Pyc and Rawson (2009) found that longer delays between practice test trials resulted in better final test performance than shorter delays between trials. In their second experiment, they measured retrieval latency as a proxy for retrieval difficulty, finding that longer delays between practice test trials resulted in increased time it took to retrieve a response, and that this increased retrieval latency was associated with improved final test performance. Other work has also found that increasing the delay between initial study and initial testing (i.e., increasing the difficulty of retrieval practice) also results in increased memory for practice tested content (Karpicke & Roediger, 2007; Whitten & Bjork, 1977). Together, these studies provide evidence that making practice testing more challenging enhances its benefits for learning.

Based on the retrieval effort hypothesis, it was thought that multiple-choice practice tests were less beneficial for learning than practice test formats that facilitated more "pure" retrieval processes, such as cued-recall tests. Indeed, many studies suggested that multiple-choice practice testing enhanced learning of directly tested content to a lesser degree than cued-recall practice testing (e.g., Foos & Fisher, 1988; McDaniel, Anderson, et al., 2007). A comprehensive meta-analysis of test effects for recognition practice tests–for which multiple-choice tests were coded as a type of recognition test–found that taking recognition tests yielded a smaller test effect ($g = 0.29$, a small-to-medium effect) than taking cued-recall practice tests ($g = 0.61$, a medium-to-large effect) which required generation of the correct answers (Rowland, 2014).

Recently, however, work by Little and colleagues has debunked multiple-choice practice tests as merely recognition tests. Their work suggests that multiple-choice practice tests can facilitate retrieval–and therefore support learning–particularly if they are constructed to include competitive (i.e., plausible) incorrect alternatives (Little & Bjork, 2010; Little et al., 2012; Little et al., 2019).

Although this work suggests that well-constructed multiple-choice tests can indeed facilitate retrieval, and that this retrieval is good for learning, multiple-choice tests still often underperform (in terms of the size of the effect), even when well-constructed, compared to cued-recall test formats (McDaniel & Little, 2019 offers a review). It is possible that this difference is due to multiple-choice tests facilitating retrieval, but that this retrieval is less effortful (and thus, less beneficial) than the retrieval facilitated by cued-recall tests.

Collaboration Can Promote Effortful Retrieval

How might learners be motivated to more effortfully retrieve during multiple-choice practice testing? One possibility is for learners to complete multiple-choice practice tests collaboratively rather than individually. A key feature of collaborative testing is the sharing of retrieved information with fellow group members. It is therefore crucial that group members engage in overt (i.e., out loud) retrieval rather than covert (i.e., internal) retrieval. Although there is still some controversy in the literature, several studies have suggested that overt retrieval may facilitate learning better than covert retrieval (Kubik et al., 2020; Tauber et al., 2018; *cf.* Smith et al., 2013).

Additionally, often group members go above and beyond simply recalling content. Rather, learners elaborate on or contextualize recalled information for their fellow group members. Research on collaborative learning more broadly suggests that socially justifying beliefs (Bruffee, 1984), offering explanations (Lou et al., 2001), or resolving controversy (Johnson et al., 1998) can encourage learners to recall information, apply knowledge in new ways, or even reorganize existing schemas. These elaborations and explanations during group discussion may foster deeper activation of target content during practice testing, facilitating durable learning.

True-False Practice Testing Does Not Always Promote Retrieval of Directly Tested and Related Content

Although it is well-established that practice testing can enhance learning of directly tested content, recent research has explored whether practice testing can enhance learning of conceptually related (but not directly tested) information. The *elaborative retrieval hypothesis* (Carpenter, 2009; Carpenter & DeLosh, 2006) suggests that retrieval practice can encourage the activation of targeted and related elaborative information. The *constructive retrieval hypothesis* (Hinze et al., 2013) additionally suggests that engaging in constructive mental processes during practice testing bolsters memory for directly tested and related content to a greater extent than engaging in processing oriented around rote retrieval. Based on these theories, it is plausible that practice testing may enhance memory for content beyond what was directly tested.

Much of the work on the benefits of practice testing for related content has examined whether certain test formats (in particular multiple-choice) may foster different retrieval processes than other formats (e.g., cued-recall). For example, Little et al. (2012) demonstrated that multiple-choice items constructed to include competitive alternatives can facilitate learning of both directly tested and related information. In a series of studies, participants read two expository text passages (e.g., on Yellowstone National Park). These passages contained content (e.g., facts about Steamboat Geyser and Castle Geyser) that were conceptually related to one another (e.g., the content related to geysers located within Yellowstone National Park).

After reading the passages, participants either took a multiple-choice practice test or a cued-recall test on content from one of the passages (the other passage served as the nontested control passage). After a 5-min distractor task, all participants then took a final cued-recall test. Some of the final cued-recall questions queried information that was directly targeted by a

practice test question (e.g., Steamboat Geyser was a correct response for a practice test item *and* a final cued-recall test item). Some of these questions queried information that was related to directly tested content on the practice test (e.g., Steamboat Geyser was a correct response for a practice test item, with Castle Geyser listed as an incorrect alternative, and then Castle Geyser was a correct response for a final cued-recall test item). The remaining questions were on content from the nontested control passage.

Although taking a cued-recall practice test and taking a well-constructed multiple-choice practice test both enhanced learning of directly tested content as compared to no practice testing, taking a multiple-choice practice test further enhanced learning of related content. Little et al. (2012) suggested that well-constructed multiple-choice practice tests may facilitate learning of related content because learners are motivated to recall information about each incorrect alternative in order to rule them out as the correct answer, whereas cued-recall tests lack the cues to prompt this additional retrieval. To test this hypothesis, Little et al. (2019) asked participants to report the information that they recalled during multiple-choice practice testing. When participants recalled information about the key incorrect alternative while practice testing, they went on to answer the related question on the final cued-recall test correctly 75% of the time. In contrast, participants did so only 35% of the time when they did not recall information about the key competitive alternative while practice testing.

Research on true-false practice testing has similarly found that this test format can facilitate learning of both directly tested and related content (Brabec et al., 2021). In Experiments 1 and 2, participants read the same text passages that were used in Little et al. (2012). They then took a true-false practice test on one of the passages, with the other passage serving as the nontested control passage. Across both experiments, a highly specific pattern of

results emerged: Evaluating true practice statements enhanced later memory for directly tested content (but not related content) whereas evaluating false practice statements enhanced later memory for related (but not directly tested) content. Brabec et al. (2021) suggested that this pattern of results demonstrated the "one-and-done" effect. The one and done effect suggests that learners retrieve only as much information as is necessary to determine whether a proposition is true or false ("one") and then stop attempting to retrieve additional information ("and done"). Consequently, when the practice test statement was true, learners generally only retrieved information about the directly tested concept, and when the statement was false, learners generally only retrieved information about the related concept.

In Experiment 3, Brabec et al. (2021) adjusted the format of the true-false practice test items by adding in competitive clauses. These competitive clauses contrasted two related pieces of information in a "this-not-that" format such that each practice test item presented the target term and the related term; e.g., Steamboat Geyser (not Castle Geyser) is the oldest geyser. In contrast to Experiments 1 and 2, true-false practice testing enhanced performance on both previously tested and previously related final cued-recall items regardless of the validity of the true-false proposition. This minor, but ultimately powerful, change suggests that the presence of both target and related information in a true-false item can encourage learners to retrieve information about both.

Taken together, the results of these studies offer three key conclusions. First, practice testing can, at times, facilitate learning of both directly tested and related content. This effect is key for the pedagogical utility of practice testing, as learners are rarely assessed on a quiz or an exam with the exact same items which appeared on a practice test. Second, different practice test formats can result in markedly different patterns of learning. Test formats that offer a variety of

cues to learners (e.g., multiple-choice and competitive-clause true-false) appear to facilitate broader patterns of learning than test formats that offer fewer cues to learners (e.g., cued-recall and traditional true-false). And third, these different patterns of learning likely stem from the patterns of retrieval that occur during practice testing. Thus, it follows that changing patterns of retrieval during practice testing may lead to different outcomes from the practice testing event– whether or not those changes stem from the format of the practice test or another element of the practice testing context.

Collaboration May Promote Retrieval of Directly Tested and Related Content During True-False Practice Testing

As Brabec et al. (2021) identified, it is possible to avoid the "one-and-done" effect by including competitive clauses into true-false items. An alternative, however, to rewriting practice test items might be to change the practice testing context such that learners work together when practice testing.

Within a group, individual members each bring their own body of knowledge and unique understandings of the material. When working together, students therefore have the opportunity to share knowledge with one another (i.e., re-exposure) or support others in retrieving their own knowledge (i.e., cross-cuing) (Blumen & Rajaram, 2008; Blumen & Stern, 2011; Nokes-Malach et al., 2019). During re-exposure, one group member retrieves a piece of information that no other group member would have otherwise retrieved. This retrieval thus offers a restudy opportunity of that content to the rest of the group. During cross-cuing, one group member says something that then prompts another group member to retrieve a piece of information that they would not have otherwise retrieved. This group member benefits from the retrieval opportunity, while other group members benefit from exposure to that retrieved content. Together, these

processes may encourage comprehensive engagement with content as learners recall target, and potentially related, information during group discussion. Additionally, in line with the elaborative and constructive retrieval hypotheses, elaborations, explanations, and other efforts to come to a consensus (i.e., sharing of evidence; Clark et al., 2000) during group discussion may encourage retrieval of and exposure to both target and related content even when taking true-false practice tests that lack competitive clauses.

Learners Do Not Always Monitor Their Learning Accurately During Flashcard-Based Practice Testing

Using flashcards is a commonly recommended way to implement practice testing (e.g., Smith & Weinstein, 2016) and students report using flashcards to study (Wissman et al., 2012). Physical flashcard decks are highly portable and are able to contain a lot of information in a small amount of space, and digital flashcard decks can be used on a wide variety of devices wherever students have an internet connection (Zung et al., 2022). Learners can include many types of information on flashcards, but most often use them to study key concepts and vocabulary (Wissman et al., 2012; Zung et al., 2022). When practice testing with flashcards, learners typically write a key piece of content on the front of the card (e.g., a vocabulary word) and then some related content (e.g., a definition) on the back of the card. Learners then use the front of the card as a cue to retrieve content from the back of the card, and may then look at the back of the card after the retrieval attempt to evaluate whether they did so successfully.

Although practice testing with flashcards offers the possibility for learners to reap the benefits of the testing effect (Pan et al., 2022), students often use flashcards suboptimally. For example, students commonly report dropping cards from study during a learning session (Zung et al., 2022). In fact, one study of undergraduate students who were instructed to practice test

with flashcards found that 63% of cards were dropped after only one successful retrieval (Kornell & Bjork, 2008). Doing so may reduce the effectiveness of retrieval practice because the decision to drop the item is premature and the information is not well-learned (Kornell & Bjork, 2008), or because dropping flashcards reduces beneficial spacing between flashcards (Kornell, 2009).

Another critical decision is whether or not to seek out feedback. Feedback has been shown to enhance the testing effect (Carpenter et al., 2022; Pan & Rickard, 2018). Overall, practice testing is more effective when feedback is provided ($g = 0.73$, considered a large effect) than when it is not, even when initial retrieval success is high (i.e., > 75%; $g = 0.56$, considered a medium effect; Rowland, 2014). When feedback is provided, even unsuccessful retrieval attempts can boost learning relative to simply reading the correct answers (Kornell et al., 2009). Feedback can assist learners in correcting errors or misconceptions, maintaining correct knowledge, and, particularly in the case of free-recall or cued-recall practice testing, feedback re-exposes learners to unrecalled content (Roediger & Butler, 2011).

Learners practice testing with flashcards may decide not to seek out feedback (i.e., view the back of the flashcard) after each retrieval attempt (Wissman et al., 2012). Instead, learners may rely on certain cues, such as ease-of-retrieval or confidence to decide whether to look at external information (*the cue-utilization view of metacognition*, Koriat, 1997; *the monitoring-affects-control hypothesis*, Nelson & Leonesio, 1988; Nelson & Narens, 1994). Reliance on fickle cues like retrieval fluency (Benjamin et al., 1998; but Bjork et al., 2013 provides a detailed discussion) can lead students to drop items from study even after *zero* successful retrieval attempts (Kornell & Bjork, 2008). In these cases, not only do learners fail to learn correct

pairings of content but may actually learn incorrect pairings that could persist to a later assessment.

Collaborative Practice Testing Offers Increased Opportunities for Feedback

During collaborative practice testing, group members have easily accessible and timely sources of feedback in the form of their fellow group members, even if feedback is not in and of itself built into the test (Molin et al., 2020; Rempel et al., 2021). Having access to immediate feedback may be especially beneficial when practice testing on more complex content (e.g., definitions of vocabulary terms), as learners may not always correctly judge the accuracy of their responses (Dunlosky & Rawson, 2012). The act of recalling aloud during collaborative practice testing offers the opportunity for group members to provide input on the completeness of the retrieval; if a learner thinks that their response was accurate or complete, a group member may at times offer a valuable alternative perspective.

Practice testing with others can also foster error correction when learners exchange information and respond to one another's responses during discussion (Barber et al., 2010). Crucially, a correct minority within a group can successfully correct errors, even if that misconception is held by the majority of the group (Smith & Tindale, 2010). Research on collaborative problem solving suggests that a correct minority of the group can often sway the incorrect majority (often by demonstrating evidence of the correct answer or by explaining their reasoning) such that the group consensus ultimately arrives at the correct conclusion (Clark et al., 2000). Remarkably, during collaborative practice testing, groups can even come to the correct answer when all group members originally begin the discussion endorsing an incorrect answer. As groups collectively discuss each group member's (incorrect) suggested response, they can

identify the faulty reasoning underlying each one and ultimately come to a consensus on the correct response (Smith et al., 2009; Vázquez-García, 2018).

In sum, a major benefit of collaborative practice testing may be the increased ability to correct errors–particularly in cases when other sources of feedback may not be readily available–which could foster both learning and the accurate assessment of one's learning.

Overview of the Current Dissertation

This dissertation explored whether individual and collaborative practice testing resulted in different learner outcomes in laboratory and classroom contexts across three types of practice testing formats. The first set of studies centered on multiple-choice practice testing and was conducted within a large undergraduate STEM course (Chapter 2). Across two experiments, I and my co-authors investigated if collaborative practice testing would result in greater long-term retention (1-week, 2-week, or 6-week delay) than individual practice testing. The second set of studies centered on practice testing with two variations of the true-false format: traditional and competitive-clause (Chapter 3). Across four experiments, the effects of individual versus collaborative practice testing for learning of previously tested and previously related content was explored in a laboratory setting and a large undergraduate STEM course. The final set of studies centered on flashcard-based practice testing and was conducted in a laboratory setting (Chapter 4). Across two experiments, I and my co-authors investigated if paired flashcard-based practice testing would lead to increased immediate and delayed test performance, and more accurate assessments of one's learning, compared to individual flashcard-based practice testing.

CHAPTER 2

The Effects of Collaborative Practice Testing on Memory for Course Content in a College Classroom

Abstract

The benefit of collaborative testing to learning has been examined via two-stage exams for high-stakes tests. The present research extends inquiry into this topic by examining whether learning benefits might arise from collaborative testing during formative stages of learning. In a large Introductory Psychology course, we investigated whether low-stakes collaborative practice testing enhanced learning compared to individual practice testing. Our data demonstrate that collaborative practice testing led to better performance on surprise individual retention tests at 1-week and 2-week delays (Exp. 1) but not after a 6-week delay on the course final exam (Exp. 2). Students' attitudes towards group work also improved from pre-course to post-course. The addition of group-building exercises prior to collaborative practice testing did not impact its efficacy. The present research suggests that collaborative practice testing can enhance long-term retention of course material and provides a potential model for implementing collaborative practice testing in large STEM classes.

Retrieving information from memory has been shown to enhance learning relative to more passive study strategies like highlighting or rereading (i.e., the testing effect; Bjork, 1975; Carpenter et al., 2008; Dunlosky et al., 2013; Glover, 1989; Roediger & Karpicke, 2006; Carpenter et al., 2022 and Pan & Rickard, 2018 offer detailed reviews). Unsurprisingly, students tend to score higher on tests that they take in groups (Cortright et al., 2003; Garaschuk, 2022; Gilley & Clarkston, 2014; Lusk & Conklin, 2003; Woody et al., 2008). More noteworthy, however, is that collaborative testing can bolster individual group members' ability to recall the tested information on a future assessment (Cortright et al., 2003; Gilley & Clarkston, 2014; Vázquez-García, 2018).

Much of the research on the benefits of collaborative testing in the classroom has focused on collaboration with respect to high-stakes exams or "exam wrappers" in which students first take an exam individually and then retake all or part of the exam in small groups. Indeed, some instructors encourage collaboration during or immediately after an exam, but additionally, students often report that they spontaneously engage in collaborative testing before exams as part of informal study sessions with their peers (Wissman & Rawson, 2016). It is not clear, however, whether this type of low-stakes collaborative practice testing is more beneficial than individual practice testing when it is used in preparation for a later criterion exam. The present studies attempt to answer this question in the context of a large Introductory Psychology course. Our aims were twofold: (1) to assess whether collaborative practice testing in preparation for an exam could lead to better learning and retention of course content compared to individual testing and (2) to examine whether facilitating constructive interactions among group members would enhance these potential benefits.

Evidence from Two-Stage Exams

As previously mentioned, one of the most common ways that collaborative testing has been implemented in classrooms is in the form of two-stage exams. For example, in the first stage, students might take a midterm exam alone, and then, in the second stage, retake all or part of that same exam in a small group (e.g., Gilley & Clarkston, 2014). Often, two-stage exams are administered as part of a high-stakes assessment for which students have prepared extensively. A number of studies have demonstrated that students tend to perform better on the group stage of the exam compared to the individual stage (Cortright et al., 2003; Gilley & Clarkston, 2014; Lusk & Conklin, 2003). Some of these studies have further assessed students' individual understanding of the tested materials after a delay and found that two-stage collaborative exams can enhance individual learning when compared to a no-test control or an individual re-test (Cortright et al., 2003; Cooke et al., 2019; Gilley & Clarkston, 2014). Other studies, however, do not show any long-term benefits of collaborative testing (Cooke et al., 2019; Woody et al., 2008). The inconsistency in these results may be due to differences in the duration of the delay before the final test (i.e., 48 hours vs. 6 weeks) or differences in students' level of preparation prior to the group stage of the exam. Even if two-stage collaborative exams do not always lead to greater learning gains compared to individual exams, there is no evidence from the two-stage literature that having students in a classroom test in groups is detrimental to their learning (LoGuidice et al., 2015 offers a review). However, as we discuss later, there could be drawbacks to inefficient group work.

Potential Reasons for Collaboration Benefits

Testing collaboratively, whether on a two-stage exam or on a practice test, could lead to enhanced learning for a variety of reasons. For instance, working with a group of students with a

nonoverlapping body of knowledge could support re-exposure and cross-cuing of information (Blumen & Rajaram, 2008; Blumen & Stern, 2011; Nokes-Malach et al., 2019). Re-exposure can occur when one group member retrieves a piece of information that no other group member would have retrieved on their own. Cross-cuing occurs when a piece of information retrieved by one group member prompts another group member to retrieve something they otherwise would not have retrieved. To illustrate, if groups are given a question about color vision, one member might retrieve information regarding one theory of color vision (e.g., the trichromatic theory), and that information then prompts another member to retrieve information regarding a different theory (e.g., the opponent-process theory). Under ideal circumstances, these additional encoding opportunities can act as a form of restudy or scaffolded retrieval practice for the group, and thereby foster better long-term learning. If, however, the test is too easy, such that every member performs so highly that each group member cannot offer any unique information to the group, then these processes are unlikely to occur and collaborative learning has the potential to be no more, or even less, effective than testing alone, a phenomenon which has been observed in other types of collaborative learning activities (e.g., problem-solving; Nokes-Malach et al., 2012; Nokes-Malach et al., 2019).

Not only do groups enhance the possibility of making several retrieval cues available, but they can also provide an opportunity for error correction. When, for example, one group member produces an incorrect answer, others in the group—if they hold accurate prior knowledge—can correct that group member's mistake (Barber et al., 2010). Through this process, individuals can benefit from the identification and refutation of their errors, and their fellow group members can retrieve and reaffirm their own understanding of the material. Students note that the opportunity for immediate peer-based feedback is a prominent benefit of

taking collaborative tests (Rempel et al., 2021). Remarkably, groups can even come to the correct answer when all group members originally begin the discussion endorsing an *in*correct answer. Through discussion of the various (incorrect) responses, group members may recognize the faulty reasoning underlying each response choice, ultimately leading them to realize the correct response (Smith et al., 2009). Even when students answer incorrectly on both the individual and the group portions of a two-stage collaborative exam, the experience of taking collaborative exams seems to facilitate an understanding that allows for students to improve their performance when given a delayed opportunity to revisit the questions that they previously answered incorrectly (Vázquez-García, 2018).

Working with others might foster additional unique processes that enrich the learning experience as compared to working alone, such as the opportunity to justify or explain their answer to the group. While working in a group, students must navigate different perspectives, which can spark constructive disagreement. Such discussions can encourage students to reevaluate their own knowledge, seek out new information to reach a consensus view, and/or refine their conclusions about a topic or task (Johnson & Johnson, 2009; Johnson et al., 1998).

Potential Barriers to Effective Collaboration

If, however, students are not motivated to work together or to rely on one another, then they may not optimally engage in collaborative learning. Students, for example, often remark that group work can be plagued by dysfunctional group dynamics, such as ineffective communication, "free-riding," or dominating group members (Gillespie et al., 2006; Hillyard et al., 2010; Woody et al., 2008). In these cases, communication between group members is stilted, one-sided, or even nonexistent. These kinds of experiences may lead students to develop negative views towards collaborative activities. As one student stated, "one bad experience with

group work can ruin it for you forever" (Gillespie et al., 2006). Such negative views may hamper students' willingness to engage in future collaborative learning activities such as collaborative testing. Promisingly, however, positive group work experiences can improve attitudes towards group work among students who previously held negative attitudes towards in-class collaboration (Wosnitza & Volet, 2014) and increase excitement for future group work (Linnenbrink-Garcia et al., 2011; Reinig et al., 2011).

Facilitating Effective Collaboration

Well-constructed group work can encourage group members to work together cohesively and constructively. Johnson and colleagues (1998) propose five key features of effective collaborative learning: positive interdependence, individual accountability, promotive interaction, social skills, and group processing. Effective collaboration occurs when group members rely on each other to produce the desired outcome (*positive interdependence*), they each feel uniquely responsible for the success of the group (*individual accountability*), they work together in trusting and mutually beneficial ways (*promotive interaction*), and they employ conflict resolution and other interpersonal strategies to support their collaboration with other group members (*social skills*). These groups also regularly monitor and adjust their group behaviors and procedures as they work to achieve their goals (*group processing*). A number of studies have examined the effectiveness of one or two of these characteristics on group success, often in the context of unstructured group problem-solving or group projects. In general, they find that heightening one or more of these five aforementioned features can make groups more successful at achieving their goals (Aramovich & Larson, 2013; Janssen et al., 2011; Johnson & Johnson, 2009; Lou et al., 1996; Perkins & Saris, 2009). To our knowledge, however, the

effectiveness of these strategies has not been systematically examined in the context of a structured collaborative testing activity.

In sum, previous work has demonstrated that collaborative testing—mainly the use of collaborative two-stage exams—can facilitate better retrieval of information in the short term (e.g., Gilley & Clarkston, 2014). Evidence for how such benefits extend to longer-term learning is mixed, and could depend (at least in part) on how well groups work together. The present research examined the benefits of collaborative practice testing for the learning and retention of material presented in an Introductory Psychology course. Unlike in prior classroom work, which has primarily focused on two-stage collaborative exams as part of formal, high-stakes tests, students engaged in both collaborative and individual practice testing as part of a low-stakes, in-class activity. The efficacy of collaborative versus individual practice testing on learning was assessed via performance on the initial practice tests, as well as performance on individual retention tests that occurred one week (Test 1) and two weeks (Test 2) after the initial activity (Experiment 1) and on a final exam that occurred 6 weeks after the activity (Experiment 2). We hypothesized that students would (1) score higher on the practice tests when working in groups and (2) information practiced collaboratively would be better retained on delayed tests compared to information practiced individually.

We also manipulated whether groups were instructed to engage in group-building activities or not before beginning the practice testing activity. We predicted that students who completed group-building activities would show greater benefits of collaborative practice testing both on the initial activity as well as on the delayed tests. Findings of the present work contribute to our understanding of the benefits of classroom-based collaborative practice testing

and inform implementation recommendations for instructors who might wish to incorporate such learning activities into their own teaching.

Experiment 1

Experiment 1 tested the hypothesis that low-stakes collaborative practice testing would enhance learning to a greater extent than individual practice testing. We implemented a 1-hour practice testing activity in two parallel sections of a large Introductory Psychology course. Whether practice tests were conducted in groups or alone was manipulated within-subjects such that all students completed two practice tests collaboratively and then two practice tests individually. Across the two sections of the course, we also manipulated whether students completed group-building activities or neutral activities prior to engaging in the practice testing activity. We hypothesized that students who engaged in group-building activities would benefit more from collaborative testing. We also incorporated measures of pre-course and post-course attitudes towards group work as a way to assess whether participating in the course improved positive attitudes towards group work and diminished discomfort with group work, and whether any such improvements were larger for those students who engaged in group-building activities.

Methods

Participants

The participants were students enrolled in two large sections of an online Introductory Psychology course in Fall 2020. The final sample consisted of 569 students who completed the testing activity on Zoom (324 from the section that completed the neutral activities and 245 from the section that completed the group-building activities). Data from 62 additional students were removed from the final sample because they did not attend all or part of the testing activity (n =

50), or they participated in a make-up activity at a different time[1] (n =12). Approval for all experiments in this manuscript was obtained from the UCLA Institutional Review Board (IRB).

Design

Both sections of the course were comprised of asynchronous instructional modules and synchronous online laboratory sessions. During the lab sessions students typically worked in small groups to complete an activity together on Zoom. One section was assigned to complete the group-building activities and the other completed the neutral activities. We randomly selected which of the two sections would receive the group-building activities and which section would not. However, students selected which section time to enroll in so assignment of students to section is not random. Given that instructions for the practice testing activity were given verbally, random assignment to group-building or neutral activities within course section was not feasible. Each small group of students completed one of six counterbalanced versions of the activity. Across these versions we counterbalanced which topics were practiced individually and which topics were practiced collaboratively.[2] Later, to assess each student's learning, they were asked to complete two surprise individual retention tests administered one and two weeks after the in-lab practice testing activity.

Materials

Samples of study materials can be found in Appendix A.

---

[1] Forty students (7%) completed the collaborative portion of the testing activity with their group on Zoom, but the individual portion of the testing activity was recorded as being completed on a different date than the testing activity. This discrepancy may have occurred either because (a) they did indeed complete the individual portion of the testing activity outside of class time or (b) they did not hit "submit" on part of the activity and the survey platform used to run the activity automatically submitted their response after a delay. As their attendance for the collaborative portion was recorded by the instructional team, data from these students were included in the final sample.

[2] The number of students who completed the Biological Psychology (n = 293), Learning (n = 269), Research Methods (n = 294), and Sensation and Perception (n = 282) practice tests collaboratively was roughly equal, as was the number of students who completed Biological Psychology (n = 276), Learning (n = 300), Research Methods (n = 275), and Sensation and Perception (n = 287) individually.

Practice tests. Four 10-item multiple choice practice tests on Research Methods, Biological Psychology, Sensation and Perception, and Learning were used as materials for the study. In line with the literature on effective multiple-choice questions (e.g., Little & Bjork, 2015), the multiple-choice questions each contained four competitive lures in addition to the correct answer.

Retention Tests. The first retention test (Test 1) contained eight questions, two from each topic of the practice testing activity. The questions were identical to items that participants had seen during the practice testing activity. The second retention test (Test 2) contained a different set of eight questions also taken directly from the practice testing activity. Thus, no question that appeared on one retention test appeared on the other. Given that only four topics were practiced tested on, questions from all practiced topics were included on each retention test. Questions were selected to be nonoverlapping in content, even within topics (i.e., a competitive alternative in one question was never the correct response for another question).

Pre-Course and Post-Course Survey Items. The pre- and post-course surveys contained a number of items that assessed students' preparation and experiences in the course. Of particular interest in the present study were measures of group work attitudes. The 17-item group work attitudes scale (Forrester & Tashchian, 2010) was administered twice: once on the pre-course survey and again on the post-course survey. This scale consisted of three factors: positive attitudes towards academic group work (7 items; e.g., "I enjoy participating in group work"); discomfort with group work (4 items; e.g., "I am often afraid to ask for help within my group"); and preference in group work (6 items; e.g., "I understand information better after explaining it to others in a group"). Students rated their agreement with the items on a scale from 1 (not at all true of me) to 5 (very true of me). The subscales have demonstrated acceptable

internal reliability in prior work (Forrester & Tashchian, 2010). The positive attitudes towards group work (Cronbach's α's = .91, .91), discomfort with group work (α's = .79, .76), and preference in group work (α's = .68, .67) subscales demonstrated acceptable internal reliability on the pre-course and post-course surveys, respectively, in Experiment 1.

Procedure

At the beginning of the course, students completed the pre-course survey, which included the group work attitudes scale (Forrester & Tashchian, 2010).

Across the two different sections of the course, we manipulated whether students engaged in group-building activities, designed to foster good group work during the practice testing activity and during the labs leading up to that activity, or not (Appendix A includes the activity materials). For example, one group-building activity asked students to begin their lab session by reading through a list of group work strengths (e.g., cooperating, clarifying) as a lab group and share out loud how each group member intended to use one of these techniques to be more productive during the lab. Students in the other section engaged in neutral activities designed to foster good learning habits more broadly but that did not focus on topics thought to be important to effective group work. For example, one neutral activity asked students to read through a list of techniques to generally improve productivity (e.g., reduce external distractions, set a time limit) and then share out loud how each group member would use one of these techniques during the lab.

During the first lab of the course, students in each section were placed randomly into small groups of 3-5 students using Zoom's breakout rooms feature. Given the size of the course and the use of online instruction due to the COVID-19 pandemic, it is unlikely that group members knew each other prior to being placed in a lab group together. The practice testing

activity occurred during the fifth lab of the course online via Zoom (https://zoom.us/). During two of the previous labs, students had completed either the group-building or neutral activities with the same small lab group that they would work with on the practice testing activity. For the other two previous labs, which had a less extensive group component, they worked with a different group of students. Thus, students had worked with their small lab group on two occasions prior to the practice testing activity.

To further foster group cohesion, students in the course section that completed group-building activities were also instructed to use a driver-navigator procedure during the collaborative rounds of the practice testing activity. That is, they were instructed to have one member of the lab group share their screen and type for the group while the other group members provided input. All group members—including the group member typing for the group—were instructed to contribute equally to the group's deliberations. In contrast, students in the section that completed neutral activities were instructed to have each member in their small group open the practice testing activity on their own screen. While students in this section were encouraged to talk to one another to figure out the answers, they were also told that each group member was supposed to submit their own practice test (which decreases positive interdependence relative to submitting as a group).

The first two practice tests were taken in their group and the second two practice tests were taken alone. Students completed the group task before the individual task to ensure that they were able to get through all the group questions during class time (and to reduce the likelihood that students would leave before they completed the group part). All students, whether taking a practice test collaboratively or individually, were asked to take these tests from memory and not use their notes.

For each topic covered in the practice testing activity, students made a global prediction of future performance (e.g., during this test you will answer 10 questions about [topic], how many questions do you think (you/your group) will answer correctly?) and a global postdiction of performance immediately after taking the test (e.g., you just answered 10 questions about [topic], how many questions do you/your group think you answered correctly?). They also "bet" between 0 and 3 points for each question, as gamifying learning tasks can foster student engagement (Faiella & Ricciardi, 2015). Students were told that betting points was just for fun and did not impact their grade for the lab. These measures were not central to our hypotheses and will not be discussed further. At the end of each practice test, students were able to review their responses and were given feedback as to which answer was the correct answer.

About one week after the practice testing activity occurred (M = 10.09 days, SD = 4.04), students received a surprise retention test (Test 1) that was placed within one of their asynchronous, self-paced weekly modules. These asynchronous modules were comprised of lecture videos separated by pages containing practice questions or reflection prompts; the link to the surprise retention test was placed on one of these pages. Students were told that completion of the retention test was for participation credit and that their test performance would not impact their course grade. The retention test contained eight questions, two per topic, taken from the prior practice testing activity. A second surprise retention test (Test 2), which included a new subset of eight questions also taken from the prior practice testing activity, was administered in the same manner about a week later (*Mean amount of time between the first and second tests* = 6.21 days, SD = 2.89)[3] (i.e., about two weeks after the practice testing activity). Following the

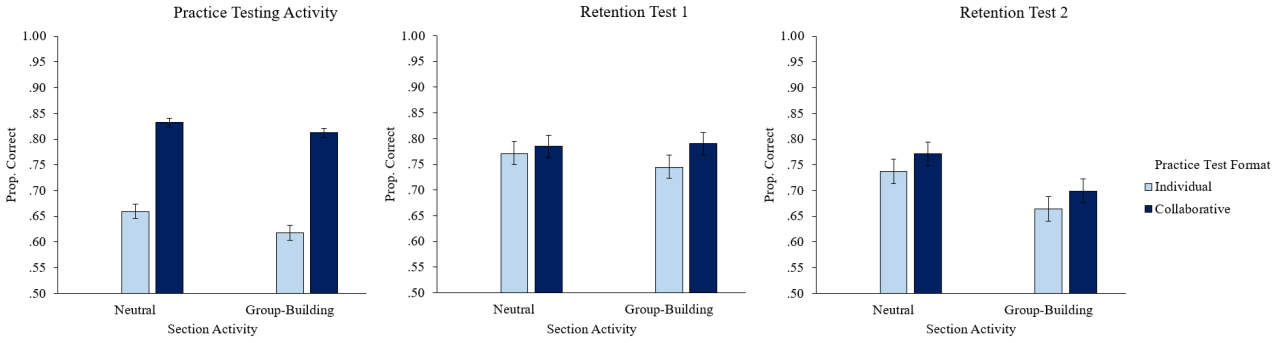
[3] Of the students that completed both retention tests, 4% (n = 18) completed the retention Test 2 before or on the same day as retention Test 1. Given that the items on retention Tests 1 and 2 were unique, we kept these students in the final sample for the retention test analyses.

end of the course, students completed the post-course survey, in which they completed the group work attitudes scale a second time.

Results

Figure 1 illustrates correct performance on practice tests that were taken collaboratively versus individually (Left Panel) as well as performance on Retention Test 1 (Middle Panel) and Retention Test 2 (Right Panel) for items that were previously practiced collaboratively versus individually.

Figure 1



Correct Performance on Practice tests (Left Panel), and on Retention Test 1(Middle Panel) and Retention Test 2 (Right Panel) as a function of section activity (Neutral vs. Group-Building) and whether the information had appeared on a previous practice test that was taken collaboratively versus individually.

Practice Test Performance

To assess whether performance on the practice tests was impacted by the practice test format (individual or collaborative) or the section activities (neutral or group-building) we conducted a 2 x 2 mixed ANOVA with practice test format as the within-subjects variable, group activity as the between-subjects variable, and practice test score as the dependent variable. For the collaborative tests, not all students in a given group submitted the same responses; therefore,

in cases where there was a discrepancy between one individual's responses and the other members' response, we averaged across the group and assigned each group member that average score for the collaborative practice test.[4]

As expected and indicated in the Left Panel of Figure 1, a main effect of practice test format was observed, such that scores were significantly higher on practice tests completed in groups (M = .82, SD = .10) as compared to practice tests completed individually (M = .64, SD = .17), F (1, 567) = 587.67, $p < .001$, $\eta_p^2 = .51$. Though we had no reason to predict systematic differences between students in one course section versus the other, the assignment of students to course section was not random as students could choose which section to enroll in before the course began. Thus, only the section activity x practice test format interaction is of interest as it tests whether group-building activities can enhance the effects of collaborative practice testing. This interaction was nonsignificant, F (1, 567) = 1.99, p = .16, η_p^2 = .003, suggesting that students who completed group-building activities and students who completed neutral activities benefited similarly from collaboration.

Retention Test Performance

Out of the 569 students who participated in the practice testing activity, 498 completed the first retention test that was integrated into an asynchronous module which opened approximately one week after the activity. Of those students who completed Test 1, 462 also completed Test 2 which opened approximately two weeks after the practice testing activity.

A 2 x 2 mixed ANOVA with practice testing activity format (individual or collaborative) as the within-subjects variable and section activity (neutral or group-building) as the between-

[4] Across both sections, 36.7% of students had a discrepancy between their individual response and fellow group members' responses on at least one of the collaborative tests. The average difference between an individual's responses and the group's responses across all practiced items was 0.003 points (SD = 0.28).

subjects variable revealed that students performed better on the one-week retention test (shown in the Middle Panel of Figure 1) for topics they previously practice tested collaboratively (M = .79, SD = .25) than for topics they previously practice tested individually (M = .76, SD = .27), F (1, 496) = 6.50, p = .01, η_p^2 = .01. No interaction between section and practice test format was observed, suggesting that the performance of students who completed either neutral or group-building activities was similarly enhanced by engaging in collaborative practice testing, F (1, 496) = 1.89, p = .17, η_p^2 = .004.

Performance on the two-week retention test (Test 2, shown in the Right Panel of Figure 1) replicated the pattern seen for the one-week retention test but with a different set of items. Again, students scored better on items that had previously been on a practice test taken collaboratively (M = .74, SD = .25) compared to items that had previously been on a practice test taken individually (M = .70, SD = .26), F (1, 460) = 5.88, p = .02, η_p^2 = .01. As was the case with Test 1 performance, no significant interaction between practice testing activity format and section activity was observed, F (1, 460) = 0.001, p = .97, η_p^2 < .001.

Group Work Attitudes

Given that positive group experiences can shift students' attitudes about group work (Wosnitza & Volet, 2014), we hypothesized that group-building activities might also improve students' overall attitudes towards group work. We examined this possibility by assessing the change in students' group work attitudes from pre-course to post-course. We measured change in group work attitudes by subtracting each student's pre-course subscale score from their post-course subscale score and then analyzed these scores using a multivariate analysis of variance (MANOVA) with section activity (group-building vs. neutral) as the between-subjects variable. The dependent variables were the change scores for the positive attitudes towards group work,

discomfort with group work, and preference in group work subscales of the group work attitudes scale.

As expected, students' (n = 534) positive attitudes towards group work increased from pre-course to post-course and discomfort with group work decreased from pre-course to post-course; positive attitudes: M_{change} = 0.20, SD = 0.84, F (1, 532) = 31.99, $p < .001$, $\eta_p^2 = .06$; discomfort: M_{change} = -0.22, SD = 0.80, F (1, 532) = 37.01, $p < .001$, $\eta_p^2 = .07$. Preference for group work stayed about the same (M_{change} = -0.03, SD = 0.55), F (1, 532) = 1.45, $p = .23$, $\eta_p^2 = .003$. There was no overall main effect of section activity, Wilks' $\lambda = 1.00$, F (3, 530) = 0.70, $p = .55$, $\eta_p^2 = .004$, and the effect of section was also nonsignificant for all subscales (all p's > .19), indicating that the magnitude of the change in group work attitudes from pre-course to post-course was similar for both sections, regardless of whether they had done group-building activities. These findings suggest that students may view group work more favorably and feel less discomfort working in groups after participating in a course where there is structured group work, which was present in both sections. It is possible that our brief group-building manipulation (with a total time of 25 minutes across three prior laboratory sessions plus the practice testing activity) was not substantial enough to impact long-term attitudes towards group work, and therefore students in the section that completed the group-building activities did not show larger improvements in group work attitudes on the post-course survey than students in the neutral group activities section.

To summarize the results of Experiment 1, we found evidence that students performed better on practice tests when they took them collaboratively, and even though corrective feedback was provided for all practice tests, the benefits of collaboration extended to retention tests administered one and two weeks later. In line with typical forgetting, students

demonstrated lower test performance on retention test 1 than during the practice testing activity for items previously practiced collaboratively. In contrast, they demonstrated higher performance on retention test 1 than during the practice testing activity for items previously practiced individually. We suggest that the increase in performance for these items is likely due to learning from corrective feedback during the practice testing activity.

Though the advantage of collaborative practice testing over individual practice testing for long-term retention would be considered modest for a laboratory-based study (corresponding to Cohen's d = ~0.2 for the first and second retention tests), the size of this effect aligns with benchmarks for practically meaningful educational interventions in authentic learning environments (Kraft, 2020). It is further noteworthy that this benefit occurred from implementing a single ~45-min in-class activity in an online course.

We did not see an effect of completing group-building activities on group test performance, but it may be the case that having these brief activities spaced across multiple weeks diminished their effects on practice testing performance. Their effects may have been more potent if all of the activities had occurred on the day of the practice testing activity, a possibility we examined in Experiment 2.

Experiment 2

The results of Experiment 1 supported our hypothesis that students' practice test performance and long-term learning would benefit from taking practice tests collaboratively as compared to individually. Further, students' positive attitudes towards group work increased while their discomfort with group work lessened from the beginning to the end of the course. We did not, however, find support for our hypothesis that group-building activities would enhance the benefits of collaboration. One possibility is that spreading out the brief group-

building activities across the quarter decreased their impact. Another possibility is that the group-building activities were not substantial enough to enhance the quality of group performance. Accordingly, we made two changes in Experiment 2 that we thought might amplify the effects of the group-building activities: 1) we offered all the group-building or neutral group activities on the day of the practice testing activity and 2) we adjusted the group activities to facilitate each of Johnson et al.'s (1998) five principles of good collaboration—positive interdependence, individual accountability, promotive interaction, social skills, and group processing—more clearly.

Due to changes in the course structure from when Experiment 1 to Experiment 2 was conducted, long-term learning was measured by assessing students' performance on final exam questions that were previously tested in the practice testing activity as opposed to surprise retention tests. Examining final exam performance gave us the opportunity to investigate the efficacy of prior collaborative practice testing on a high-stakes summative assessment that was more consequential for students. It also allowed us to assess the efficacy of this activity at a longer (6-week) delay. As practice testing is generally intended to help students prepare for such high-stakes examinations, we expected Experiment 2 to provide additional clarity as to the potential educational benefit of collaborative practice testing.

Method

Participants

In Winter 2021, 452 students in two large sections of Introductory Psychology participated in the present study online via Zoom as part of an in-class lab activity. Ninety-nine students were removed from the study because they did not complete all or part of the practice testing activity, leaving 353 students in the final sample (159 who completed neutral group

activities and 194 who completed group-building activities)[5]. Of those students, 339 ultimately completed the final exam.

Design

The design of Experiment 2 was identical to that of Experiment 1 with the exception of the following changes. First, the instructors switched conditions such that the instructor who administered group-building activities in Experiment 1 offered the neutral activities in Experiment 2, and vice versa. Again, given that the group-building and neutral activities required different verbal instructions, section activity could only be manipulated across course sections. Second, due to the timing of the midterm exam, the practice testing activity occurred in Week 5 (instead of Week 7) of the quarter and all group-building activities occurred during the practice testing activity to maximize their impact on group cohesion. Third, learning was assessed via an open-book final exam administered six weeks after the practice testing activity.

Given that the practice tests were well-matched in difficulty, we opted for a simpler version of the counterbalancing, that switched which two topics appeared on practice tests taken collaboratively and which two topics appeared on practice tests taken individually. Students in even numbered groups answered practice questions on Research Methods and Sensation and Perception in a collaborative manner, and answered practice questions on Biological Psychology and Learning in an individual manner. Odd numbered groups did the reverse.[6]

[5] Unlike in Experiment 1, students were explicitly told that they were allowed to leave the Zoom call after the collaborative practice testing session (though they were encouraged to stay and complete the individual practice tests if they could). As a result, 34% of students completed all practice tests on the same day, and 66% completed at least one of the individual practice tests on a different day.

[6] This counterbalancing resulted in roughly equal numbers of students answering questions about Biological Psychology (n = 178), Learning (n = 178), Research Methods (n = 175), and Sensation and Perception (n = 175) appearing on practice tests taken in a collaborative manner as those answering questions about Biological Psychology (n = 175), Learning (n = 175), Research Methods (n = 178), and Sensation and Perception (n = 178) appearing on practice tests taken in an individual manner.

Materials

The practice tests used in Experiment 2 were identical to those used in Experiment 1. Four questions from these practice tests (two that were previously tested individually and two that were previously tested collaboratively) were included on an open-book final exam administered six weeks after the practice testing activity; these questions served as the retention test. The same measures to assess attitudes towards group work were used as in Experiment 1. All three subscales demonstrated acceptable internal reliability on the pre-course and post-course surveys, respectively, in Experiment 2 (positive attitudes towards group work: α's = .90, .89; discomfort with group work: α's = .82, .78; preference in group work: α's = .72, .63).

Procedure

The section activity materials were similar to Experiment 1's but in the section that was assigned the group-building activities they were adjusted to more clearly facilitate Johnson et al.'s (1998) criteria for effective collaborative learning (Appendix A includes the full activity materials). We again implemented *positive interdependence* through the use of a driver-navigator procedure by requiring each group to submit only one copy of the answers, thereby indicating to group members that they were expected to come to a consensus prior to submitting their response. We facilitated *promotive interaction* by giving groups a description of productive group behaviors (e.g., clarifying, harmonizing) and asking students to write down one such behavior that they planned to use during the collaborative portion of the practice testing activity before they began the activity. We encouraged appropriate use of *social skills* by asking group members to find one thing they all had in common and write it down. We promoted *group processing* by including a prompt between the first and second collaborative practice test that asked groups to consider whether each member was participating equally, and encouraged them

to talk amongst themselves about changes to the group's procedures that they would like to implement when taking the second practice test. Finally, we heightened *positive interdependence* by informing students before they actually began the activity that (a) during the last five minutes of the lab session the entire class was going to answer a few questions about topics that had been covered on the collaborative practice tests and that (b) if 90% or more of the class answered these questions correctly, the entire class would earn one of the points for this lab (in actuality, however, students earned the point regardless of class performance). It was therefore important that students work together during the collaborative rounds to make sure that *all* group members understood the material. The neutral section activities were similar to those used in Experiment 1 (i.e., focusing on general study skills), except in Experiment 2 they were administered during the same lab session in which the practice testing activity occurred.

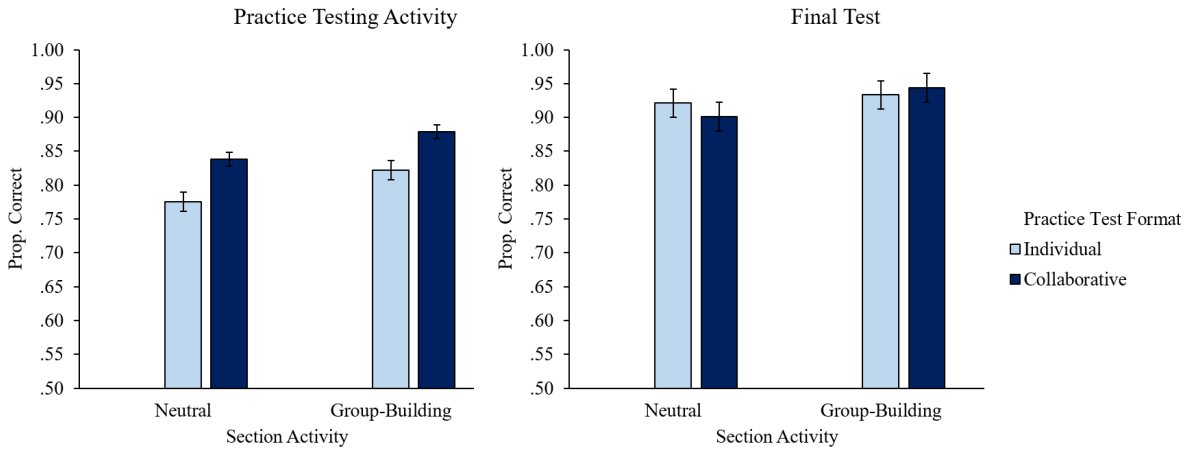
Additionally in Experiment 2, in order to ensure that students agreed upon a single answer, all students used a driver-navigator procedure when taking a practice test in a collaborative manner. Only those students engaging in group-building activities, however, were prompted at the end of their first collaborative practice test to discuss whether every group member participated equally during the practice test and, in an effort to encourage them to think carefully about their group process, they were also asked to think about 1 or 2 specific things their group could do to encourage group members to participate equally when taking the second collaborative practice test.

Finally, long-term learning was assessed via performance on four questions on the final exam that were previously answered on practice tests taken collaboratively versus individually during the practice test activity.

Results

Figure 2 includes an illustration of students' practice test and final exam performance.

Figure 2



Effect of practice test format and section activity on students' ability to correctly answer questions during the practice testing activity (Left panel) and on the final exam administered six weeks after the practice testing activity (Right panel).

Practice Test Performance

As in our analyses of the data collected in Experiment 1, we assessed the impact of collaboration on practice testing activity performance. As illustrated in the Left panel of Figure 2, we found that students performed better on practice tests that were taken collaboratively versus practice tests taken individually (M = .86, SD = .09 versus M = .80, SD = .13), F (1, 351) = 69.85, $p < .001$, η_p^2 = .17, replicating the benefit of group practice testing that was observed in Experiment 1. Again, given that students chose which section time to enroll in (and thus assignment of students to course section was not random), we focus on the section activity x practice test format interaction. As in Experiment 1, no significant interaction between section group activity and practice test format was observed, suggesting that the benefit of taking

practice tests in a collaborative manner was similar for students regardless of section group activity, F (1, 351) = 0.22, p = .64, η_p^2 = .001.

Final Exam Performance

To assess the long-term impact of collaborative versus individual practice testing, we examined students' scores on four questions (one per topic) which appeared on the final exam administered six weeks later. As illustrated in the right panel of Figure 2, students scored extremely well on the repeat practice test questions that occurred on the final exam, and their ability to answer these questions did not differ in relation to whether the questions had appeared on practice tests taken individually (M = 0.93, SD = 0.19) versus collaboratively (M = 0.92, SD = 0.20), F (1, 337) = 0.09, p = .76, η_p^2 < .001. Additionally, we did not observe an interaction between section and practice test format, F (1, 337) = 1.03, p = .31, η_p^2 = .003.

There are several plausible reasons why we did not replicate the long-term benefits of collaborative practice testing seen in Experiment 1. First, it is possible that the delay between the practice testing activity and the final exam was simply too long (six weeks) to observe sustained benefits of collaborative versus individual practice testing. Second, whereas students were unprepared for the surprise Experiment 1 retention tests, they had likely studied extensively for the high-stakes final exam, contributing to high overall performance. Third, while both the final exam and the ungraded retention tests from Experiment 1 were open-book tests, students were discouraged from using their notes in the instructions for the ungraded retention test (i.e., they were asked to try to answer from memory) whereas they were encouraged to do so in the instructions for the final exam, and were likely more motivated to use available course resources to look up information or check their answers when taking the final exam.

Group Work Attitudes

To assess whether group-building activities improved students' attitudes towards group work, we again conducted a MANOVA with section activity (group-building or neutral) as the between-subjects variable and change in score on the positive attitudes towards group work, discomfort with group work, and preference in group work subscales of the group work attitudes scale as dependent variables.

As expected, students' (n = 311) positive attitudes towards group work increased, M_{change} = 0.15, SD = 0.76, F (1, 309) = 10.68, p = .001, η_p^2 = .033, and discomfort with group work decreased, M_{change} = -0.35, SD = 0.81), F (1, 309) = 57.66, p < .001, η_p^2 = .16, from pre-course to post-course. Again, preference in group work stayed about the same (M_{change} = -0.05, SD = 0.55), F (1, 309) = 1.51, p = .22, η_p^2 = .005.

Unlike in Experiment 1, however, we did observe an overall significant effect of section activity, Wilks' λ = 0.97, F (3, 307) = 3.36, p = .019, η_p^2 = .03. The effect of section was significant for the subscales of positive attitudes towards group work, F (1, 309) = 8.12, p = .005, η_p^2 = .03 and preference in group work, F (1, 309) = 5.47, p = .02, η_p^2 = .02, but not discomfort with group work, F (1, 309) = 0.11, p = .74, η_p^2 < .001.

Students who completed group-building activities showed a slightly larger increase in positive attitudes towards group work (M_{change} = 0.26, SD = 0.77) than students who did not (M_{change} = 0.02, SD = 0.72). Similarly, students who completed group-building activities reported a slight increase (M_{change} = 0.11, SD = .53) in preference in group work from pre-course to post-course, whereas students who completed the neutral activities demonstrated a slight decrease (M_{change} = -0.03, SD = 0.55). Together, these findings align with our expectation that students would view group work more favorably after participating in a course that emphasizes

group work. In the section that completed group-building activities the magnitude of this change was larger than in the section that completed neutral activities, which highlights the potential benefits of providing greater support and structure to group work activities.

General Discussion

In two classroom-based studies, collaborative practice testing during a synchronous online learning session yielded better performance than individual practice testing on the practice testing activity. Additionally, in Experiment 1, students demonstrated benefits of collaboration on surprise individual retention tests that were administered 1-week and 2-weeks after the practice testing activity. Overall, students in both experiments reported increased positive attitudes towards group work and decreased discomfort in group work from the beginning of the course to the end of the course. Taken together, these findings suggest that a well-structured collaborative testing activity may enhance long-term learning and positively impact students' opinions about group work.

Implications of Assessing Retention via Low-Stakes versus High-Stakes Tests

The patterns of performance on delayed retention tests differed between Experiments 1 and 2. In Experiment 1, students' long-term retention of previously practiced content was enhanced by collaborative practice testing when assessed via a surprise retention test given 1-week and 2-weeks after the testing activity. In contrast, in Experiment 2, we found no difference in long-term retention of information previously tested individually versus collaboratively on a high-stakes final exam given six weeks later. We propose several possibilities for the difference in this pattern of results. First, the retention intervals in Experiment 1—though long relative to many other studies—were only 1-week and 2-weeks as compared to six weeks in Experiment 2. Thus, it is possible that the benefits of collaborative testing for retention of information

decreased over time and did not last through a six-week retention interval, as has been observed in other studies of collaborative testing (e.g., Ives, 2014).

Second, the method of assessing information retention varied considerably between the two experiments. In Experiment 1, the retention tests were administered within self-paced, asynchronous online modules that students completed each week. Students did not expect to take retention tests and therefore had no reason to prepare for them. Thus, the surprise retention tests given in Experiment 1 can be thought to represent a relatively "pure" assessment of differences in students' retention of information. In contrast, in Experiment 2, long-term learning was assessed via the inclusion of questions previously presented on the practice tests as part of an open-book final exam. This difference poses a number of potential issues. Students in Experiment 2 may have prepared extensively for the open-book final exam which may have "washed out" any benefits of the collaborative practice testing activity. Additionally, the open-book format of the exam in Experiment 2 offered students the option to look up answers. Although students could have also looked up answers while taking the retention tests given in Experiment 1 (despite being asked to rely only on their memories), the fact that the retention tests were a surprise and were not going to contribute to the students' course grade probably decreased the likelihood of their looking up answers compared to the high-stakes final exam given in Experiment 2. Differences in overall performance observed on the retention tests of Experiment 1 and the corresponding final exam questions in Experiment 2 further support the possibly that students prepared and potentially looked up answers more for the high-stakes final exam. Students on average earned 78% and 72% correct on retention Test 1 and retention Test 2, respectively, whereas performance on the previously practiced questions that appeared on the final exam was much higher (i.e., above 90%). That looking up answers on the final exam

potentially engages information search processes rather than retrieval processes presents an additional challenge in interpreting the results of Experiment 2.

Reconciling with Prior Work

In Experiment 1 of the present research, we found that collaborative testing enhanced recall on delayed retention tests, whereas prior studies on collaborative testing have produced mixed results, especially when learning was assessed at longer delays (Cooke et al., 2019; Cortright et al., 2003; Woody et al., 2008). There are multiple possible reasons why the practice testing activity used in Experiment 1 allowed us to observe long-term effects of collaborative testing. In the present research, students completed the practice testing activity with their lab groups; thus, a level of familiarity and comfort with each other may have generally enhanced the quality of information exchange and collaborative processing that occurred during their collaborative activity.

Our collaborative testing activity also occurred relatively early in the learning process as part of a low-stakes formative assessment. In contrast, much of the classroom-based literature on collaborative testing involves two-stage exams, that often occur after students have taken a high-stakes midterm or exam (LoGuidice et al., 2015 offers additional discussion). The experience of participating in collaborative testing long before a required formal assessment is administered may have created a low-stakes environment in which students felt more comfortable exchanging ideas. Furthermore, students in Experiment 1 were not maximally prepared for the practice testing activity, and therefore may have been more likely to have had gaps in their knowledge (or even misconceptions) that could have been filled in and/or corrected in the act of exchanging their ideas with other students during collaborative practice testing.

An additional possibility for why students benefitted more from the collaborative practice test than the individual practice test is that, due to the logistical constraints of a large course, the collaborative test always preceded the individual test. As a result, students may have been more fatigued for the individual practice test or rushed in completing that portion of the activity. Gilley and Clarkston (2014), however, found collaborative testing to be more effective than individual testing even though all students completed the individual testing phase prior to the collaborative testing phase. Another potential consequence of taking the collaborative practice test first is that students may have found the subsequent individual practice test less interesting or engaging, potentially reducing their learning from that testing opportunity. Yet, Bloom (2009), using a between-subjects design, found that collaborative retesting following a midterm exam yielded superior retention of information over individual retesting. There, the benefits of collaborative testing for learning could not be attributed to students' comparisons of collaborative and individual testing.

Lastly, there is some variability in the degree to which previous studies have used assessments that ask students to transfer their knowledge to a new problem/situation. In the present research, we used questions on the retention tests that were identical to the questions presented during the practice activity. It is possible that the impact of collaborative testing may be less pronounced when subsequent questions require greater extrapolation or generalization of knowledge. Or perhaps, given that collaboration might create greater discussion of alternative responses, group testing might actually enhance students' ability to answer questions about related topics on a future test. Although the current study was not designed to assess how well this learning transfers to related questions, or to extricate the varied potential mechanisms which

may contribute to the benefits of collaborative practice testing, this is an important avenue for future research.

The Role of Group-Building Activities

Across both of the present experiments, no consistent effects of the group-building exercises on the effectiveness of the collaborative learning activities were observed, whether these exercises were spread out across the quarter or implemented all on the same day. One possibility as to why the group-building exercises did not improve scores on the collaborative practice test was that the collaborative testing activity was so structured that groups worked well together regardless of whether they engaged in the group-building vs. neutral activities prior to competing the practice tests. Students, for example, had a clear time limit to complete the practice testing activity, and they were motivated to finish the activity on time so that they could earn their full attendance grade (earned by completing all the practice tests). This feature may have reduced some of the off-topic discussion and coordination issues that often plague group work. The instructor also went through the procedure of the activity and answered student questions before the activity began so that students were well-prepared for each step of the activity. Additionally, the general structure of multiple-choice questions (i.e., five answer options with the mandate to select just one) may have clearly guided discussion and encouraged group members to stay on-task. Finally, as discussed above, students had previously worked twice with the other members of their groups and thus may have already had reasonable opportunity to form effective group dynamics, regardless of the section activities that they completed.

The lack of an effect for these group-building activities may indicate that they may not be necessary when promoting quality collaboration within highly-structured activities, but they still

might be useful in other collaborative contexts. Prior research has largely examined group-building activities within the setting of more complex or open-ended tasks, such as completing a course-long group project or solving challenging scenario-based problems (Aramovich & Larson, 2013; Janssen et al., 2011; Kim & Ryu, 2013). When completing open-ended tasks, students may be more susceptible to less efficient and efficacious group work, and thus the additional support of prior group exercises may genuinely help those students. In the present research, however—although they certainly did not harm the quality of group work—these prior group exercises seem to have been less necessary.

Concluding Comments

Taken together, the present pattern of results suggests that collaborative practice testing can be an effective strategy for enhancing students' learning and long-term retention of course content. In our first experiment, the implementation of a relatively short (1-hour) activity produced memory gains that were sustained one and two weeks later. Furthermore, the enhanced long-term retention gained from collaborative practice testing was comparable whether or not students had engaged in extra group-building exercises prior to the testing activity, suggesting that students can benefit from the experience of collaborative practice testing even when they have not been explicitly prepared for cohesive group work. Thus, these types of brief low-stakes testing activities may be a way, in general, to foster better attitudes in learners for engaging in group work. Future work could explore which features of the collaborative testing activity are most facilitative of long-term retention of content. Overall, however, the demonstrated effectiveness of collaborative practice testing for enhancing learning in the present classroom setting of a large Introductory Psychology course would seem easily adaptable for the teaching of a wide range of subject matters and thus make it possible for instructors of a large variety of

courses to incorporate such activities into their individual instructional plans. The next chapter explores whether collaboration might enhance learning of not only directly tested content, but also conceptually related content, both in the laboratory and in the same large undergraduate STEM course.

CHAPTER 3

Patterns of Learning Following Collaborative and Individual True-False Practice Testing

Abstract

In four experiments, we investigated the impact on learning of individual versus collaborative practice testing across two variations of the true-false test format in an online laboratory and an undergraduate STEM course. In Experiments 1 and 2, participants read two passages, took a true-false practice test alone or in small groups, and then completed a final individual cued-recall test which queried knowledge of previously practiced and conceptually related content. Experiment 1 employed traditional true-false items whereas Experiment 2 used competitive-clause true-false items which contrasted tested and related content in a "this-not-that" format. Experiment 3 (traditional true-false) and Experiment 4 (competitive-clause true-false) were conducted in a large undergraduate STEM course. To-be-learned content was presented in previous course meetings, so learners first completed a collaborative practice test, then an individual practice test, and finally an individual cued-recall test. Overall, our results suggest that collaborative practice testing with traditional true-false items can elicit broader learning benefits than individual practice testing (as evidenced by performance on previously tested and previously related final test items). Additionally, when learners have robust prior knowledge (Experiment 3), those who practice test collaboratively may be less susceptible to learning incorrect pairings of content from false practice test items than those who practice test alone. Competitive-clause true-false practice testing, on the other hand, broadly facilitated learning and led to low rates of learning incorrect associations from false practice items regardless of whether the practice test was completed individually or collaboratively. Together, these results offer

insights into how various formats and implementations of true-false practice testing may impact learning.

Collaborative and Individual Practice Testing Lead to Different Patterns of Learning Across Two Variations of The True-False Test Format

Practice testing, sometimes referred to as retrieval practice, can be a potent enhancer of memory (Bjork, 1975; Roediger & Karpicke, 2006; but Pan & Rickard, 2018 and Rowland, 2014 offer meta-analytic reviews). Research conducted in classroom settings suggests that engaging in practice testing can substantially enhance later memory for tested content (McDaniel, Roediger, & McDermott, 2007; Schwieren et al., 2017 offer detailed reviews). Practice testing is one of the most efficacious study strategies currently known (Dunlosky et al., 2013), while also being a learning technique that may be particularly effective when done in groups (e.g., Imundo et al., under review).

With the benefit of practice testing for directly tested information well-established in a variety of learning contexts, recent research has examined whether practice testing with multiple-choice questions can enhance learning of not only the correct answer, but also information pertaining to the incorrect alternatives offered by the question (i.e., related content; e.g., Little & Bjork, 2010; Little et al., 2019). Only recently, however, has there been consideration of whether *true-false* practice testing can also enhance learning of related content. One recent study suggests that the construction of the true-false item matters, with true-false practice testing having differential effects on memory for previously related content depending on whether the initial statement was true or false and whether the practice item is modified to include both the directly tested and related content in a "this-not-that" format (i.e., competitive-clause true-false items; Brabec et al., 2021). Overall, this body of work indicates that whether practice testing enhances learning for related content depends on the practice item retrieval and how it may guide learners' retrieval.

In the current line of research, we examine the effect of practice testing on memory for previously tested and previously related content across traditional and competitive-clause true-false test items, first in a laboratory setting and then within the context of a large STEM course. As previous work suggests that collaboration during practice testing can encourage beneficial retrieval processes (Blumen & Rajaram, 2008; Blumen & Stern, 2011; Barber et al., 2010; but LoGuidice et al., 2015 offers a detailed discussion), we additionally examined the impact of taking these true-false practice tests individually versus in small groups on learning.

Different Test Formats May Elicit Retrieval with Varied Characteristics

A variety of hypotheses offer explanations as to why retrieval can be beneficial for learning. Bjork (1975), for example, suggests that retrieval can be a memory modifier, increasing the number of retrieval pathways or strengthening currently existing routes to stored information (also Carrier & Pashler, 1992). In line with the desirable difficulties framework (Bjork, 1994), the *retrieval effort hypothesis* (Pyc & Rawson, 2009) suggests that more difficult retrieval potentiates learning to a greater extent than less difficult retrieval when that retrieval is either successful (Carpenter & DeLosh, 2006; Imundo et al., 2021; Kornell et al., 2011) or accompanied by feedback (Kornell et al., 2009; Roediger & Butler, 2011; Smith & Handy, 2016). Other explanations as to why retrieval can be beneficial for learning highlight the potential for practice testing to enhance learning of both directly tested and related information. The *elaborative retrieval hypothesis* (Carpenter, 2009) suggests that retrieval practice can encourage the activation of both directly tested and related elaborative information. Relatedly, the *constructive retrieval hypothesis* (Hinze et al., 2013) proposes that engaging in constructive mental processes during practice testing bolsters memory for directly tested and related content to a greater extent than engaging in processing oriented around rote retrieval.

Multiple-Choice Practice Tests Can Bolster Learning of Directly Tested and Related Content

Research has explored whether certain test formats are more likely to elicit beneficial mental processes during practice testing than others. For example, a recently conducted meta-analysis found that recall-based test formats led to greater benefits of testing than recognition tests (Carpenter & DeLosh, 2006; Glover, 1989; Rowland, 2014). For years, multiple-choice tests were categorized as a type of recognition test and were therefore considered less beneficial for learning than short-answer or free-recall tests (Foos & Fisher, 1988; McDaniel, Anderson, et al., 2007). Indeed, several studies have found that multiple-choice tests are inferior to short-answer tests when learning is assessed using questions on previously targeted content (Butler & Roediger, 2007; McDaniel, Anderson, et al., 2007; McDaniel, Roediger, & McDermott, 2007). Little et al. (2012), however, demonstrated that multiple-choice items which are constructed to include competitive (i.e., highly plausible) incorrect alternatives can facilitate learning. In their series of studies, undergraduate participants read expository text passages (e.g., on Yellowstone National Park). Each passage contained several topics that included highly related content (e.g., facts about various *geysers* located within the park such as Steamboat Geyser, Castle Geyer, and Daisy Geyser). After reading the passages, participants took a multiple-choice practice test or a cued-recall test on some of the presented content from one passage, with content from the other passage serving as non-practiced control information. Shortly thereafter, participants took a final cued-recall test. Their results suggest that practice testing with well-constructed multiple-choice questions—with or without feedback—can enhance learning of directly tested content to a greater extent than practice testing with cued-recall test items. Furthermore, multiple-choice items consistently bolstered memory

for previously related content on the final test (e.g., information pertaining to Steamboat Geyser when Castle Geyser was practice tested on) whereas cued-recall practice test questions did not. The authors argued that the finding that multiple-choice practice tests can enhance memory for both directly tested and related content is particularly notable because in many educational contexts practice tests are not identical to the later test; rather, they are intended to facilitate students' recall of content when presented with similar questions on that content.

The benefit of multiple-choice practice testing to later memory for previously related content is thought to stem from the prompting of retrieval of information of both the correct response and the competitive, incorrect alternatives that are presented alongside the correct answer. When learners are offered noncompetitive multiple-choice questions during practice testing, no benefit to previously related content is observed, presumably because learners do not need to retrieve information pertaining to the incorrect competitive alternatives in order to select the correct response (Little & Bjork, 2010). In fact, when undergraduate participants taking a multiple-choice practice test recalled information pertaining to an incorrect competitive alternative, they answered the related question on the final cued-recall test correctly 75% of the time, whereas they did so only 35% of the time when they did not recall any pertinent information (Little et al., 2019).

The Pedagogical Utility of True-False Practice Tests May be Underappreciated

If multiple-choice tests have previously been criticized for their pedagogical utility, true-false tests have been positively vilified for (1) being a poor measure of learning and (2) not facilitating learning in the same way that multiple-choice or short-answer questions do.

True-false test questions are easy to construct, can be answered more rapidly than questions of other test formats, and are generally straightforward to grade compared to essay-

based test questions (Cocks, 1929). Yet, true-false testing has been criticized for being a poor measure of learning given the high likelihood of answering an item correctly at chance (Hevner, 1932), and because a single true-false item tends to discriminate between learners of different knowledge levels to a lesser extent than a multiple-choice test item (Ebel, 1975). When considering true-false tests as a whole (that is, across all items on the test), however, highly inflated scores due to guessing are actually quite unlikely and true-false tests can exhibit good psychometric properties (Burton, 2001; Ebel, 1970; Ebel, 1975).

True-false tests have also been criticized for failing to promote long-term learning. After finding that administering a multiple-choice pretest enhanced learning in a classroom setting, but administering a true-false pretest did not, Jersild (1929) declared that true-false tests have "dubious value as a pedagogical instrument" (pp. 608). This study, however, was contrasted by follow-up work that offered some evidence that true-false tests may offer learning benefits comparable to other "objective" tests like multiple-choice tests (e.g., Cocks, 1929; Hertzberg et al., 1932; Roberts & Ruch, 1928). For example, Standlee and Popham (1960) administered weekly true-false quizzes accompanied by a mid-semester multiple-choice assessment and an end-of-semester multiple-choice assessment. Weekly true-false quizzes enhanced performance on the mid-semester assessment compared to merely hearing the instructor read the questions aloud and answer them himself, but this benefit was only numerically—not significantly—observed on the end-of-semester assessment.

Though these studies offer some evidence that true-false tests might enhance learning, there were still doubts as to whether they, as recognition tests, could promote retrieval of content or durable learning. Consequently, there was generally little examination of the learning benefits from taking true-false tests in the second half of the twentieth century.

Recently, however, Brabec et al. (2021) has reopened investigation into the pedagogical utility of true-false practice tests. In Experiments 1 and 2, undergraduate learners read two expository text passages used previously in Little et al. (2012) and then took a true-false practice test on the content presented in one of the passages. After a brief delay, learners then took a final cued-recall test on content that was previously tested, previously related to tested content, or was untested (control).

Similar to multiple-choice practice testing, true-false practice testing did at times enhance later memory of previously tested and previously related content. Unlike multiple-choice practice testing, however, whether testing on the true-false practice item bolstered memory for previously tested or previously related content was highly dependent on whether the evaluated proposition was true or whether it was false. Evaluating true practice statements enhanced later memory for directly tested content (but not related content), whereas evaluating false practice statements enhanced later memory for previously related content (but not directly tested content). The authors suggested that this pattern of results demonstrated the "one-and-done" effect; in other words, that learners retrieved only as much information as was necessary to determine whether the proposition was true or whether it was false, and it was this specific pattern of retrieval during practice testing that resulted in the observed pattern of final test performance. When the practice test statement was true, learners generally only retrieved information about the directly tested concept to determine that it was true, and when the statement was false, learners generally only retrieved information about the related concept to determine that the statement was false.

In Experiment 3, Brabec et al. (2021) sought to undo the one-and-done effect by adding in competitive clauses to the true-false practice test items. These competitive clauses contrasted

two pieces of information in a "this-not-that" format yielding true-false practice test items that presented both directly tested and related content; e.g., Steamboat Geyser (not Castle Geyser) is the oldest geyser. Learners who practice tested using competitive-clause true-false items demonstrated broad benefits to both previously tested and previously related content regardless of whether the proposition was true or whether it was false, suggesting that the incorporation of competitive clauses may elicit broader retrieval than traditional true-false items.

True-False Practice Tests May Promote Learning of Inaccurate Information

Although there is now evidence that both traditional and competitive true-false tests can elicit retrieval processes and foster learning, an ongoing concern is whether false items within those tests could potentially lead learners to learn inaccurate information on a subsequent assessment (i.e., negative suggestion; Remmers & Remmers, 1926). Unlike short-answer questions, true-false tests by their very nature present false pairings of concepts roughly half the time. Doing so may leave a "residual of misinformation" with the learner that can be observed when the learner retrieves an intrusion or rates a false statement as true on a later test (Roberts & Ruch, 1928, pp. 112).

In one study, learners read text passages about US presidents and then were presented with a true-false test and asked to rate the items on a scale from *definitely false* to *definitely true* (Toppino & Brochin, 1989). One week later, learners were again presented with a true-false test which contained some repeated and some nonrepeated items and rated the truthiness of the presented statements. In line with the illusory truth effect (i.e., that repeated statements are perceived as more truthful than novel statements; Hasher et al., 1977), learners rated repeated statements as more truthful than nonrepeated statements, even if those statements were false. Further, learners demonstrate evidence of negative suggestion even when the statements

are rephrased, suggesting that negative suggestion is not tied to the mere repetition of identical statements (Toppino & Luipersback, 1993).

Toppino and Brochin (1989) also found that administering a two-alternative multiple-choice test resulted in significant negative suggestion, indicating that negative suggestion can occur from taking a multiple-choice test (likely due to the presence of incorrect competitive alternatives in each question; Butler, 2018; Butler & Roediger, 2007; Butler & Roediger, 2008 for offer detailed discussions). Subsequent research has found that negative suggestion from true-false practice questions can be reduced by providing feedback (Uner et al., 2021) and by incorporating competitive clauses into true-false items (Brabec et al., 2021).

The literature on negative suggestion has largely centered on the effect of a practice test presenting false pairings of content on learning of inaccurate associations. We would, however, like to offer a discussion on the distinction between this *test-driven* negative suggestion and *learner-driven* negative suggestion.

Test-driven negative suggestion (which is often simply labeled as negative suggestion in prior work) during practice testing can occur when an incorrect association between two pieces of information is "suggested," such as when two concepts are paired incorrectly to create a false proposition in a true-false test. Even if the learner retrieves accurate information during practice testing and identifies this false proposition as false (or is offered corrective feedback), its appearance during practice testing may still alter one's memory for content at the later final test (Brown et al., 1999; Uner et al., 2021; *the continued influence effect*, Johnson & Seifert, 1994; Rich et al., 2023). We can detect evidence of test-driven negative suggestion after-the-fact during a final test when learners respond in line with the inaccurate pairing that was previously suggested. For example, for the question *"What dwarf planet is located in the asteroid belt?,"* a

learner answers “Eris” (i.e., the closely matched response) rather than “Ceres” (i.e., the targeted response) following exposure to the relevant false practice item.

Alternatively, *learner-driven* negative suggestion can occur when the learner recalls an incorrect pairing of two concepts during practice testing. As an example, if presented with the correct practice test item *True or False? Ceres is a dwarf planet located in the asteroid belt*, the learner may incorrectly retrieve “Eris” as the dwarf planet located in the asteroid belt, leading them to erroneously evaluate this true statement as false and form an incorrect association between “Eris” and “asteroid belt.”

Retrieving incorrect pairings of content on a later test would most obviously be evidence of test-driven negative suggestion following evaluation of a *false* true-false proposition, whereas it would most obviously be evidence of learner-driven negative suggestion following evaluation of a *true* true-false proposition.

In summary, true-false tests have been criticized for being poor measures of learning and failing to prompt retrieval (and associated learning) to the same extent as other test formats. Although competitive-clause true-false items might facilitate broader retrieval of both previously tested and previously related content, traditional true-false items are simpler to write and are more commonly administered in educational contexts. Additionally, there are concerns as to the potential negative consequences of presenting learners with incorrect information during true-false practice testing.

Collaborative Practice Testing May Promote Beneficial Retrieval Processes

Although there are flaws with true-false practice tests, from an instructional perspective they are much easier to write than competitive multiple-choice questions. Thus, we asked whether it is possible to encourage productive retrieval and reduce negative suggestion in

learners taking true-false practice tests. Collaborative practice testing (i.e., testing in groups) has been shown to prompt a variety of potentially relevant processes compared to individual practice testing. When learners practice test in groups, there is the potential for re-exposure to and cross-cuing of information (Blumen & Rajaram, 2008; Blumen & Stern, 2011; LoGuidice et al., 2015). During re-exposure, a group member recalls information and shares it with the group, offering the rest of the group a restudy opportunity. During cross-cuing, a group member recalls information that prompts another group member to retrieve additional information, creating an opportunity for retrieval practice. More generally, studies of collaborative learning (of which collaborative testing is a type) suggest that working with others prompts learners to share knowledge by offering explanations (Aramovich & Larson, 2013), exchanging evidence (Clark et al., 2000), engaging in social justification of responses (Bruffee, 1984), and resolving conflict (Johnson et al., 1998). Engaging in these behaviors may facilitate activation of knowledge as well as constructive processes such as the reorganization of schemas (Chi & Wylie, 2014). Together, these processes may encourage learners to retrieve both directly tested and related content during true-false practice testing, even if those true-false practice items do not contain competitive clauses.

Further, working in groups can offer increased opportunity for the correction of errors, (i.e., error pruning; Hyman et al., 2013; Rajaram & Pereira-Pasarin; Ross et al., 2008). This error pruning may be particularly facilitated by open-ended, unstructured conversation (Rajaram, 2011). In one study conducted in a large STEM course, students answered an in-class conceptual question during a lecture break (Smith et al., 2009). Next, students discussed the question with their neighbors and then individually answered an isomorphic question with different surface details but testing the same concept as the initial question. Overall, students

learned from peer discussion of the initial question such that they were more likely to answer the subsequent isomorphic question correctly. Perhaps most notably, however, this benefit was present even if all members of the discussion group began peer discussion endorsing an incorrect alternative. It is possible that through discussion of each member's (incorrect) proffered response, students were able to correct one another's erroneous thinking and converge on the correct answer. Together, these studies suggest that collaboration during a true-false practice test could encourage error pruning and reduce the carryover of misinformation to a later test.

Given the potential benefits of collaboration, we examined whether taking true-false practice tests collaboratively can enhance students' recall for both tested and related information on a later test. In Experiments 1 and 2, we sought to extend investigation into the effects obtained in Brabec et al. (2021) in a controlled laboratory setting, and then in the authentic learning context of a large undergraduate STEM course (Exps. 3 and 4). We expected that, when practice testing individually, learners would show evidence of the one-and-done effect after taking a traditional true-false practice test (Exps. 1 and 3) and broader benefits to memory for previously tested and previously related content after taking a competitive-clause true-false practice test (Exps. 2 and 4).

Although no other work has examined the effect of collaborative practice testing with true-false items, based on the existing literature we hypothesized that collaboratively evaluating traditional true-false propositions may elicit retrieval of a greater variety of information than working individually, and thus increase final test performance on both previously tested and previously related items. Further, we expected that collaboration may offer some of the benefits of feedback when feedback is not provided (Exps. 1 and 2), or even potentiate the benefits of offering feedback after a practice test (Exps. 3 and 4), such that rates of negative suggestion

would be lower following collaborative practice testing as compared to individual practice testing.

Experiment 1

Experiment 1 occurred in a controlled laboratory setting and examined whether collaboration would undo the "one-and-done" effect and encourage learners to engage in broad retrieval during traditional true-false practice testing. It also investigated if practice testing with others would facilitate error correction (evidenced by lower rates of negative suggestion) as compared to working alone. As the one-and-done effect is identified by differences in the impact of true versus false initial practice, the validity of the true-false proposition was considered as a factor in this and all subsequent experiments. True practice refers to evaluation of a test item which is true (i.e., the correct response is to select "True") and false practice refers to evaluation of a test item which is false (i.e., the correct response is to select "False").

In this experiment, participants read two text passages and then took a traditional true-false practice test on one of those passages either (1) individually or (2) in small groups. Shortly thereafter, their knowledge for previously tested, previously related, and non-practiced control content was assessed individually using a final cued-recall test. Previously tested and previously related questions were differentiated by whether the cues that were presented during the true-false practice test were also presented during the final cued-recall test. If they were, then that item was classified as previously tested. If they were not, then that item was classified as previously related.

Methods

Participants

In line with the sample sizes used in Brabec et al. (2021), we aimed to test 50 participants per practice test setting. Participants were recruited from the psychology department subject pool at UCLA to participate in this study in exchange for course credit. The final sample (n = 110) included 61 participants in the Individual condition and 49 participants in the Collaborative condition. Data from 33 additional participants were excluded for the following reasons: for missing one or more attention checks (n = 17), for erroneously selecting the wrong version of their assigned final test (n = 6), for self-reported familiarity with the experimental materials (n = 3), and for self-reporting that their attention was divided during the study (i.e., they were texting during the study; n = 7).

The final sample included 89 women (80.9%) and 21 men (19.1%), and the average age of participants was 20.49 years (SD = 2.89). Participants most commonly identified as white (n = 35, 31.8%) and Asian/Pacific Islander (n = 33, 30.0%), followed by Latino/a/x (n = 22, 20.0%), bi/multiracial (n = 9, 8.2), Black (n = 4, 3.6%), and American Indian/Alaskan Native (n = 1, 0.9%). Six (5.5%) participants indicated that they identified as a race or ethnicity not listed.

Design

This study used a 2 (*Initial Practice*: True or False) x 2 (*Question Type*: Previously Tested or Previously Related) x 2 (*Practice Test Setting*: Individual or Collaborative) design. Initial practice and question type were within-subjects factors, and practice test setting was a between-subjects factor. Participants in the same Zoom session were all randomized to either complete the true-false practice test individually or complete it collaboratively.

Materials

Text Passages. Two ~1100-word passages on educational content (i.e., ferrets and the solar system) adapted from Little (2011) were used as stimuli. Each passage included at least eight categories of information (i.e., facts about *dwarf planets*, a type of planetary body), and within each category there were at least four distinct propositions per category (e.g., Ceres is… in the asteroid belt). These propositions were distributed throughout the passage (i.e., these propositions did not necessarily appear all together in a single paragraph).

True-False Practice Test. The practice test was drawn from a list of 64 true-false items (Table 1). These true-false items were created in sets of four, with two "true" and two "false" items per set, by combining two statements about the same category of information from one of the passages in different ways. For example, the solar system passage included two statements regarding dwarf planets: "Ceres is...in the asteroid belt" and "Eris is in the scattered disc." These two statements became the two "true" true-false items for this category of information. The false statements were created by combining the subject of each statement with the predicate of the other, yielding "Eris is a dwarf planet located in the asteroid belt" and "Ceres is a dwarf planet located in the scattered disc." Each passage contained eight categories of information, so there were 32 true-false items per passage for a total of 64 true-false items.

Cued-Recall Final Test. The final test was drawn from a list of 32 cued-recall items (Table 1 includes examples). These test items encompassed one set of two test items for each of the categories per passage. Each two-item set was created using the same two statements per category of information that were used to create the true-false items.

Table 1. Example items from the true-false practice tests and cued-recall final tests for Experiment 1 and Experiment 2.

	Practice Test Items		Final Test	
Exp.	Initial Practice	Example	Question Type	Example *(answer)*
1	True	Ceres is a dwarf planet located in the asteroid belt.	Previously Tested	What dwarf planet is located in the asteroid belt? *(Ceres)*
			Previously Related	What dwarf planet is located in the scattered disc? *(Eris)*
	False	Eris is a dwarf planet located in the asteroid belt.	Previously Tested	What dwarf planet is located in the scattered disc? *(Eris)*
			Previously Related	What dwarf planet is located in the asteroid belt? *(Ceres)*
2	True	Ceres (not Eris) is a dwarf planet located in the asteroid belt.	Previously Tested	What dwarf planet is located in the asteroid belt? *(Ceres)*
			Previously Related	What dwarf planet is located in the scattered disc? *(Eris)*

False	Eris (not Ceres) is a dwarf planet located in the asteroid belt.	Previously Tested	What dwarf planet is located in the scattered disc? *(Eris)*
		Previously Related	What dwarf planet is located in the asteroid belt? *(Ceres)*

Procedure

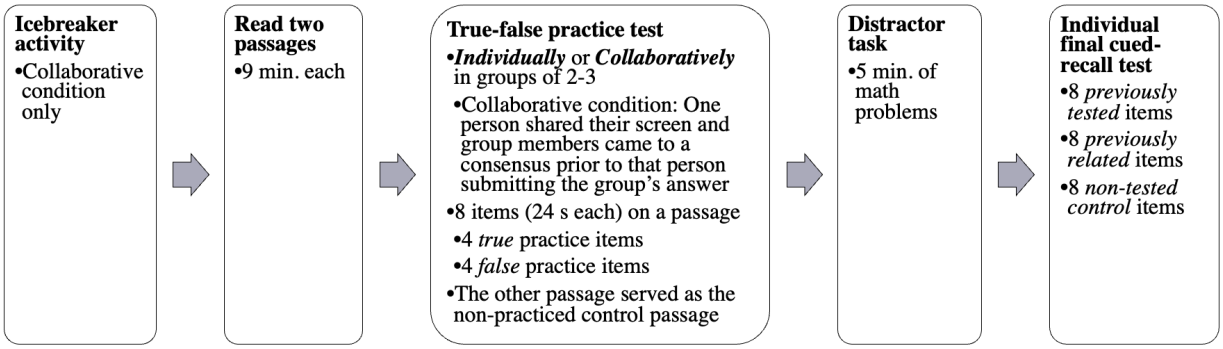


Figure 3. Diagram of the procedure used in Experiments 1 and 2.

The experiment was conducted in one session online via Zoom in groups of 2-3 (Figure 3). The procedure used here closely aligned with the procedure of Brabec et al. (2021). Experimental materials were presented using the open-source software Collector (Garcia & Kornell, 2014). To maintain participant privacy, participants changed their Zoom moniker from their name to their assigned participant ID before joining the Zoom session. Participants assigned to the Individual condition proceeded immediately to the initial study phase of the experiment, whereas students assigned to the Collaborative condition first completed a 2-min icebreaker activity (i.e., find one thing you all have in common) to facilitate more effective group

work during the collaborative practice test. In the initial study phase, participants in both conditions individually studied the experimental passages for nine minutes each, one after the other. Which passage was presented first was counterbalanced across participants.

The next phase of the experiment was the true-false practice test. Prior to beginning the practice test, participants in the Collaborative condition were placed into breakout rooms so that they were not observed by the experimenter and one group member was randomly assigned to share their screen with the other group members. Those in the individual condition remained in the main room. The 8-item true-false practice test was on one of the two passages, with the other passage serving as the non-practiced control passage. Which passage was practice tested on and which passage was assigned to be the control passage was counterbalanced across participants. Participants spent exactly 24 s on each item and were not given feedback. Selection of practice test items was constrained to ensure that for four of the items the correct response was "True" and for the other four items the correct response was "False," and such that only one practice question per category of information from the tested passage was included. Which items were presented was counterbalanced across participants.

After they finished the practice test, those in the collaborative condition returned to the main session. From this point on, the rest of the experiment was completed individually regardless of a participant's condition. Participants next solved math problems for five minutes as a distractor task. Finally, all participants completed the 24-item cued-recall final test individually.

The cued-recall test included questions on previously tested information and on previously related (but not directly tested) information. Questions were presented in separate blocks of 12 items (i.e., a previously tested block and a related block. The previously tested

block included eight items which featured descriptions that were previously tested during the true-false practice test. Half of these items corresponded to practice test items that were true and half to practice test items that were false. In the related block, eight of the items featured descriptions that were not previously tested during the true-false practice test but belonged to the same categories of information as descriptions that were directly tested. Again, half of these items corresponded to practice test items that were true and half to practice test items that were false.

The remaining four final test items in each block assessed memory for content from the non-practiced control passage. For non-practiced control questions, the distinction between previously tested and previously related was artificial (as there was no prior testing on that content). Therefore, whether a control question was considered previously tested or previously related for the purposes of computing net effects of practice was based on whether the control item was presented in the previously tested block or the previously related block. The order of items within each block and the order of blocks was counterbalanced across participants to account for possible order effects.

Final cued-recall test items were answered one at a time and participants were required to spend 20 seconds on each item.

Results and Discussion

Practice test performance for Experiments 1-4 is available in Appendix B.

Final Test Performance

In order to examine whether participants who completed the true-false practice test in groups versus alone showed different patterns of recall on the subsequent cued-recall test, we conducted a 2 (*Initial Practice*: True or False) x 2 (*Question Type*: Previously Tested or

Previously Related) x 2 (*Practice Test Setting:* Individual or Collaborative) ANOVA using IBM SPSS 28.0 software (Figure 4). Initial practice and question type were within-subjects factors and practice test setting was a between-subjects factor. The net effect of practice was the dependent variable, and it was computed by subtracting final test performance on non-practiced control items from final test performance on practiced items (i.e., *net effect of practice = final test performance on practiced items – final test performance on non-practiced control items*).

There was a significant 3-way interaction, $F\ (1, 108) = 36.02, p < .001, \eta_p^2 = .25$, so we looked at the 2-way interaction between initial practice and question type separately within each level of practice test setting.

Individual. These analyses included only data from participants who completed the practice test alone. For participants who completed the practice tests individually, the interaction between initial practice and question type was significant, $F\ (1, 60) = 45.43, p < .001, \eta_p^2 = .43$, and visually appears to be a cross-over interaction. Therefore, follow-up paired-samples t-tests were conducted to examine the effect of question type within each level of initial practice on the net effect of practice versus control. For true initial practice, participants performed significantly better on previously tested questions ($M = .17, SD = .29$) than on previously related questions ($M = .00, SD = .29$), $t\ (60) = 3.72, p < .001, d = 0.48$, 95% CI [0.21, 0.74]. This pattern demonstrates a considerable benefit of true initial practice for previously tested questions, with a near-zero net effect of practice for previously related questions. Conversely, for false practice, participants performed significantly better on previously related questions ($M = .17, SD = .32$) than on previously tested questions ($M = -.04, SD = .28$), $t\ (60) = -5.26, p < .001, d = -0.67$, 95% CI [-0.95, -0.39]. In other words, false initial practice led to substantial benefits to performance on previously related questions but a near-zero net effect of practice for previously tested

questions. This pattern of results is fully aligned with our hypothesis, replicates the pattern obtained in Brabec et al. (2021), and offers new evidence of the one-and-done effect as an outcome of individual true-false practice testing.

Collaborative. These analyses included only data from participants who completed the practice test in small groups. For the subset that completed the practice test in groups, the interaction between initial practice and question type was nonsignificant, $F\ (1, 48) = 3.07$, $p = .086$, $\eta_p^2 = .06$, so main effects were examined.

The nonsignificant main effect of question type suggested that true initial practice similarly benefitted performance on previously tested ($M = .13$, $SD = .30$) and previously related ($M = .17$, $SD = .34$) final test items, $F\ (1, 48) = 0.02$, $p = .89$, $\eta_p^2 < .001$. Likewise, false initial practice was comparably beneficial to performance on previously tested ($M = .07$, $SD = .30$) and previously related ($M = .02$, $SD = .29$) questions. There was, however, a main effect of initial practice, $F\ (1, 48) = 8.52$, $p = .005$, $\eta_p^2 = .15$. On average, true initial practice had a significantly higher net benefit to final test performance than false initial practice.

The pattern of final test performance following collaborative practice testing condition only partially supports our hypothesis. We anticipated that collaborative practice testing would benefit learning across both types of initial practice and for both types of questions. Although true initial practice broadly enhanced learning, false practice did not.

Retrieval of The Incorrect Competitive Alternative on The Final Test

To better understand the nature of learners' memory for studied content, we conducted an exploratory analysis of retrieval of the incorrect competitive alternative on the final cued-recall test. As discussed in the Introduction, this is often thought to occur due to negative suggestion during practice; i.e., the pairing of two incorrect concepts in memory.

The analyses on retrieval of the incorrect competitive alternative used a net effect as the dependent variable just like the analyses on final test performance described above. This net effect was computed using the equation: *net retrieval of incorrect competitive alternative = retrieval of the incorrect competitive alternative on practiced questions – retrieval of the incorrect competitive alternative on non-practiced control questions*.

In an ANOVA examining the impact of initial practice, question type, and practice testing setting on net retrieval of the incorrect competitive alternative, there was a significant 3-way interaction, $F\ (1, 108) = 25.13, p < .001, \eta_p^2 = .19$. Two-way interactions between initial practice and question type were therefore examined within each level of practice test setting separately.

Individual. For those who completed the practice test items individually, the initial practice x question type interaction on net retrieval of the incorrect competitive alternative was significant, $F\ (1, 60) = 43.49, p < .001, \eta_p^2 = .43$, so follow-up paired-samples t-tests were conducted to examine the effect of question type following true initial practice and following false initial practice separately.

Overall, participants who practice tested individually demonstrated net retrieval of the incorrect competitive alternative that was substantially greater than zero only for previously tested questions after false initial practice. Following true initial practice, participants demonstrated similarly low net retrieval of the incorrect competitive alternative for previously tested questions ($M = .004, SD = .13$) and previously related questions ($M = .05, SD = .19$), $t\ (60) = -1.52, p = .14, d = -.19$, 95% CI [-.45, .06]. In contrast, following false initial practice, participants demonstrated significantly higher net retrieval of the incorrect competitive

alternative for previously tested questions (M = .18, SD = .22) than for previously related questions (M = -.05, SD = .14), t (60) = 6.77, p < .001, d = .87, 95% CI [.57, 1.16].

Collaborative. In the subset of data from participants who completed the true-false practice test in groups, the initial practice x question type interaction on net retrieval of the incorrect competitive alternative was nonsignificant, F (1, 48) = 0.68, p = .42, η_p^2 = .01, so main effects were examined. Participants demonstrated similar net retrieval of the incorrect competitive alternative across previously tested and previously related items, F (1, 48) = .04, p = .84, η_p^2 = .001, but demonstrated greater net retrieval of the incorrect competitive alternative following false initial practice as compared to true initial practice, F (1, 48) = 7.89, p = .007, η_p^2 = .14.

Table 2. Descriptive statistics for Exps 1-4

Practice Test Setting	Initial Practice	Question Type	Final Test Performance (Net)[a] *M (SD)*				Retrieval of Incorrect Competitive Alternative (Net) *M (SD)*			
			Exp 1	Exp 2	Exp 3	Exp 4	Exp 1	Exp 2	Exp 3	Exp 4
Individual	True	Previously Tested	.17 (.29)	.15 (.30)	.17 (.47)	.17 (.46)	.004 (.13)	.11 (.21)	-.001 (.28)	-.007 (.26)
		Previously Related	.00 (.29)	.13 (.27)	-.03 (.50)	.10 (.49)	.05 (.19)	.08 (.17)	.13 (.38)	.03 (.32)
	False	Previously Tested	-.04 (.28)	.12 (.31)	-.01 (.49)	.12 (.49)	.18 (.22)	.10 (.20)	.12 (.39)	.01 (.31)
		Previously Related	.17 (.32)	.11 (.28)	.15 (.49)	.13 (.51)	-.05 (.14)	.07 (.16)	.00 (.27)	.02 (.32)
Collaborative	True	Previously Tested	.13 (.30)	.14 (.31)	.09 (.48)	.15 (.49)	.05 (.17)	.01 (.17)	.02 (.31)	.003 (.30)

	Previously Related	.17 (.34)	.14 (.35)	.06 (.51)	.10 (.52)	.02 (.18)	-.02 (.20)	.04 (.32)	.04 (.32)
False	Previously Tested	.07 (.30)	.18 (.33)	.07 (.47)	.12 (.50)	.08 (.17)	.02 (.21)	.04 (.33)	.02 (.32)
	Previously Related	.02 (.29)	.14 (.33)	.09 (.48)	.12 (.50)	.10 (.22)	.01 (.22)	.04 (.31)	.008 (.29)

[a]Net effect = practice – control

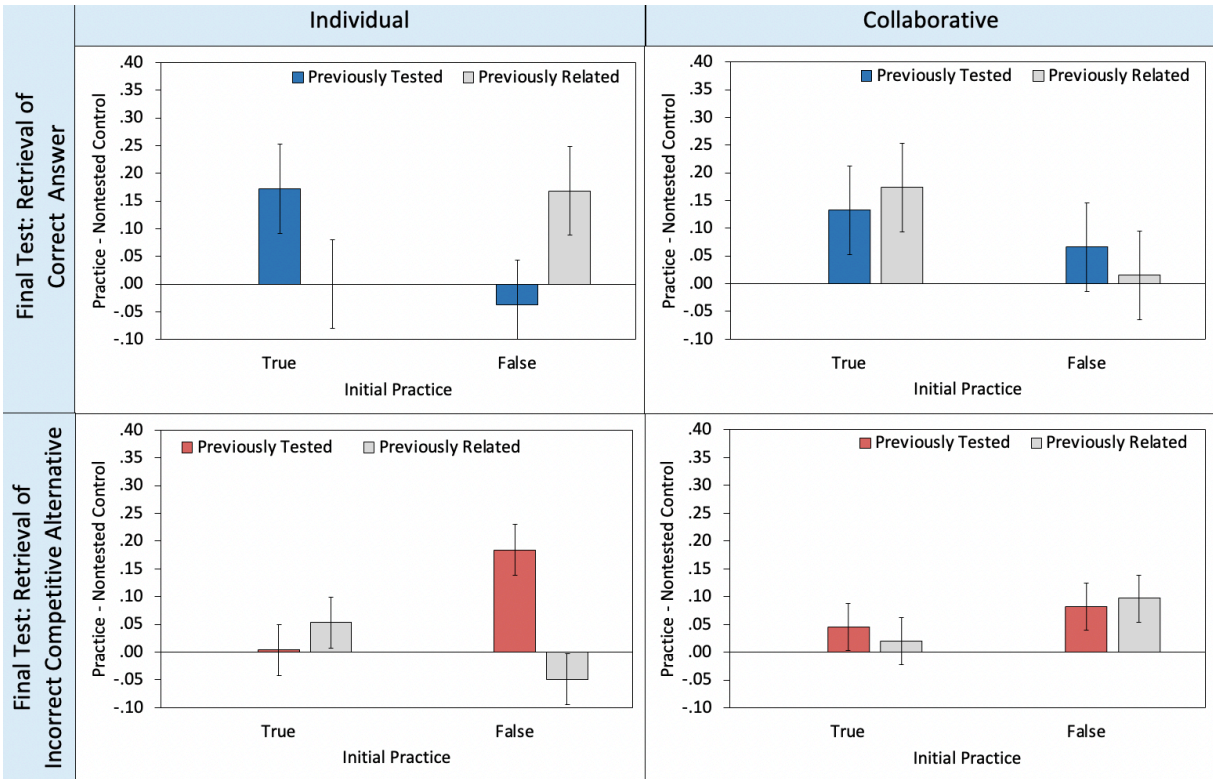


Figure 4. Net effects of practice testing on the final cued-recall test in Experiment 1. Error bars represent 95% confidence intervals.

In Experiment 1, collaborative practice testing with true items enhanced learning relative to no practice testing, but practice testing with false items yielded no net benefit to final test performance.

Participants who tested individually demonstrated evidence of the one-and-done effect, replicating the results of Brabec et al. (2021, Exp. 1). This pattern of results suggests that, when participants practice tested individually, they retrieved only as much information as was necessary to determine whether the statement should be evaluated as true or as false. For true statements, this tendency enhanced memory for the directly tested concept on the final cued-recall test, as participants may have retrieved information directly targeted in the practice question to determine that the statement was true (e.g., *True or False? Ceres is a dwarf planet*

located in the asteroid belt. "Yes, this is true, Ceres is a dwarf planet located in the asteroid belt"). In contrast, for false statements, this tendency enhanced memory for previously related concepts on the final cued-recall test, as false statements may have prompted participants to retrieve information conceptually related to the target information to determine that the statement was false (e.g., *True or False? Eris is a dwarf planet located in the asteroid belt*. "No, that is not true, Ceres is a dwarf planet located in the asteroid belt").

Participants who practice tested collaboratively in Experiment 1 demonstrated a substantially different pattern of final test results. First looking at final test performance following collaborative, true practice, the pattern does not align with the one-and-done effect. Participants demonstrated enhanced final test performance for both previously tested and previously related questions following true practice. This difference suggests that participants may have engaged in different patterns of retrieval during collaborative as compared to individual practice testing. Perhaps discussing with others encouraged participants to retrieve a broader array of information than they would have otherwise if they had practice tested alone. If this broader array of information included key related content, doing so may have bolstered memory for both previously tested and previously related information.

Although we found broad learning benefits for collaborative practice of true items, we were surprised to find that there was no net benefit to learning of collaborative practice of false items. Why did participants not demonstrate a net benefit of practice relative to non-practiced control for either question type following false practice? The pattern of final test results regarding retrieval of the incorrect competitive alternative may offer a potential explanation. As seen in Figure 4, participants demonstrated low rates of retrieving the incorrect competitive alternative (which were not different from zero) following collaborative, true practice. In

contrast, participants demonstrated positive net retrieval of the incorrect competitive alternative following collaborative, false practice. In other words, participants' final test responses demonstrated evidence of test-driven negative suggestion, but not learner-driven negative suggestion.

The presence of test-driven negative suggestion here is notable because, if participants in the collaborative condition were showing complete retrieval failure following false practice (i.e., participants simply could not recall information presented in the practice test item and/or retrieved during the discussion of that practice test item), then we would expect both final test performance *and* rates of retrieving the incorrect competitive alternative to be low (i.e., participants would output nothing or a completely unrelated concept during the final cued-recall test). Instead, we see that the decrease in final test performance roughly corresponds to an increase in retrieving the incorrect competitive alternative.

Possibly, the increased amount of information accessed during discussion between group members made it difficult for individuals to track which discussed information was true and which discussed information was false. Prior work suggests that individuals can demonstrate errors in source memory. For example, even when individuals read information that is labeled as "false," they may still output that information on a later test due to a failure in maintaining the association between the information and its "false" label (Skurnik et al., 2005; also *the sleeper effect*; Hovland et al., 1949).

This phenomenon may have been exacerbated by the nature of the traditional T/F practice item. For these false practice items (e.g., *Eris is a dwarf planet located in the asteroid belt*), the incorrect pairing of "Eris" and "asteroid belt" is suggested. Participants may be able to recall that, in fact, *Ceres* is a dwarf planet located in the asteroid belt and correctly evaluate that

statement as false, but this correction may not be sufficiently memorable—especially if it comes from another group member not one's own retrieval—as the correct concept for that pairing (*Ceres*) is not presented to participants during practice testing.

The pattern of results involving test-driven negative suggestion following individual practice testing supports the explanation that decreases in final test performance may be driven by participants incorrectly replacing a correct pairing for a false one. Following true practice, evidence of negative suggestion was relatively rare. This result is expected, as a true practice test item does not suggest an incorrect pairing between two pieces of information, and we would generally expect spontaneous errors (i.e., learner-driven negative suggestion) to occur infrequently. Following false practice, however, evidence of test-driven negative suggestion was quite high for previously tested questions and relatively rare for previously related questions. As was the case following collaborative practice testing, an increase in net test-driven negative suggestion roughly corresponded to a decrease in final test performance, suggesting that the two are related.

Offering both the directly tested and the key incorrect alternative by embedding competitive clauses in true-false practice items may lead to a different pattern of results; e.g., *Ceres (not Eris) is a dwarf planet located in the asteroid belt*. Including both concepts in the practice test item may facilitate broader retrieval of information during practice testing, as observed in Brabec et al. (2021, Exp. 3), because the explicit contrasting of these two concepts may prompt participants to retrieve information regarding both. This prompting may be particularly impactful for those practice testing individually, as they demonstrated a very specific pattern of learning benefits in Experiment 1. Offering a competitive-clause practice test item may also reduce negative suggestion at final test, especially following false initial practice, as the

competitive clauses present the correct information that could be used to "replace" the incorrect suggested pairing in memory. This additional information may be especially impactful in reducing test-driven negative suggestion in the Collaborative condition.

Experiment 2

Experiment 2 investigated whether adding competitive clauses to true-false practice test items would (1) encourage broad retrieval of both tested and related content during practice testing by learners in the Individual condition and (2) reduce negative suggestion following false initial practice by learners in the Collaborative condition. Experiment 2 closely followed the approach of Experiment 1 except that participants practice tested with competitive-clause true-false items rather than traditional true-false items. In this experiment, participants read two text passages, took a competitive-clause true-false practice test on one of those passages individually or in small groups, and then answered previously tested, previously related, and non-practiced control questions on a final cued-recall test.

Methods

Participants

One hundred and seventeen participants were included in the final sample (Individual: n = 63; Collaborative: n = 54). Data from an additional 22 participants were excluded because the participant missed one or more attention checks (n = 9), there were technical issues with the study (n = 7), the participant reported prior familiarity with the experimental materials (n = 5), or because the participant did not complete the full study (n = 1).

Participants in the final sample were on average 20.30 years old (SD = 2.41). Most of the sample (n = 90; 76.9%) identified as women, followed by men (n = 25, 21.4%), and those who identified as another gender identity (n = 2, 1.7%). The sample most commonly identified as

Asian/Pacific Islander (n = 48, 41.0%), white (n = 36, 30.8%), Latino/a/x (n = 20, 17.1%), bi/multiracial (n = 7, 6.0%), and Black (n = 3, 2.6%); three participants (2.6%) reported that they identified with a race or ethnicity not listed.

Design, Materials, and Procedure

The only change from Experiment 1 to Experiment 2 was the incorporation of competitive clauses within the true-false items (Table 1 includes examples).

Results and Discussion

Final Test Performance

Just as in Experiment 1, a 2 (*Initial Practice*: True or False) x 2 (*Question Type*: Previously Tested or Previously Related) x 2 (*Practice Test Setting*: Individual or Collaborative) 3-way ANOVA with net effect of practice relative to no practice on the final cued-recall test as the dependent variable was used to evaluate participants' learning of passage content (Figure 5). The descriptive statistics of the net practice effects suggest that substantial learning did occur from practice testing. This learning, however, did not seem to be impacted by whether initial practice was true or false, the practice test setting was individual or collaborative, or the question type was previously tested or related; in other words, the interactions and main effects in the model were nonsignificant (all p's > .32).

The pattern of results obtained in Experiment 2 reflect broad benefits to learning following competitive T/F practice testing as compared to no practice testing. There were similar and positive net effects for final test performance across all levels of initial practice, question type, and practice test setting.

Retrieval of The Incorrect Competitive Alternative on The Final Cued-Recall Test

In line with the analyses of Experiment 1, we again examined the impact of experimental factors on negative suggestion on the final cued-recall test. As the interactions in the model were nonsignificant (all p's > .36), main effects of question type, initial practice, and practice test setting were examined.

The nonsignificant main effect of question type suggests that net retrieval of the incorrect competitive alternative was similar across previously tested and previously related items on the final cued-recall test, F (1, 115) = 1.38, p = .24, η_p^2 = .01. In contrast to Experiment 1, where net retrieval of the incorrect competitive alternative was higher following false initial practice than true initial practice, the main effect of initial practice was nonsignificant, suggesting that net negative suggestion was similar regardless of whether the practice item was true or false, F (1, 115) = 0.07, p = .79, η_p^2 = .001. The significant main effect of practice test setting, however, suggests that net retrieval of the incorrect competitive alternative was significantly higher following individual practice testing than following collaborative practice testing: In fact, participants who practice tested collaboratively demonstrated near-zero net retrieval of the incorrect competitive alternative across all levels of initial practice and question type, F (1, 115) = 17.33, p < .001, η_p^2 = .13.

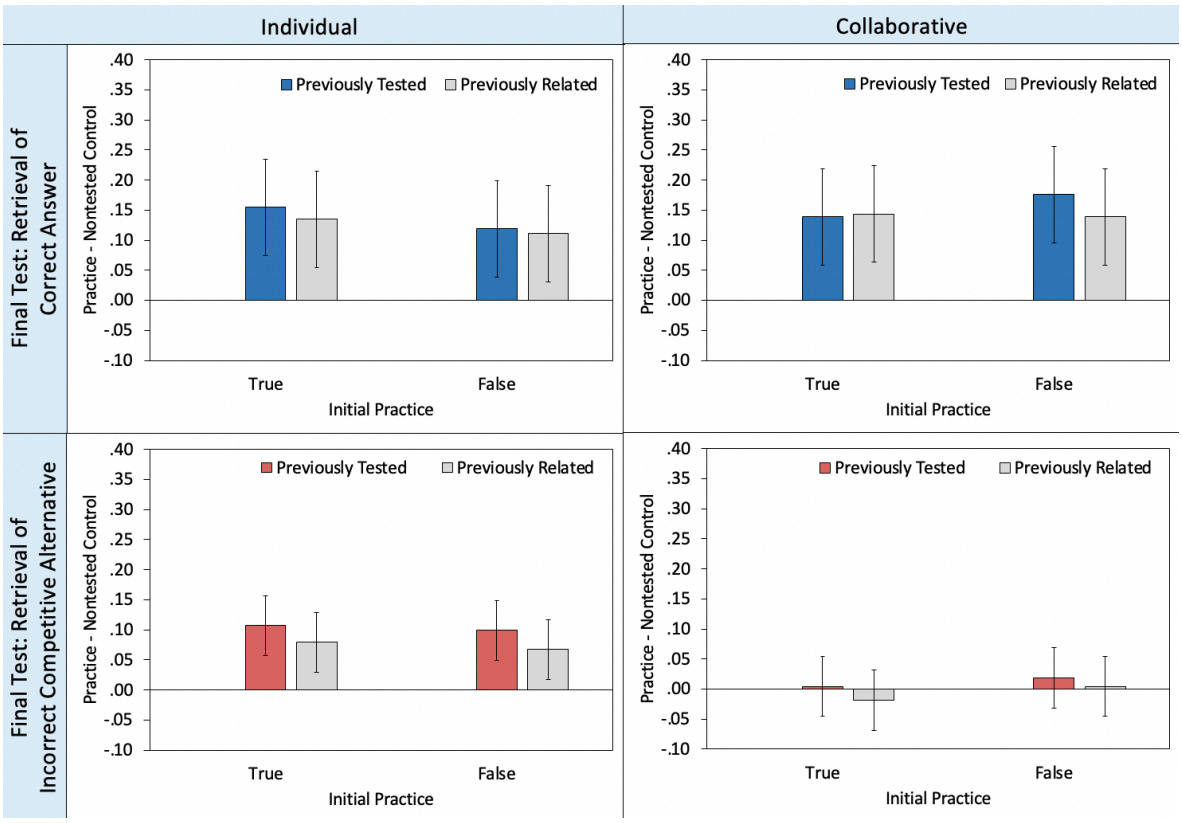


Figure 5. Net effects of practice testing on the final cued-recall test in Experiment 2. Error bars represent 95% confidence intervals.

The results of Experiment 2 demonstrate broad benefits of practice testing to learning compared to no practice testing regardless of initial practice, type of question, or practice test setting. These results extend those obtained in Brabec et al. (2021, Exp. 3) and suggest that competitive true-false items may elicit more comprehensive retrieval than traditional true-false items when participants practice test individually and when they practice test collaboratively.

The patterns obtained in Experiment 2 following collaborative practice testing are noticeably different than those obtained in Experiment 1. In Experiment 1, collaborative practice testing with true items enhanced learning relative to no practice testing, but practice testing with false items yielded no net benefit to final test performance. Additionally, collaborative practice testing with false items resulted in evidence of test-driven negative suggestion on the final cued-

recall test for both previously tested and previously related questions. In Experiment 2, collaborative practice testing yielded a net benefit to learning following both true and false initial practice, and participants demonstrated very little (i.e., not significantly different from zero) net negative suggestion. It is possible that the inclusion of competitive clauses in the true-false practice items used in Experiment 2 supported participants in tracking correct and incorrect pairing of information.

Strikingly, rates of both learner-driven and test-driven negative suggestion were lower following collaborative practice testing than following individual practice testing. One potential benefit of collaborative testing is error correction (LoGuidice et al., 2015). Why then was this benefit not apparent in Experiment 1? It is possible that participants' knowledge of the topic in both Experiment 1 and Experiment 2 was not particularly robust—they were only given nine minutes to read each passage. Group members may therefore have struggled to leverage the information they read in the passage to strongly correct errors during practice testing, particularly when confronted with false practice test items. In Experiment 2, however, the information required for participants to construct the correct pairing of information was included in the false practice items. This extra information may have fostered more successful evaluation of false items during practice (as evidenced by the increase in practice test performance for false items from Experiment 1 to Experiment 2) and yielded longer-lasting error correction following collaborative practice testing (Ecker et al., 2010; Johnson & Seifert, 1994).

Overall, the results of Experiments 1 and 2 suggest that participants' memory for content is highly influenced by the items that they practice test on. They also suggest that individual and collaborative practice testing can yield different patterns of memory. A limitation of these experiments, however, is that they were conducted under highly controlled conditions.

Participants were unfamiliar with the materials prior to the study and were given limited opportunity to learn passage content prior to practice testing. Those in the collaborative condition were also unfamiliar with each other, which may have implications for how group members interacted. The experimental session was also conducted online via Zoom and participants could only spend 24 s on each practice question, which may have constrained how group members interacted with one another or navigated the interpersonal aspects of the practice testing session (e.g., interpreting body language, indicating disagreement with a proposed response). Finally, participants did not receive feedback, potentially limiting their ability to learn from errors made during practice testing.

Experiment 3

In Experiments 3 and 4, we tested if the patterns of results obtained in Experiments 1 and 2 would be obtained in an authentic learning context. These experiments were run during a class session of Introductory Psychology at UCLA, offering the opportunity to examine the effect of both collaborative and individual true-false practice testing in a classroom setting where (1) groups collaborated in-person, (2) group members knew each other, (3) students were likely to have a strong understanding of tested content, (4) students were not limited to a particular amount of time per practice test question, and (5) feedback was provided following each practice test.

Methods

Participants

Undergraduate students (n = 508) across two sections of a large Introductory Psychology course at UCLA participated in this study as part of an in-class activity. Thirty additional

students participated in the study, but their data were excluded because students had missing data on one or more of the tests.

Design

Though the factors used in this study were the same as in Experiments 1 and 2, whether students practice tested individually versus collaboratively was a within-subjects factor rather than a between-subjects factor in Experiment 3. As in Experiments 1 and 2, true practice refers to the evaluation of a test item which is true (i.e., the correct response is to select "True") and false practice refers to the evaluation of a test item which is false (i.e., the correct response is to select "False"). Likewise, previously tested and previously related questions are differentiated by whether the cues that were presented during the true-false practice test were also presented during the final cued-recall test. If they were, then that item was classified as previously tested. If they were not, then that item was classified as previously related (Table 3). Course topics were counterbalanced such that each topic was either practiced tested on collaboratively, practice tested on individually, or was not practice tested on (i.e., the control topic). Practice test items within each course topic were counterbalanced across participants with the constraints that half of the items would become previously tested and half of the items would become previously related, half of the items were true and half of the items were false, and that only one practice test item pertained to a matched pair of concepts (e.g., participants could receive one practice question on EEG or one practice question on fMRI but never both).

Materials

The materials used in this study were centered on three course topics: Biological Psychology, Sensation and Perception, and Research Methods. Within each topic, six pairs of concepts were selected to be used in this study, such as fMRI versus EEG (Biological

Psychology), experimenter bias versus demand characteristics (Research Methods), and top-down processing versus bottom-up processing (Sensation and Perception). These pairs of concepts were selected because the course instructors (M.P. and C.M.C) noticed that students often confused them with one another on course assessments.

True-False Practice Test. Practice test questions incorporated a brief cover story (e.g., a description of a study where there is experimenter bias) followed by a traditional T/F statement for students to evaluate as either "True" or "False." Each practice test was drawn from a list of 24 true-false items. These true-false items were created in sets of four, with two "true" and two "false" items per set, by combining two statements about the same pair of concepts from one of the course topics in different ways (Table 3). One true statement tested one concept within the pair (e.g., EEG) and the other true statement tested the other (e.g., fMRI). The false statements were created by taking the true statement and replacing the correct member of the pair with the incorrect member of the pair (e.g., replacing EEG with fMRI or vice versa). There were six sets of true-false items (one for each pair of concepts within a topic), yielding a total of 24 true-false items per course topic and 72 total practice items across the three course topics.

Cued-Recall Final Test. Final cued-recall questions were based on selected course concepts and required participants to type in a 1–2-word answer (Table 3). For each pair of concepts, one was chosen to be tested on in the final cued-recall test, resulting in an 18-item final cued-recall test, six items pertaining to each course topic. Unlike in Experiments 1 and 2, this approach allowed all students to take the same final cued-recall test.

Table 3. Example items from the true-false practice tests and cued-recall final test for Experiments 3 and 4.

Exp.	Initial Practice	Question Type	Practice Question	Final Test Question
3	True	Previously Tested	Quen wants to know where activation increases within the visual cortex while people are reading. True or False? The best methodology for Quen's study is fMRI.	Joe wants to know where activation increases within the frontal lobe when solving a logic puzzle. Which methodology would be best suited to Joe's study. *(fMRI)*
		Previously Related	Danielle wants to know if the time-course of activation in response to viewing faces differs from that of household objects. True or False? The best methodology for Danielle's study is EEG.	
	False	Previously Tested	Quen wants to know where activation increases within the visual cortex while people are reading. True or False? The best methodology for Quen's study is EEG.	
		Previously Related	Danielle wants to know if the time-course of activation in response to viewing faces differs from that of household objects. True or False? The best methodology for Danielle's study is fMRI.	
4	True	Previously Tested	Quen wants to know where activation increases within the visual cortex while people are reading. True or False? The best	

		methodology for Quen's study is fMRI (not EEG).
	Previously Related	Danielle wants to know if the time-course of activation in response to viewing faces differs from that of household objects. True or False? The best methodology for Danielle's study is EEG (not fMRI).
False	Previously Tested	Quen wants to know where activation increases within the visual cortex while people are reading. True or False? The best methodology for Quen's study is EEG (not fMRI).
	Previously Related	Danielle wants to know if the time-course of activation in response to viewing faces differs from that of household objects. True or False? The best methodology for Danielle's study is fMRI (not EEG).

Procedure

The three course topics practice tested on in this study were covered in lecture prior to the in-class testing activity. After being introduced to the practice testing activity, students were put in their lab groups of 3-5 students, which they had worked with previously during other in-class activities. All students first completed a collaborative practice test, then an individual practice test, and then an individual final cued-recall test. All tests were presented using Qualtrics

(https://www.qualtrics.com/) and were self-paced. Due to the logistical challenges of manipulating the order of the group and individual tasks (e.g., providing different verbal instructions to different groups of students in a 250-person classroom), the collaborative test always occurred first. In setting up the study in this way, we inherently stack the deck against collaborative testing given the longer delay between practice and the final cued-recall test for those items as compared to items practice tested on individually.

For the collaborative practice test, one student had the practice test open on their own computer. Students were told that they should discuss all answers with their group members and come to a consensus on each answer before moving onto the next question. Each question was presented one at a time and students could not go back after answering the question. After answering the six practice test items, students received feedback. After completing the collaborative practice test, students immediately proceeded to the individual practice test on a different course topic. Students were instructed to open the individual test on their own laptop computers and to work alone. They answered the six practice test items, then received feedback. Once both practice tests were completed, students opened the link to the individual final cued-recall test. Students then completed a 5-min distractor task in which they answered short answer questions about course topics not included in this study (e.g., operant conditioning). Then, students took the individual final cued-recall test. The final cued-recall test questions were presented one at a time and students could not go back after answering the question. After completing the final test, students answered one question about how their group worked together and then received feedback.

Results and Discussion

Final Test Performance

To investigate the effect of taking competitive-clause true-false practice tests on learning, a 2 (*Initial Practice*: True or False) x 2 (*Question Type*: Previously Tested or Previously Related) x 2 (*Practice Test Setting*: Individual or Collaborative) 3-way ANOVA was conducted (Figure 6). All factors were within-subjects. As in the preceding experiments, net effect of practice was the dependent variable, which was computed by subtracting final test performance on non-practiced control items from final test performance on practiced items.

There was a significant 3-way interaction, F (1, 507) = 37.69, $p < .001$, $\eta_p^2 = .07$. Due to this interaction, we examined the 2-way interaction between initial practice and question type separately within each level of practice test setting.

Individual. For the subset of items that were practiced individually, there was a significant 2-way initial practice x question type interaction, F (1, 507) = 107.36, $p < .001$, $\eta_p^2 = .18$. Therefore, follow-up paired-samples t-tests were conducted to examine the effect of question type within each level of initial practice.

For true initial practice, participants performed significantly better on previously tested questions ($M = .17$, $SD = .47$) than on previously related questions ($M = -.03$, $SD = .50$), t (507) = 8.06, $p < .001$, $d = 0.36$, 95% CI [0.27, 0.45]. In contrast, for false initial practice, participants performed significantly better on previously related questions ($M = .15$, $SD = .49$) than on previously tested questions ($M = -.01$, $SD = .49$), t (507) = -6.45, $p < .001$, $d = -0.29$, 95% CI [-0.38, -0.20]. This pattern of results replicates that obtained in Experiment 1—in a controlled laboratory setting—in a real-world classroom and offers evidence of the one-and-done effect even when feedback is provided after the practice test.

Collaborative. For the subset of items that were practiced collaboratively, practice testing yielded positive net effects (relative to no practice) for all combinations of initial practice and practice question type (Table 2). For the items that were previously practiced in groups, the initial practice x question type interaction was nonsignificant, F (1, 507) = 2.62, p = .106, η_p^2 = .005. We therefore report main effects. Participants demonstrated similar amounts of learning for previously tested and previously related content, F (1, 507) = 0.20, p = .66, η_p^2 = < .001, and similar amounts of learning following true initial practice and false initial practice, F (1, 507) = 0.22, p = .64, η_p^2 = < .001.

Retrieval of the Incorrect Competitive Alternative

A 3-way within-subjects ANOVA was conducted with initial practice, question type, and practice test setting as the factors and net retrieval of the incorrect competitive alternative (i.e., outputting the closely related concept rather than the target information) as the dependent variable. Again, net retrieval of the incorrect competitive alternative was calculated using the formula: *net retrieval of the incorrect competitive alternative = proportion retrieval of the incorrect competitive alternative (practiced questions) – proportion retrieval of the incorrect competitive alternative (non-practiced control questions)*. Positive values indicate greater rates of negative suggestion after practice than after no practice and negative values indicating lower rates of negative suggestion after practice than after no practice. The 3-way interaction was significant, F (1, 507) = 29.76, p < .001, η_p^2 = .055. We therefore examined the results of the individual and collaborative conditions separately.

Individual. For the items that students practiced individually, the 2-way question type x initial practice interaction was significant, F (1, 507) = 75.93, p < .001, η_p^2 = .13, so follow-up paired-samples t-tests were conducted to analyze the simple effects. For true initial practice,

rates of net retrieval of the incorrect competitive alternative were near-zero for previously tested questions (M = -.001, SD = .28) but positive for previously related questions (M = .13, SD = .38), t (507) = -6.30, p < .001, d = -0.28, 95% CI [-0.37, -0.19]. As we will discuss more in-depth in the discussion of this experiment, it is possible that students' prior misconceptions interfered during the final cued-recall test (i.e., this is evidence of *learner*-driven negative suggestion).

False initial practice led to the opposite pattern of results: Net retrieval of the incorrect competitive alternative was near-zero for previously *related* questions (M = .00, SD = .27) and positive for previously tested questions (M = .12, SD = .39), t (507) = 6.15, p < .001, d = 0.27, 95% CI [0.18, 0.36]. This result is potentially evidence of test-driven negative suggestion, as students outputted the incorrect pairing of concepts suggested by the false practice test items.

Collaborative. Overall, results suggest that collaborative practice testing led to similar, near-zero rates of net retrieval of the incorrect competitive alternative regardless of initial practice or question type. For the items that were previously practice tested collaboratively, the 2-way question type x initial practice interaction was nonsignificant, F (1, 507) = 1.02, p = .31, η_p^2 = .002, as was the main effect of question type, F (1, 507) = 0.57, p = .45, η_p^2 = .001, and the main effect of initial practice, F (1, 507) = 0.59, p = .45, η_p^2 = .001.

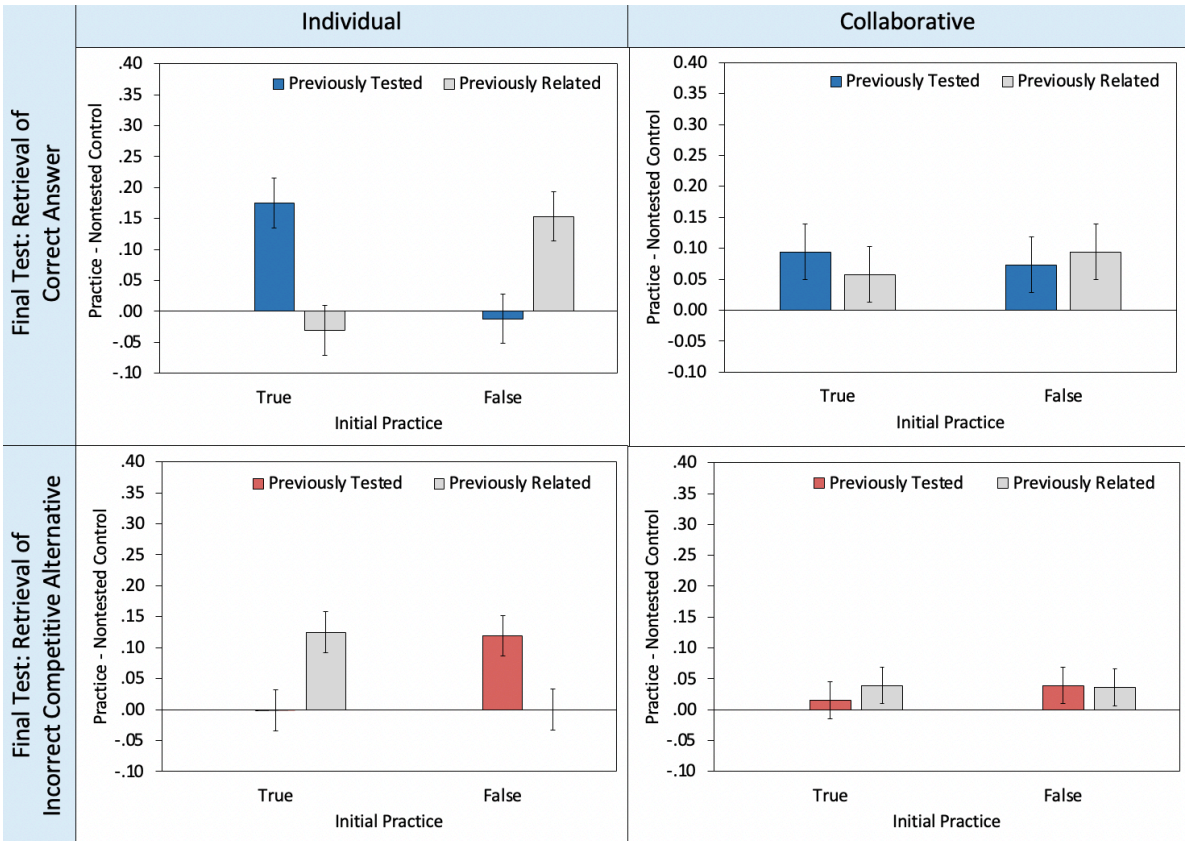


Figure 6. Net effects of practice testing on the final cued-recall test in Experiment 3. Error bars represent 95% confidence intervals.

The results of Experiment 3 suggest that practice testing has a meaningful impact on memory for course content in an authentic learning context. Extending the results of Experiment 1, students demonstrated evidence of the one-and-done effect for topics which they practice tested on individually, even though—unlike in Experiment 1—students received feedback at the end of the individual practice test. Also like in Experiment 1, there was substantial evidence of test-driven negative suggestion following false practice for previously tested questions, suggesting that students tended to output the incorrect suggested pairing following false practice.

Unlike Experiment 1, however, there was also evidence of learner-driven negative suggestion following *true* practice for previously related questions. We suggest that this result is

due to students' (incorrect) prior understanding of course content. The pairs of concepts used in this experiment were selected because students in previous iterations of the course often confused members of the pair with one another on course assessments (e.g., selecting fMRI rather than EEG on a midterm exam). Therefore, it is understandable that students sometimes produced the incorrect member of the pair during on the cued-recall test, even if they never saw that incorrect pairing during practice testing.

That explanation, however, does not fully account for the fact that what we see here is positive *net* rates of learner-driven negative suggestion following practice testing as compared to following no practice testing. If retrieval of the incorrect competitive alternative following true practice was purely driven by students' incorrect prior understandings of course material, then we would expect that students would do so at similar rates regardless of whether the item had appeared on the practice test or not (i.e., yielding zero net learner-driven negative suggestion).

So, we ask why would students output the incorrect competitive alternative on the final test for previously related questions following true practice? We believe the most likely explanation is that the evaluation of the true practice item (e.g., *True or False? The best methodology for Danielle's study is fMRI*) made the targeted concept (*fMRI*) highly available in memory (which was reinforced at the end of the practice test when they viewed that term again as part of their feedback). Thus, on the final test when students were presented with a question about research methodologies, fMRI had particularly high retrieval strength and was therefore more likely to be (incorrectly) produced instead of the correct answer (*EEG*). This suggestion is in line with prior research that people rely on what they have recently read, even when they hold accurate prior knowledge (Rapp, 2016; Rapp & Salovich, 2018 offer additional discussion).

As in Experiment 1, collaborative practice testing yielded substantially different patterns of learning than individual practice testing. Replicating the result obtained in Experiment 1, following collaborative practice testing there was a net benefit on the final test for both question types following true initial practice. Unlike Experiment 1, however, there was also a net benefit to final test performance following false initial practice.

Similarly, collaborative practice testing again led to different patterns of evidence of learner- and test-driven negative suggestion than individual practice testing. Overall, rates of negative suggestion were low and generally not different from zero across all levels of initial practice and type of question following collaborative practice testing. Notably, test-driven net negative suggestion was not significantly different from zero following false initial practice, a substantially different result from that obtained in Experiment 1.

It is possible that students had more robust prior knowledge in Experiment 3 than in Experiment 1, and that this prior knowledge may have reduced the frequency of source monitoring errors, as it may have been less effortful for students to distinguish between conceptually related content during or after group discussion. Further, working freely in a familiar lab group may have given participants greater opportunity to or comfort in sharing that existing knowledge with one another. This increased opportunity to share knowledge may have also led to more elaborative explanations of answers, potentially supporting learning and correction of misconceptions. This explanation is supported by practice test performance data: Unlike in Experiments 1 and 2, students did significantly better when they worked in groups than when they worked alone (full analyses are available in supplementary materials).

Net test-driven negative suggestion may have also been reduced by offering feedback following the collaborative practice test. Although this feedback was also offered following

individual practice testing, students may have been more motivated to attend to and learn from this feedback when working in groups than when working alone.

Experiment 4

In Experiment 3, some key patterns of results from Experiment 1 were again obtained, whereas other new patterns emerged. Experiment 4 sought to extend the results of Experiment 2 by implementing practice testing with competitive-clause true-false items in the same authentic learning context as in Experiment 3. Overall, we expected the broad benefits of competitive-clause true-false practice testing to replicate in Experiment 4.

Methods

Participants

As in Experiment 3, students (n = 473) across two new sections of a large Introductory Psychology course participated in this study as part of in-class activity. Data from an additional nine students who participated in the in-class activity were excluded because they were missing data from one or more of the tests.

Design, Materials, and Procedure

The only change from Experiment 3 to Experiment 4 was that test items in Experiment 4 incorporated competitive clauses within the true-false items (Table 3).

Results and Discussion

Final Test Performance

To assess the impact of collaborative versus individual practice testing and true versus false initial practice on previously tested and previously related final cued-recall questions, we conducted an ANOVA with practice test setting, initial practice, and question type as factors and net effect of practice as the outcome (Figure 7). The descriptive statistics presented in Table 2

suggest substantial positive net effects of practice on learning for all combinations of these factors. The question type x initial practice x practice test setting 3-way interaction was nonsignificant, $F\ (1, 472) = 0.26, p = .61, \eta_p^2 = .001$, so 2-way interactions were examined. The practice test setting x question type and practice test setting x initial practice interactions were also nonsignificant, (all p's > .79), but there was a significant question type x initial practice interaction, $F\ (1, 472) = 5.81, p = .016, \eta_p^2 = .012$. This interaction suggests that the net effect of practice on previously tested versus previously related cued-recall final test questions varied depending on whether the practice tested item was true or false.

Given this interaction, the net effect of initial practice on previously tested and previously related cued-recall test questions was examined for true initial practice and false initial practice separately using paired samples t-tests.

For the subset of cued-recall test items that followed practice testing with *true* items, the net effect of practice was greater for previously tested items ($M = .16, SD = .38$) than for previously related items ($M = .10, SD = .40$), $t\ (472) = 3.38, p < .001, d = 0.16$, 95% CI [0.07, 0.25]. For the subset of cued-recall test items that followed practice testing with *false* items, the net effect of practice was similar across previously tested items ($M = .12, SD = .39$) and previously related items ($M = .12, SD = .40$), $t\ (472) = 0.03, p = .98, d = 0.001$, 95% CI [-0.09, 0.09]. Together, these results suggest that true initial practice benefitted the learning of previously tested content more so than previously related content, but that false initial practice facilitated learning of both types of content similarly.

Retrieval of the Incorrect Competitive Alternative

Mirroring the pattern of final test performance, the question type x initial practice x practice test setting 3-way interaction was nonsignificant, as were the setting x initial practice

and question type x initial practice 2-way interactions (all p's > .49). The 2-way question type x initial practice interaction was significant, F (1, 472) = 5.34, p = .021, η_p^2 = .011, and was therefore followed up by paired samples t-tests.

For the subset of cued-recall test items that followed true initial practice, net retrieval of the incorrect competitive alternative was near-zero for previously tested items (M = 0.002, SD = .22) and slightly positive for previously related items (M = .03, SD = .23), t (472) = -2.71, p = .007, d = -0.13, 95% CI [-0.22, -0.03]. For the subset of cued-recall test items that followed false initial practice, net retrieval of the incorrect competitive alternative was similar across previously tested items (M = .02, SD = .25) and previously related items (M = .01, SD = .23), t (472) = 0.42, p = .68, d = 0.02, 95% CI [-0.07, 0.11].

As in prior experiments, this pattern suggests that decreases in final test performance tended to be associated with increases in retrieval of the incorrect competitive alternative. Overall, however, there was little evidence of net learner- or test-driven negative suggestion in Experiment 4.

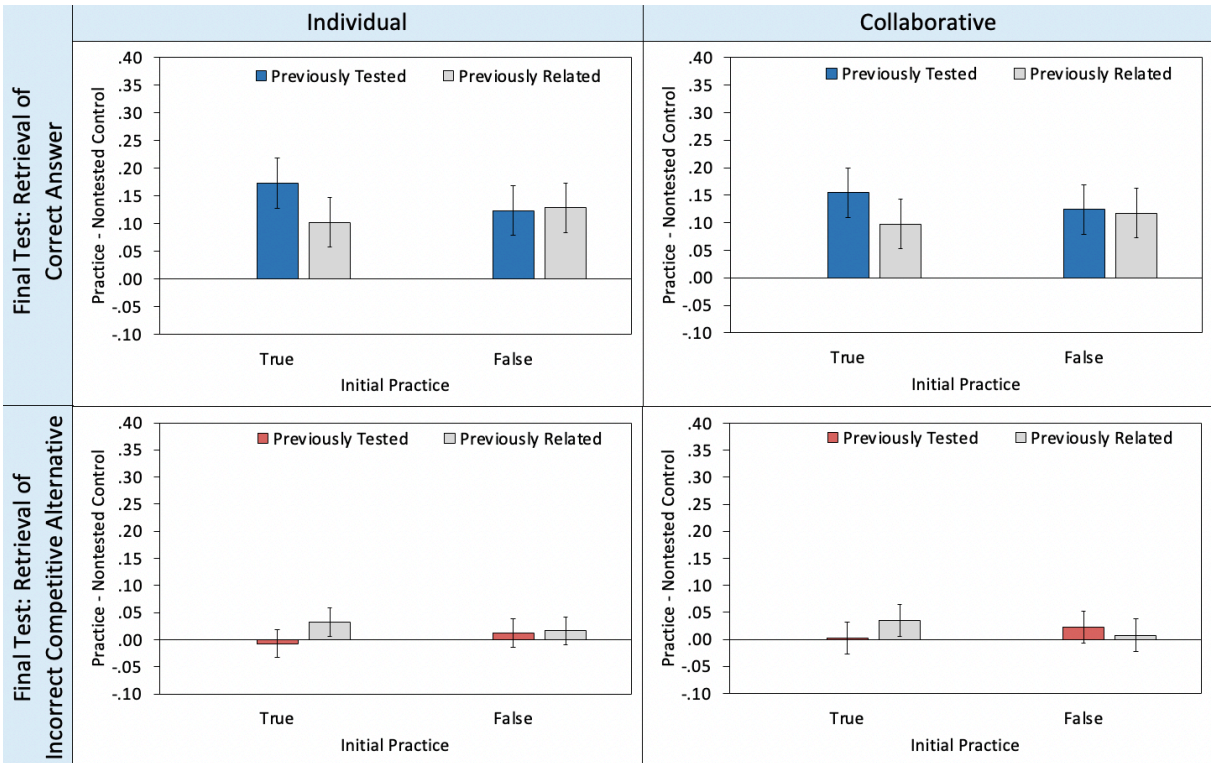


Figure 7. Net effects of practice testing on final cued-recall test in Experiment 4. Error bars represent 95% confidence intervals.

As anticipated, the results of Experiment 4 indicated that engaging in competitive true-false practice testing broadly benefits later memory for course content, regardless of whether that practice test was completed collaboratively or individually. Unlike in Experiment 2, the memory benefit of true initial practice was stronger for previously tested than for previously related questions, for both individual and collaborative practice testing. It is possible that, because students had stronger prior knowledge of the content in Experiment 4 as compared to Experiment 2, students more easily confirmed the accuracy of "true" statements and therefore less frequently carefully considered the information contained within the competitive clause. In contrast, for "false" statements, students may still have been motivated to carefully consider both the target information and the information contained within the competitive clause so that they could verify that the statement was indeed incorrect by using the competitive clause to construct the correct

pairing. It should be noted, however, that true practice did not at all benefit learning of related content in Experiment 3 (i.e., when participants practice tested using traditional true-false items), suggesting that the inclusion of competitive clauses did support retrieval of related content during true initial practice, even if that retrieval was potentially not as effortful or as consistent as retrieval of directly tested content.

Additionally, though competitive-clause true-false practice test items present both accurate and inaccurate pairings within each to-be-evaluated proposition, as in Experiment 2, net negative suggestion was near-zero for nearly every combination of initial practice, question type, and practice test format in Experiment 4. This result provides additional evidence from students in an authentic learning environment that practice testing with competitive-clause true-false statements may (a) promote memory for accurate content and (b) reduce confusion between pairs of highly related concepts.

General Discussion

In four experiments, we examined the impact on learning of practice testing alone versus practice testing with others across two forms of true-false practice testing: (1) traditional true-false and (2) competitive-clause true-false. We investigated the learning impact of these variations on implementing true-false practice testing first under highly controlled laboratory conditions (Experiments 1 and 2) and then within the authentic learning context of a large undergraduate STEM course (Experiments 3 and 4). Overall, our results suggest that true-false practice testing can promote learning, and that these patterns of learning can profoundly differ following individual versus collaborative true-false practice testing, especially when learners practice test using traditional true-false items.

Reconciling The Effects of True-False Practice Testing on Learning in the Laboratory Versus in The Classroom

Experiment 1 (traditional true-false test items) and Experiment 2 (competitive-clause true-false items) compared the effect of individual and collaborative true-false practice testing on learning in a carefully controlled online laboratory setting. In these experiments, participants briefly studied passage content, took a true-false practice test either alone or in a small group (without feedback), completed a brief distractor task, and then took an individual final cued-recall test. Under these conditions, participants had limited ability to create robust prior knowledge on the to-be-learned content, spent only 24 s on each practice test item, did not receive corrective feedback after the practice test, and worked with individuals that they were unlikely to be familiar with (although we did not specifically ask participants if they knew one another, groups were formed based on who signed up for a particular time slot and it was highly unlikely given the large participant pool that participants would have had prior familiarity with one another).

In contrast, Experiment 3 (traditional true-false test items) and Experiment 4 (competitive-clause true-false items) compared the effect of individual and collaborative true-false practice testing on learning within a large in-person undergraduate STEM course. In these experiments, there was no initial study phase, as participants received instruction on the to-be-learned content in prior course meetings. Instead, students worked with a familiar group of students (their laboratory groups who they had collaborated with previously in the course) to take a collaborative practice test (with feedback). Then students took an individual practice test (with feedback), completed a brief distractor task, and then took an individual final cued-recall test. Under these conditions, participants possessed a more-established body of prior knowledge on

the to-be-learned content, worked with familiar others, were allowed to spend as long as they would like on each practice test item, and received corrective feedback after each practice test. Despite these substantial differences in study procedures, several results were obtained in both learning contexts.

The Effect of Individual Versus Collaborative Traditional True-False Practice Testing on Learning

First, participants in both learning contexts demonstrated evidence of the one-and-done effect (Brabec et al., 2021) following individual practice testing with traditional true-false items. Here, true initial practice (i.e., evaluating a true proposition) yielded a net benefit to learning (as compared to no practice testing) of previously tested content but not previously related content. Oppositely, false initial practice (i.e., evaluating a false proposition) yielded a net benefit to learning of previously related content but not previously tested content. This pattern suggests that learners only recalled as much information as was necessary to ascertain the validity of the true-false proposition: When the proposition was true, learners retrieved information about the target concept, but when the proposition was false, learners retrieved information about the incorrect competitive alternative.

Notably, receiving correct-answer feedback (i.e., being told whether a statement was "true" or "false") in Experiment 3 did not ameliorate the one-and-done effect, which suggests that the learning is facilitated by retrieval of that content during practice testing and not merely knowing the validity of the proposition. Whether feedback that includes additional content, such as feedback which offers the true version of the false proposition, might reduce participants' tendency to exhibit the one-and-done effect is a question for further research; to our knowledge, the only work which has systematically offered feedback following true-false practice testing

provided correct-answer feedback and did not explicitly compare providing feedback to not providing feedback (Uner et al., 2021).

In contrast to the pattern obtained following individual practice testing, learners did not demonstrate evidence of the one-and-done effect following traditional true-false practice testing in either a controlled laboratory context or an authentic learning setting. In a laboratory context, true initial practice seemed to yield learning of both previously tested and previously related content, but false initial practice did not yield any net benefit to learning relative to no practice testing for either question type. In an authentic learning context, however, learners benefitted broadly from both true and false initial practice.

Why might false initial practice have led to such differential impacts on learning across these two contexts? That increases in rates of retrieving the incorrect competitive alternative roughly correspond to decreases in final test performance (based on visually inspecting differences in final cued-recall test performance following true initial practice versus false initial practice) may offer some insight. This pattern suggests that it is not that false practice led to *no* learning, but rather that it led to learning of incorrect pairings of information. The learning of inaccuracies following false practice was especially prevalent in the laboratory context.

It is possible that group discussion may have facilitated broader retrieval of content during practice testing. Doing so was positive during true initial practice when learners were only exposed to correct information. When the test suggested incorrect pairings of concepts in the form of a false proposition, however, this increase in recalled information may have negatively impacted learners' ability to track which pairings of content were true and which pairings of content were false. If learners retrieved content brought up during discussion while taking the final cued-recall test, but could not remember whether that content was previously

tested or previously related, then learners would produce fewer correct answers and more incorrect competitive alternatives on the final test—with this decrease and increase, respectively, roughly similar in magnitude. This suggested result matches our obtained result in Experiment 1. It also fits the pattern of results obtained in the authentic learning context in Experiment 3. If prior knowledge reduces the cognitive load of tracking incorrect versus correct pairings of content, then learners would be more likely to output the correct answer on the final cued-recall test and less likely to produce the incorrect competitive alternative. Again, this suggested result matches our obtained result in Experiment 3.

The Effect of Individual Versus Collaborative Competitive-Clause True-False Practice Testing on Learning

Overall, competitive-clause true-false practice testing benefitted learning of previously tested and previously related content when that practice testing was done individually and when it was done collaboratively. This result was consistent across the online laboratory and the classroom learning contexts. The consistency in the pattern of obtained results suggests that the benefits of competitive-clause true-false practice testing may be robust to variations in its implementation.

Evidence of Test-Driven and Learner-Driven Negative Suggestion Following True-False Practice Testing

Across the four experiments, learners sometimes demonstrated evidence of increased test-driven and learner-driven negative suggestion following practice testing as compared to after no practice testing. We considered retrieval of the incorrect competitive alternative on the final cued-recall test to be evidence of test-driven negative suggestion following false initial practice as, in those instances, the practice test item "suggested" an incorrect pairing of two concepts. In

contrast, we considered retrieval of the incorrect competitive alternative on the final cued-recall test to be evidence of learner-driven negative suggestion following true initial practice. In the case of learner-driven negative suggestion, the learner spontaneously produces the intrusion on the final cued-recall test, as the practice test item itself only presented correct information. In Experiment 1, learners demonstrated evidence of substantial test-driven negative suggestion for previously tested items following false initial practice. This result suggests that learners were susceptible to the misinformation presented by false practice test items (Butler, 2018; Toppino & Luipersbeck, 1993). Even when learners had more robust prior knowledge and were offered correct-answer feedback, learners still demonstrated substantial test-driven negative suggestion following individual, false initial practice (Experiment 3). This result bolsters concerns that true-false practice items can promote learning of falsehoods in an authentic learning context and that these incorrect associations may persist to later assessments.

Collaboration during practice testing with traditional true-false items tended to lead to less test-driven negative suggestion than practice testing alone. Working with others offers the possibility of exchanging knowledge with one another. Doing so may have facilitated not only the identification of false items as false when learners had adequate prior knowledge (Experiment 3, supplementary materials include the practice test results) but also the exchange of elaborative explanations. As socially justifying beliefs (Bruffee, 1984) and exchanging evidence (Clark et al., 2000) are two key processes groups engage in when collaboratively learning, group members may have gone beyond simply announcing that a proposition was true or false. Instead, they may have been motivated to explain their reasoning by retrieving information to correct the false statement or by offering additional elaboration on the topic. These behaviors may have promoted a deeper understanding as to why that proposition offered a false pairing of

content, and perhaps even created a stronger memory trace of the correct versus the incorrect pairing of concepts, which may have reduced test-driven negative suggestion on the final cued-recall test.

Additionally, although feedback can reduce evidence of negative suggestion following multiple-choice practice testing (Butler, 2018), that was not the case here. Feedback offered during multiple-choice testing typically presents the correct answer such that learners are told that their response was either correct or incorrect and can restudy the correct information if needed. Here, feedback simply told participants whether true-false proposition was true or whether it was false. Possibly, simply providing correct-answer feedback following true-false practice testing does not sufficiently support learners in resisting misinformation presented by false propositions.

Evidence from research on the continued-influence effect (i.e., the phenomenon that false information can continue to influence understandings even after being corrected) suggests that refutations of false information are strongest when the correct information is provided along with the refutation (Ecker et al., 2010; Johnson & Seifert, 1994). Exposure (or re-exposure, as is the case here) to correct information offers the possibility of updating one's knowledge beyond simply tagging something as false by actually replacing the incorrect information with the correct information in memory. As associations between content and a "false" label can fade with time, the refutation may be more durable if learners can reconstruct their memory with new, accurate understandings. This suggestion is supported by the tendency for there to be lower rates of negative suggestion following practice testing with false competitive-clause items than when practice testing with false traditional items. The competitive clauses may have offered learners the necessary information to correct the false statement and also to construct a new, correct

pairing of content, which may have resulted in a stronger refutation of the incorrect pairing offered by the false proposition.

Evidence of considerable learner-driven negative suggestion was only obtained for previously related questions on the final cued-recall test following individual practice testing in an authentic learning context. It is likely that this result stems from learners' prior misconceptions or confusions about course content. The pairs of concepts used in Experiments 3 and 4 were selected because the course instructors noticed that students often mixed them up on course assessments. When learners were presented with a true statement containing the directly tested concept (e.g., EEG) it likely increased the availability of that term in memory. Being offered a question on the final cued-recall test targeting the related term (e.g., fMRI) may have led to spreading activation which increased the retrieval strength (Bjork & Bjork, 1992) of both the target concept (fMRI) and the previously practiced content (EEG). Since students had recently seen the key term EEG on the practice test, it likely already had high retrieval strength, and thus was often erroneously recalled on the final cued-recall test. Although there was some suggestion of learner-driven negative suggestion under the laboratory conditions of Experiment 1, it was much less prevalent, suggesting that perhaps when learners' prior knowledge is less organized or key terms in a schema are less strongly connected to one another, participants are more likely to experience simple retrieval failure (i.e., recalling nothing or a completely unrelated key term) rather than retrieve the incorrect competitive alternative.

Limitations and Future Directions

Although the current studies offer new insights into the effects of collaborative versus individual and traditional versus competitive-clause true-false practice testing on learning, there are some limitations to this work. First, an immediate test rather than a delayed test was

employed to assess learning. Measuring learning at a delay could offer unique insights as to the impact of the experimental manipulations on retention of practice tested content. Second, the materials in these studies were selected such that pairs of concepts were highly related and easily confused by learners. A question for future research is whether evidence of negative suggestion would be as common or the benefit of competitive clauses for learning be as powerful if students practiced tested on content that was not as highly confusable. Third, although prior work informs our speculation as to the cognitions and behaviors that participants in these studies engaged in while practice testing, attempting to observe these processes via either videorecording (Marquez et al., 2023) or think-aloud procedures (Little et al., 2019; Uner et al., 2021) could offer a window into how participants actually engage with various forms of true-false practice tests. Finally, there were several changes between the procedures and learning contexts of Experiments 1 and 2 and Experiments 3 and 4. Consequently, it is challenging to identify the exact change(s) that may have contributed to the differences in the patterns of results. Additional investigation which disentangles the effects of working with familiar versus unfamiliar others, or being offered or not offered feedback, could inform recommendations for the implementation of true-false practice testing.

Concluding Comments

The current work indicates that collaborative true-false practice testing can indeed yield different patterns of learning than individual true-false practice testing. Further, investigation of the effect of incorporating competitive clauses into true-false practice items obtained substantial evidence that doing so can elicit broad benefits to learning and even reduce the learning of incorrect associations from false propositions presented during practice testing. These patterns suggest that practice testing with competitive-clause true-false items, especially when done so

collaboratively, can powerfully enhance learning. The next chapter investigates whether collaboration continues to potentiate the benefits of practice testing when learners are given less structure during the practice testing activity and use a format commonly employed by students during self-regulated learning: flashcards.

CHAPTER 4

When Two Learners Are Better Than One:

Using Flashcards with a Partner Improves Metacognitive Accuracy

Abstract

We investigated the benefits of two ways to use flashcards to perform retrieval practice: alone versus with a partner. In two experiments, undergraduate students learned word-definition pairs using flashcards alone (Individual condition) or with another student (Paired condition), made judgments of learning (JOLs), and then completed a final cued-recall test after a 5-min delay (Experiments 1-2) and a 24-hour delay (Experiment 2). In Experiment 1, students in the Paired condition dropped flashcards less often than in the Individual condition (dropping was prohibited entirely in Experiment 2). In addition, although final test performance was similar across conditions in both experiments, inaccurate JOLs 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. Thus, 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. Overall, these findings have implications for self-regulated learning and effective exam preparation.

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*; Pan & Rickard, 2018; Roediger & Butler, 2011; Rowland, 2014 offer comprehensive reviews). Indeed, a recent 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; Zung et al., 2022). 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 studying behaviors (e.g., time per item, decisions 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 as 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; Pan et al., 2022; Senzaki et al., 2017; Zung et al., 2022 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 (Pan et al., 2022). 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 (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 own 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 three reasons why using flashcards in pairs may be beneficial. First, learners cannot engage in covert retrieval. The need for overt responses, which can be more effective than covert responses (e.g., Kubik et al., 2020; Tauber et al., 2018), may prevent learners from "cheating themselves" by not fully articulating a response to a given question or cue. Retrieval attempts can be more potent and more informative for metacognitive judgments as a result. A partner might even offer explanations and correction of errors, further facilitating learning (Johnson et al., 1998; LoGuidice et al., 2015). Second, the presence of others may affect learners' emotional states positively, such as by increasing 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. 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). All of these reasons suggest that using flashcards with a partner—which has yet to be extensively investigated—may be beneficial.

The Present Studies

We investigated the hypothesis that flashcard-based retrieval practice is better for learning and metacognitive accuracy when it is implemented with a partner as opposed to individually. We also examined potential differences between individual and paired flashcard learning in terms of the mechanics of flashcard use (e.g., cycles through flashcard stacks), associated study decisions (e.g., dropping cards), and affective states. Across two 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 Experiment 2 it was prohibited. Additionally, while in Experiment 1 both Individual and Paired learners engaged in cycles of study and retrieval practice, in Experiment 2, all learners engaged in an initial study period such that Individual learners then only engaged in retrieval practice, which we believe to be more aligned with students' own behaviors when using flashcards in daily life (Zung et al., 2022). Importantly, across conditions, we controlled for total time, used the same flashcards and learning environments, and gave similar instructions.

Experiment 1

The first experiment addressed a scenario wherein learners have 20 minutes each to study a set of vocabulary words and perform retrieval practice on those vocabulary words. They can do so by themselves or with a partner. In the case of Individual learners, such learning involves 20 minutes of studying followed by 20 minutes of practice. For Paired learners, the logistics are somewhat more complex: One partner must serve as the "tester" and the other partner as the "testee" before the roles are switched. Hence, in the Paired condition, one partner engages in 20 minutes of practice testing from the outset, whereas the other partner does so after those 20 minutes have 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 a 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-participants factorial design with factors of Condition (Individual vs. Paired) and First Learning Activity (Study First vs. 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 Appendix C.

Each word-definition pair was printed on a 4 x 6 in. white index flashcard. For the *standard* flashcard set, which was designed for studying and testing, 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.

Procedure

The experiment was run in 2-hr timeslots involving up to four participants each and using three nearly-identical laboratory testing rooms. 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 learning phase involving flashcards. Then they completed a series of survey questions—which included providing a judgment of learning (JOL)—, 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.

Learning 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. Verbal feedback was disallowed to minimize off-task conversations. 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 learning 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 (JOL; 0-100% likelihood of remembering the words on an upcoming test), (d) answer a mind-wandering probe (reported focus during the learning phase, from 0-100%), and (e) answer questions regarding their activities during the learning phase and their own flashcard use in everyday study sessions. 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 Pan & Rickard, 2017). The experiment concluded afterwards.

Results

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 reported in Appendix D, 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.

Learning Phase

Number of Learning Cycles. Participants indicated the number of learning cycles (i.e., practicing through the entire flashcard stack) they completed per 20-min period. These responses were summed for a total number of cycles in the entire learning phase; if participants indicated an incomplete cycle, then 0.50 was added (this method, albeit somewhat imprecise, was applied across conditions for consistency). Individual learners typically completed one more learning cycle ($M = 5.36$, $SD = 1.64$) across the entire learning 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 kappa = .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%), $\chi^2\,(2) = 44.43$, $p < .001$. 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 on the basis of 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 (supplementary materials include these analysis). 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], which indicates that recall of the GRE words was no different at a short delay after individual or paired flashcard learning (Table 4 presents the descriptive statistics for each condition).

Table 4

Cued-Recall Test Performance in Experiments 1 and 2

Condition	Experiment 1[7]		Experiment 2			
			Immediate Test		Delayed Test	
	M	*SD*	*M*	*SD*	*M*	*SD*
Individual	.48	.24	.49	.28	.40	.28
Paired	.43	.23	.44	.25	.35	.24

[7] 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, a series of bivariate correlations related JOL and final test performance for both conditions (Figure 8). Individual learners demonstrated moderate-to-large correlations between their JOL and final test performance when learning individually, r (61) = .59, $p < .001$, as did Paired learners, r (86) = .60, $p < .001$. Although the magnitude of the relationships between JOL and test performance was similar between the Paired and Individual conditions, Figure 8 clearly shows that the intercepts of the regression lines between the two conditions (computed by regressing test performance onto JOL data) differ, prompting further analyses of participants' metacognitive calibration.

Metacognitive Calibration. We computed metacognitive calibration by subtracting participants' actual test performance from their JOLs, 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, JOL-test performance and metacognitive calibration scores provide complementary, but distinct, information about learners' metacognitive judgments.

An independent samples t-test compared Individual and Paired learners' metacognitive calibration scores (Figure 8). 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].

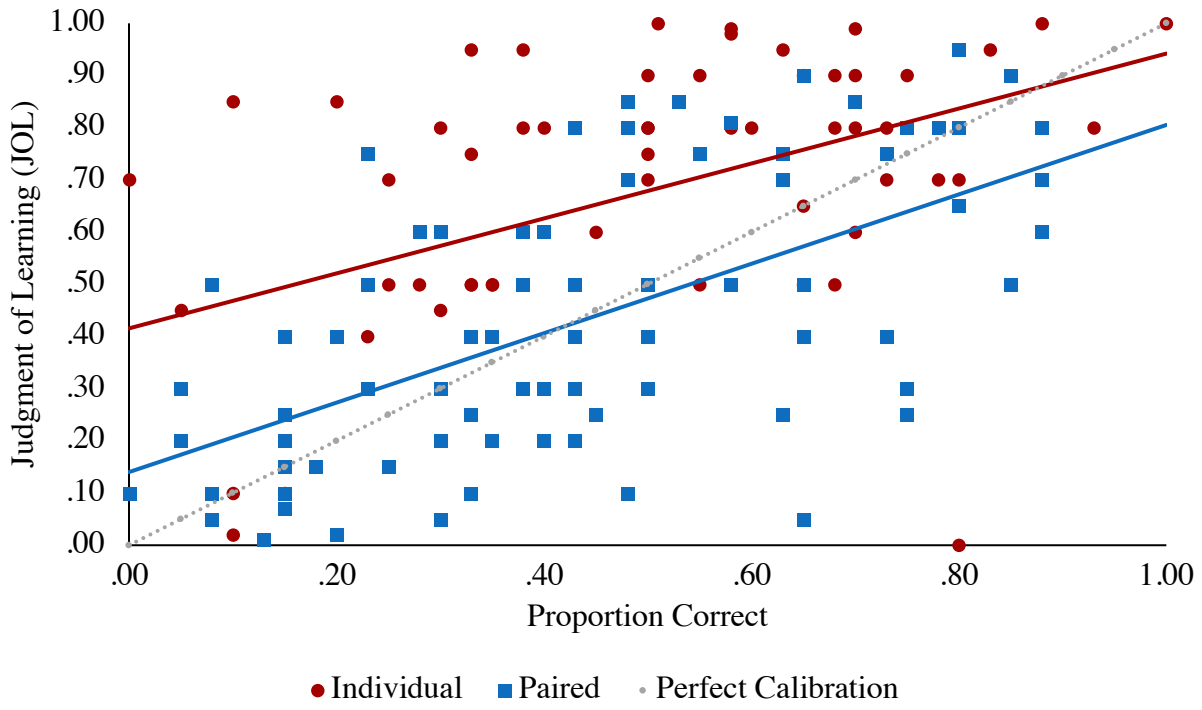


Figure 8. Metacognitive calibration demonstrated by those in the Individual flashcard learning and the Paired flashcard learning conditions. Each panel displays the correlation between final test performance and global judgments of learning (JOLs). A dotted line represents the hypothetical case of perfect calibration between JOLs and test scores; crucially, participants in the Individual condition tended to substantially overestimate their final test performance.

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]. However, those that studied in pairs reported significantly higher negative affect (M = 16.93, SD = 6.54) than those that 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 focus during the experimental tasks did not significantly differ between the Individual (M = 78.44, SD = 16.48) and Paired (M = 80.75, SD = 19.14) conditions, t (150) = -0.78, p = .44, d = 0.13, 95% CI [-8.18, 3.55].

Experiment 1 Discussion

Contrary to our predictions, the results of Experiment 1 suggest that collaborative and individual practice of difficult vocabulary terms using flashcards yield comparable test performance after a 5-min delay. It is 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).

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. 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 learning 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 for other reasons (perhaps because their partner was holding the flashcard deck, adding friction to the drop decision), 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.

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 have been often 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; Imundo et al., submitted). 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 continued to compare 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. To rule out if paired learners were 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 another in the Paired condition, a brief icebreaker 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 learning phase or on 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 learning phase—here, renamed the *practice 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), and for reporting that they dropped flashcards from study during the practice phase ($n = 15$).

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.

Procedure

Aside from the following changes listed below, the procedure of Experiment 2 was the same as Experiment 1.

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 seconds, for an overall study time of approximately 10 minutes.

Practice Phase. Given that learners received approximately 10 minutes 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) and renamed from the *learning* phase to the *practice* phase. During the practice phase, all participants solely used the standard flashcard set.

Individual condition. Participants were instructed to test themselves during the entirety of the practice phase. They were told that the experimenter would check in on them after 15 minutes. 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 practice phase, Paired learners were given two minutes to complete an icebreaker activity. During this icebreaker, 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 practice phase, participants were told that they could provide brief verbal feedback or 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 cards from study was removed.

Final Cued-Recall Test

Immediate (5-min). Twenty definitions were presented.

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 reported a JOL.

Results

Number of Practice Cycles

Participants indicated the number of practice cycles (i.e., practicing through the entire flashcard stack) they completed per 15-min period of the practice 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, an ANOVA with Condition (Individual or Paired) as the between-subjects factor, Test Delay (5-min or 24-hrs) as the within-subjects factor, and test performance as the dependent variable was conducted. Six participants did not have a delayed test[8] 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 (1, 133) = 1.44, $p = .23$, $\eta_p^2 = .01$[9]. The nonsignificant Condition x Test Delay interaction suggests this similarity did not change between the immediate test and the delayed test, F (1, 133) = 0.003, $p = .96$, $\eta_p^2 < .001$.

Metacognitive Judgments

Of the 141 participants in the final sample, 84 (59.6%) offered a delayed JOL. 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 JOL (in accordance with our IRB protocol, participants were not required to answer any

[8]Additionally, seven 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 does not change the pattern of results.
[9]An independent samples t-test examining the effect of Condition at immediate test only (n = 141) obtained the same result.

question)[10]. 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 JOL ahead of the delayed cued-recall test did not appear and thus they had no opportunity to report a JOL.

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 JOL and final test performance for both conditions and for both test timings (presented in Figure 9).

As in Experiment 1, for the immediate test, participants demonstrated moderate-to-large correlations between their JOL 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 JOLs and performance on the delayed test, Individual: r (51) = .43, p = .001, Paired: r (29) = .36, p = .047.

Metacognitive Calibration. In order 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

[10] It is not clear why so many participants chose not to report a delayed JOL. 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 JOL and were unsure how to respond.

(*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].

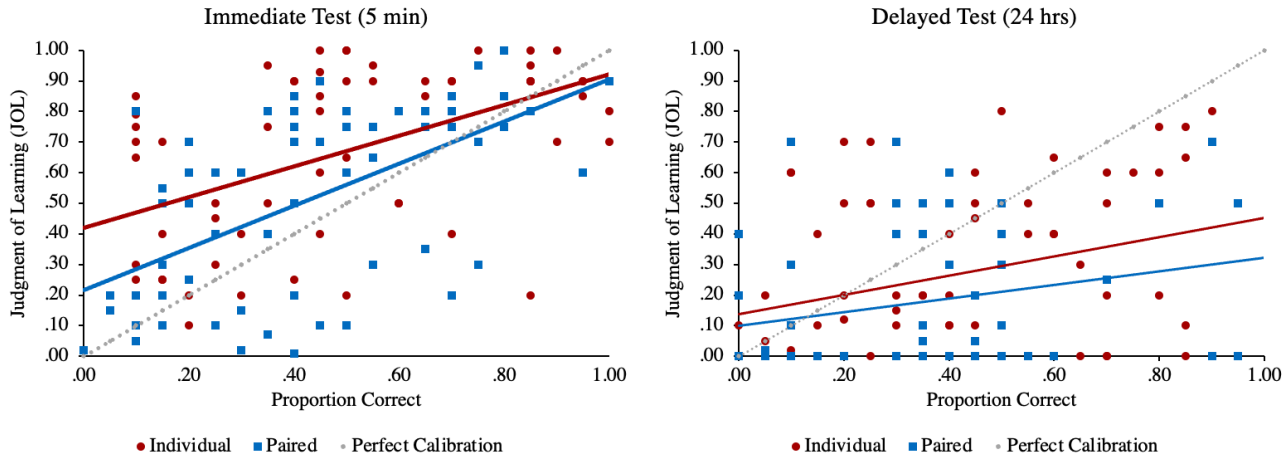


Figure 9. Metacognitive calibration for the Immediate test (left panel) and the Delayed test (right panel). Each panel displays the correlation between test performance and global judgments of learning (JOLs). The red and blue lines represent least squares regression fits to Individual and Paired data, respectively. A dotted line represents the hypothetical case of perfect calibration between JOLs 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

Overall, participants self-reported far more positive (*M* = 27.30, *SD* = 8.05) than negative affect (*M* = 15.09, *SD* = 4.23), *t* (140) = 17.29, *p* < .001, *d* = 1.46, 95% CI [10.81, 13.60]. 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]. In contrast to Experiment 1, however, 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 ice breaker 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 focus during the experimental tasks did not significantly differ between the Individual (*M* = 86.33, *SD* = 14.55) and Paired (*M* = 87.76, *SD* = 12.83) learning conditions, *t* (139) = -0.61, *p* = .54, *d* = -0.10, 95% CI [-6.05, 3.20].

Self-Reported Flashcard Use in Experiments 1 and 2

Table 5 summarizes data on participants' self-reported use of flashcards for exam preparation. In both experiments, most students reported using flashcards at least sometimes when studying, with roughly 1 in 5 using flashcards frequently when preparing for an exam. 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.

Table 5

Frequency of Self-Reported Flashcard Use When Preparing for Exams

Frequency	When studying generally				When studying with a partner			
	Exp. 1		Exp. 2		Exp. 1		Exp. 2	
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Never	19	12.5	18	12.8	25	16.4	36	35.5
Almost never	37	24.3	41	29.1	54	35.5	52	36.9

Sometimes	65	42.8	72	51.1	59	38.8	50	35.5
Almost every time	26	17.1	7	5.0	9	5.9	3	2.1
Every time	5	3.3	3	2.1	5	3.3	52	0.0
Total	152	100.0	141	100.0	152	100.0	141	100.0

General Discussion

Across two experiments, using flashcards to learn with a partner did not yield greater learning compared to using flashcards alone. Despite our expectation that collaboration might serve as a "desirable difficulty" and better promote long-term learning in Experiment 2, Individual and Paired flashcard use was equally beneficial for learners when learning was assessed at both a brief (5-min) and long (24-hr) delay. Although performance did not differ 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: Whereas Individual learners were often overconfident—overestimating learning by approximately 20% in both experiments—Paired learners were more accurate. These benefits of paired flashcard practice might be particularly important for those using flashcards during self-regulated learning.

Why is Paired Flashcard Learning Advantageous for Metacognition?

Our findings appear to stem from characteristics of using flashcards with a partner (which, in the present studies, resembled how two learners might use flashcards): (a) overt responses were required, (b) feedback occurred only after a complete retrieval attempt, and (c) feedback was consistently provided. Unlike their counterparts in the Individual condition, Paired

learners had to clearly articulate a response before feedback was provided, possibly resulting in more effortful retrieval processes (Pyc & Rawson, 2009 offers a discussion about the benefits of effortful retrieval) which were not shortchanged by any "cheating" and peeking at the answers. Further, although Paired learners had to wait for partner-provided feedback, its consistent occurrence obviated any issues with insufficient checking of answers (Wissman et al., 2016). Inconsistent 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. Such dropping commonly occurred because a given word-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 word-definition pairs. Consequently, Individual learners based their JOL on impoverished information relative to Paired learners, whom could rely on more consistent item-level evidence.

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. Another possibility is that the experience of taking the immediate test in Session 1 of Experiment 2 offered participants insight into their learning ahead of predicting their delayed test performance, and that this information was particularly useful for Individual learners.

Affective Considerations

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 (although students sometimes opt to work with strangers, including in large classes, in assigned groups, or with "friends of friends"). This explanation is supported by the lack of evidence for elevated negative affect in Paired learners in Experiment 2, which incorporated a brief icebreaker to facilitate participants getting to know each other (if only superficially) and eased restrictions on verbal communication during the flashcard practice phase. Although students would likely work with those they know if engaging in paired flashcard learning in everyday life (although our survey results suggest that students may be unlikely to do so, in line with Zung et al., 2022), 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 asked 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. As a key driver of collaborative benefits is the exchange of knowledge through explanations and elaborations, it further possible that the use of less-complex materials (word-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. To address some of these possibilities, future work might employ experimenter observation, 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

Our finding that it is advantageous to use flashcards in pairs has important practical implications for self-regulated learning and effective exam preparation. This work also suggests solutions for some common pitfalls of flashcard-based retrieval practice. Further, 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.

CHAPTER 5

Summary and Discussion

Considerable research has attested to the fact that practice testing can enhance learning of content more powerfully than many other learning strategies (Dunlosky et al., 2013). Possible variations in form and implementation of practice testing offer learners numerous possibilities in how to engage in practice testing. Consequently, understanding how these variations in practice testing may facilitate different patterns of retrieval, and subsequent patterns of learning, offers considerable practical and educational utility. In an increasingly interconnected world in which collaboration is considered a key skill (Mashek, 2022), collaborative practice testing may elicit processes which could benefit learning and learning-relevant outcomes (e.g., metacognition). The aim of this dissertation was to investigate whether collaborative practice testing would yield different patterns of learning as compared to individual practice testing. We focused on structured practice test formats (e.g., multiple-choice) as they tend to more clearly guide learners' retrieval than unstructured test formats (e.g., free-recall) and therefore offer the possibility to more clearly interpret differential patterns in learning as evidence of differential patterns of retrieval during the practice testing event. Additionally, we explored whether such similarities or differences might be different across several types of practice test formats. Finally, we assessed whether individual versus collaborative practice testing might impact learning-relevant outcomes, such as attitudes towards group work (Chapter 2) and evaluation of one's learning (Chapter 4).

Collaborative Practice Testing Yields More Durable Learning Than Individual Practice Testing

Overall, our results revealed that individual and collaborative practice testing can in fact yield different patterns of learning or learning-relevant outcomes under certain conditions. In Chapter 2, collaborative multiple-choice practice testing with feedback fostered more durable learning than individual multiple-choice practice testing with feedback when learning was assessed on a surprise retention test one week and two weeks later (Experiment 1), but not when retention was assessed on the open-book course final exam six weeks later (Experiment 2). Interpretation of the differences in performance on the course final exam following individual and collaborative practice testing, however, is challenging because students engaged in considerable outside studying for the exam and students were allowed to look up answers. These opportunities resulted in very high test performance on the retention items and on the exam more generally. Together, these findings offer some evidence that collaborative multiple-choice practice testing can foster more durable learning than individual multiple-choice practice testing.

Collaborative Practice Testing With Traditional True-False Items Facilitates Broader Learning of Previously Tested and Previously Related Content Than Individual Practice Testing

Chapter 2 provided evidence that collaborative multiple-choice practice testing can facilitate more durable learning of directly tested content than individual multiple-choice practice testing. Chapter 3 extended the investigation of these learning differences by exploring if practice testing with others might facilitate differential learning of both directly tested and conceptually related content than practice testing alone. As considerable research has already established that individual practice testing with multiple-choice items can enhance learning of

previously tested and previously related content (Little et al., 2012; Little et al., 2019), the series of studies in Chapter 3 employed two forms of the true-false test format: traditional and competitive-clause. While traditional true-false items simply offer a proposition that is true or false (e.g., Eris is a dwarf planet located in the asteroid belt), competitive-clause true-false items modify that proposition to contrast target and related content in a this-not-that format; e.g., Eris (not Ceres) is a dwarf planet located in the asteroid belt (Brabec et al., 2021).

As did Chapter 2, Chapter 3 offers evidence that collaborative and individual practice testing can yield different patterns of learning, particularly following practice testing with traditional true-false items. Both under highly-controlled online laboratory conditions (Experiment 1) and within a large in-person undergraduate STEM course (Experiment 3), learners who practice tested individually demonstrated evidence of the one-and-done effect (Brabec et al., 2021). True initial practice benefited learning of previously tested content (but not previously related content) whereas false initial practice benefited learning of previously related content (but not previously tested content). Together, these findings suggest that individual practice testing with traditional true-false items may offer limited learning benefits, even when students hold considerable prior knowledge of the to-be-learned content and are offered feedback after learning (Experiment 3). To our knowledge, this chapter also offers the first demonstration of the one-and-done effect in an authentic learning context.

Collaborative practice testing, in contrast, resulted in a markedly different pattern of learning than individual practice testing with true-false items. Under laboratory conditions when learners had limited study time prior to practice testing and did not receive feedback, evaluating true propositions facilitated learning of both previously tested and previously related content, whereas evaluating false propositions did not facilitate learning compared to no practice testing

for either question type. In fact, rates of retrieval of the incorrect competitive alternative (i.e., evidence of negative suggestion) on the final cued-recall test suggest that collaboration may have encouraged learning of the incorrect association proffered by the false proposition. These results suggest a double-edged sword of collaboration: Working with others to evaluate these propositions might facilitate discussion of a greater variety of to-be-learned content, but also may add cognitive load such that it is more challenging for learners to establish and maintain which pairings of concepts are correct and which pairings of concepts are incorrect. When learners had stronger prior knowledge and received feedback (Experiment 3), however, this tendency to learn false associations was reduced and learners broadly benefitted from collaborative practice testing. Taken together, these results suggest that collaboration can facilitate broader retrieval of information than working alone when practice testing with traditional true-false items, but that measures should be taken to ensure that discussion is not so taxing on learners that they struggle to form and/or maintain correct associations of content when presented with false practice test items.

Fortunately, Chapter 3 suggests that incorporating competitive clauses into true-false practice test items may facilitate both broad retrieval of content when practice testing individually and support tracking of which associations of concepts are correct versus incorrect amongst learners practice testing collaboratively. In Experiments 2 and 4, practice testing with competitive-clause true-false items resulted in learning of both previously tested and previously related content and generally little evidence of negative suggestion. In fact, in Experiment 2 learners who practiced tested collaboratively with competitive true-false items demonstrated less evidence of negative suggestion than those who practice tested individually, indicating that the effects of collaboration and adding competitive clauses to true-false items might synergize to

provide a particularly useful implementation of practice testing. One potential explanation for this effect is that collaboration offered the opportunity for learners to share knowledge and correct each other's errors, and that these processes were facilitated by both the target information and the competitive incorrect alternative being presented within the true-false practice test item.

Paired Flashcard Practice Supports More Accurate Metacognitive Judgments Than Individual Flashcard Practice During Self-Regulated Learning

Chapter 2 and Chapter 3 investigated the impact on learning of collaborative versus individual practice testing for test formats that students may not necessarily use while practice testing in daily life, as constructing multiple-choice or true-false test items oneself is time-consuming and students may not have access to high-quality practice test items in these formats (although there are efforts to change that; Paquette-Smith et al., 2023). Students do, however, report frequently practice testing with flashcards in their own study sessions (Wissman et al., 2012; Zung et al., 2022), so Chapter 4 investigated the differential learning effects of collaborative (paired, in this case) and individual flashcard practice on learning. To more closely emulate the conditions of self-regulated learning, learners in these experiments were not closely supervised and, aside from a few basic instructions, generally allowed to practice how they would like, in terms of item order and time spent on each item. As our goal was to investigate a test format that students commonly use during self-regulated learning, we additionally explored whether learners might more accurately assess their learning following collaborative as compared to individual flashcard usage, as accurate monitoring of one's learning can have important implications for whether or not learners make optimal decisions during self-regulated learning (Nelson & Leonesio, 1988).

The results of Chapter 4 somewhat contrast the general trends in the results of Chapters 2 and 3. In Chapter 4, individual and collaborative flashcard practice resulted in similar rates of learning challenging word-definition pairs when learning was assessed using an immediate individual cued-recall test (Experiments 1 and 2) and a 24-hr delayed test (Experiment 2). Metacognitive calibration on the immediate test, however, substantially differed across the two implementations of practice testing in both experiments. Learners who used flashcards individually significantly overestimated future cued-recall test performance–on average by roughly 20%–whereas learners who used flashcards in pairs were much more accurate in predicting their future test performance. This tendency by individual learners to considerably overestimate their future test performance resolved after a delay. Overestimating immediate test performance, however is particularly important for optimal management of one's learning because current evaluations as to the state of one's learning may inform decisions about when to drop cards from study (which is associated with less learning than not dropping; Kornell & Bjork, 2008) or when to terminate the learning session altogether. In fact, in Experiment 1, when dropping cards from study was allowed, individual learners were far more likely to drop cards from study than paired learners (and often reported that they did so because they believed that they had sufficiently learned the information).

Experiment 2 ruled out the possibility that differential rates in dropping flashcards from study was responsible for individual learners' poor metacognitive calibration by prohibiting dropping. Even when dropping was prohibited individual learners still demonstrated greater overconfidence on their learning than paired learners, suggesting that some other feature(s) of working with others such as overt retrieval, feelings of accountability, or motivation to seek out consistent feedback, may be driving these observed differences in metacognitive calibration.

Together, the findings of Chapter 4 suggest that collaborative flashcard practice might yield similar learning to individual flashcard practice. But, using flashcards with a partner may promote behaviors and processes that could improve management of one's learning.

What Might Drive the Benefits of Collaborative Practice Testing?

One of the exciting–but also challenging–aspects of studying collaborative practice testing is that collaboration can foster a rich and dynamic collection of behaviors, interpersonal processes, and cognitions simultaneously. In a single session of collaborative practice testing, for example, learners may produce overt retrieval responses, offer and/or receive elaborative explanations, monitor learning, engage in conflict resolution, and more, which might all simultaneously impact learning. Existing research on best practices in the implementation of retrieval practice (Roediger & Butler, 2011) and collaborative learning (Johnson et al., 1998) point to factors–such as possessing non-overlapping bodies of knowledge (Chapter 2), offering and receiving elaborative explanations (Chapter 3), and engaging in overt retrieval (Chapter 4)–which may be key drivers of these benefits of collaboration during practice testing. The work presented here offers compelling evidence that testing together can enhance learning and learning-related outcomes in ways that are distinct from testing alone, but there are myriad future avenues for research to explore the mechanisms underlying these effects (e.g., Marquez et al., 2023).

Concluding Comments

A robust body of work suggests that practice testing is a uniquely powerful tool for learning. Much of that research, however, has centered on individual practice testing rather than collaborative practice testing. This work offers evidence that testing together can benefit learning (Chapters 2 and 3) and learning-relevant outcomes (Chapter 4) across multiple-choice,

true-false, and flashcard-based practice test formats in the laboratory and in the classroom. Although future work should continue efforts to disentangle the effects on learning of the myriad processes and behaviors elicited during collaborative practice testing, the research presented here suggests that learners may have much to gain by practice testing with others.

APPENDIX A

Group-Building and Neutral Activities Used in Chapter 2

Section activities completed in Fall 2020

Week	Neutral Activity	Group-Building Activity
1 (First lecture)	This course uses small-group work during synchronous laboratory sections of the course. Laboratory groups are intended to help you complete lab assignments and ask questions to your TAs.	This course uses small-group work during synchronous laboratory sections of the course. Laboratory groups are intended to help you complete lab assignments in a manageable amount of time and help you learn class material. Laboratory groups are a great resource to help you succeed in the course and getting to know your groups and being engaged during laboratory group sessions will help you do better in the course.
2	Think about all you've learned about the science of learning. From your own experiences, or from what you've learned, list two ways this information could help you succeed in the course.	Think about the group work you will do during laboratory sections this quarter. From your own experiences, come up with at least two potential ways group work could help you succeed in the course.

3	Read through the list of techniques to be more productive: **-Reduce screen clutter**. Close all other tabs besides the ones needed for the activity **-Reduce external distractions.** Put your phone away or put your phone on "do not disturb mode" **-Read first.** Look through the whole assignment before beginning to work on it and note parts of the activity that may take more time or effort **-Delegate tasks.** Delegate tasks ("divide and conquer") when appropriate **-Set a time limit.** Set an approximate time limit for each important task or question **-Monitor your progress.** Periodically assess how many major tasks/questions you have	Read through the list of the following strengths: **-Cooperating** (interested in the views of other group members) **-Clarifying** (listening and summarizing discussion) **-Inspiring** (encourages participation and progress) **-Harmonizing** (encouraging cohesion and collaboration) **-Risk taking** (take a chance on trying something new for group success) **-Process checking** (checking the agenda and timing) In your group, give each group member the opportunity to share (out loud) how they intend to use one of techniques above to help the group be more productive during today's meeting. Write down the technique number that each person is using (it's fine if group members select the same technique).

completed and how many more you have to complete.

In your group, give each group member the opportunity to share (out loud) how they intend to use one of techniques above to help the group be more productive during today's meeting. Write down the technique number that each person is using (it's fine if group members select the same technique).

4

In your next lab you'll be working with the same group members as last week, so think back to your group work strengths:

-Cooperating (interested in the views of other group members)

-Clarifying (listening and summarizing discussion)

-Inspiring (encourages participation and progress)

		-Harmonizing (encouraging cohesion and collaboration) **-Risk taking** (take a chance on trying something new for group success) **-Process checking** (checking the agenda and timing) Write down one thing you plan to do to help your group work well together during the next lab.
6	Reflect on something you've learned about in this class that you could use in your own life. Have each group member write down their response.	Reflect on how working with a group helped or hindered your performance during this lab. What could you do next time to make your group work better together? Have each group member write down their response.
7 (Practice testing activity)	When collaborating, each group member should have the survey open on their screen. Talk to one another and help each other figure out the answers, but each group member should submit their own quiz round.	When collaborating, use a driver-navigator procedure. In other words, one person should be sharing their screen and typing for the group with the other group members helping to answer the questions. A different person should be the driver for each quiz round.

7 (End of first group practice test)	*Discuss the following question with your group. There is no need to write anything down!* Did every group member participate equally during the review round? Think about 1 or 2 SPECIFIC things **your group** could do to encourage group members to participate equally.

Section activities completed in Winter 2021

Johnson et al.'s (1998) principles of collaboration	Group-Building Activity
Positive interdependence	When collaborating, use a driver-navigator procedure. In other words, one person should be sharing their screen and typing for the group with the other group members helping to answer the questions.
Individual accountability	At the end of the lab, we're going to come back to the main room and have you complete a few questions about the topics covered on the GROUP rounds. If **90% or more of the class** answers

these questions correctly, the entire class will earn one of their points for this lab.

Therefore, it is important that you all work together during the first two review rounds to make sure that all group members understand the material well.

Promotive interaction

Collaboration is not always easy but there are behaviors or actions that can support good group work. Here are some productive group behaviors you may use to help your group work go well.

-Cooperating (interested in the views of other group members)

-Clarifying (listening and summarizing discussion)

-Inspiring (encourages participation and progress)

-Harmonizing (encouraging cohesion and collaboration)

-Risk taking (take a chance on trying something new for group success)

-Process checking (checking the agenda and timing)

Think about one of those productive group behaviors you plan on using today. Write it below.

Social skills

Get to know your group members! Find out one thing all of you have in common and write it below (e.g., we all have a dog, we've all visited a certain place).

Group processing	*Discuss the following question with your group. There is no need to write anything down!* Did every group member participate equally during the review round? Think about 1 or 2 SPECIFIC things **your group** could do to encourage group members to participate equally.

Sample of Practice Testing Activity Questions

Biological Psychology

Which of these statements best describes the difference between the forebrain and the hindbrain?

-The hindbrain is twice as large as the forebrain

-Damage to the forebrain impairs respiration and damage to the hindbrain affects language

-The forebrain controls higher level processing while the hindbrain controls basic biological functions*[11]

-The forebrain contains the reticular formation and the hindbrain contains the thalamus

-The hindbrain is connected to the forebrain by the corpus callosum

Learning

Ralph has been conditioned to fear a stuffed teddy bear by a researcher who has paired the presentation of the teddy pair with very loud music. Ralph begins to show fear to other stuffed teddy bears that share many similar features to the original stuffed teddy bear. Which principle is Ralph showing?

---

[11] * indicates the correct answer.

-Assimilation

-Generalization*

-Accommodation

-Discrimination

-Second-order conditioning

Research Methods

Tamar is testing whether women's decrease in body satisfaction is larger after looking at cosmetics advertisements versus looking at neutral images. She randomly assigns a group of women to look at either advertisements for cosmetic products or pictures of trees, and then complete a survey about their body satisfaction. What type of design is Tamar using?

-Experimental*

-Quasi-experimental

-Correlational

-Deceptive

-Double-blind

Sensation and Perception

Bo is a search-and-rescue dog. He can detect the scent of a person on a scrap of cloth 30 feet away 50% of the time. This is demonstrating Bo's _____.

-Difference threshold

-Absolute threshold*

-Decision criteria

-Top-down processing

-Sensory adaptation

APPENDIX B

Practice Test Performance in Experiments 1-4

Experiment 1

In order to examine whether individual and groups performed differently on the True vs. false practice we conducted A 2 (*Initial Practice*: True or False) x 2 (*Practice Test Setting*: Individual or Collaborative) ANOVA was conducted with initial practice as a within-subjects factor, practice test setting as a between-subjects factor, and practice test performance as the dependent variable. Question type was not included in the model because that factor is only relevant for final test questions. The interaction between initial practice and practice test setting was nonsignificant, F (1, 108) = 0.02, p = .89, η_p^2 < .001. The main effect of initial practice was significant, F (1, 108) = 30.67, p < .001, η_p^2 = .22. Across both individual and collaborative practice testing, a greater proportion of true statements (M = .71, SD = .18) were evaluated correctly as compared to false statements (M = .58, SD = .24). The effect of practice test setting was nonsignificant, F (1, 108) = 2.97, p = .088, η_p^2 = .03. Numerically, collaborative practice testing (M = .67, SD = .17) resulted in slightly higher practice test scores than individual practice testing (M = .62, SD = .16), but, again, this difference was not statistically reliable.

Experiment 2

A 2 (*Initial Practice*: True or False) x 2 (*Practice Test Setting*: Individual or Collaborative) ANOVA was conducted with initial practice as a within-subjects factor, practice test setting as a between-subjects factor, and practice test performance as the dependent variable. The initial practice x practice test setting interaction was nonsignificant, F (1, 115) = 0.34, p = .56, η_p^2 = .003, as was the main effect of initial practice, F (1, 115) = 0.002, p = .96, η_p^2 < .001,

True: $M = .68$, $SD = .21$, False: $M = .68$, $SD = .26$. Unlike in Experiment 1, participants did equally well on true and false practice test items, a change driven by the increase in practice test performance on false items from Experiment 1 to Experiment 2. The main effect of practice test setting was nonsignificant, $F\ (1, 115) = 3.73$, $p = .056$, $\eta_p^2 = .03$; in line with the pattern of results obtained in Experiment 1, numerically collaborative practice testing ($M = .71$, $SD = .15$) produced slightly higher practice test performance than individual practice testing ($M = .65$, $SD = .19$), but this difference was not statistically reliable.

Experiment 3

A 2 (initial practice: True or False) x 2 (Practice Test Setting: Individual or Collaborative) fully within-subjects ANOVA was conducted with practice test performance as the dependent variable. Again, question type was not included in the model because that factor is only relevant to final test questions. The main effect of setting was significant, $F\ (1, 507) = 143.65$, $p < .001$, $\eta_p^2 = .22$, as was the main effect of initial practice, $F\ (1, 507) = 562.80$, $p < .002$, $\eta_p^2 = .53$. As in Experiment 1, true statements ($M = .87$, $SD = .14$) were more often evaluated correctly than false statements ($M = .64$, $SD = .19$). Unlike in Experiment 1, groups ($M = .82$, $SD = .18$) demonstrated higher practice test scores than individuals ($M = .68$, $SD = .19$). These main effects were qualified by a significant initial practice x practice test setting interaction, $F\ (1, 507) = 41.74$, $p < .001$, $\eta_p^2 = .076$. Follow-up paired samples t-tests revealed that students performed better on true practice statements than false practice statements both when testing individually and when testing collaboratively (all p's $< .001$) but that this difference was larger when practice testing individually ($M_{diff} = .29$) than when practice testing collaboratively ($M_{diff} = .17$).

Experiment 4

A 2 (*Initial Practice*: True or False) x 2 (*Practice Test Setting*: Individual or Collaborative) fully within-subjects ANOVA was conducted with practice test performance as the dependent variable. There was a main effect of initial practice, $F\,(1, 472) = 148.66, p < .001$, $\eta_p^2 = .24$ and a main effect of practice test setting, $F\,(1, 472) = 62.93, p < .001, \eta_p^2 = .12$. As in previous experiments, participants evaluated more true practice statements correctly ($M = .86$, $SD = .14$) than false practice statements ($M = .75, SD = .19$). As in Experiment 3, participants evaluated more practice statements correctly when they worked in groups ($M = .85, SD = .17$) than when they worked alone ($M = .76, SD = .20$). The initial practice x practice test setting interaction was nonsignificant, $F\,(1, 472) = 0.12, p = .74, \eta_p^2 < .001$.

APPENDIX C

List of GRE Vocabulary Word-Definition Pairs

No.	GRE Word	Definition
1.	Abeyance	temporary inactivity, cessation, or suspension
2.	Abjure	to renounce, repudiate, or retract, especially with formal solemnity; recant
3.	Anodyne	a medicine that relieves or allays pain
4.	Canard	a false or baseless, usually derogatory story, report, or rumor
5.	Cosset	to treat as a pet; pamper; coddle
6.	Ebullient	overflowing with fervor, enthusiasm, or excitement; high-spirited
7.	Ersatz	serving as a substitute; synthetic; artificial
8.	Expiate	to atone for; make amends or reparation for
9.	Fracas	a noisy, disorderly disturbance or fight; riotous brawl; uproar
10.	Fusillade	a simultaneous or continuous discharge of firearms
11.	Gainsay	to deny, dispute, or contradict
12.	Hermetic	made airtight by fusion or sealing
13.	Impugn	to challenge as false (another's statements, motives); cast doubt upon
14.	Lachrymose	suggestive of or tending to cause tears; mournful
15.	Lambaste	to beat or whip severely
16.	Maelstrom	a large, powerful, or violent whirlpool
17.	Monolithic	made of only one stone
18.	Munificent	extremely liberal in giving; very generous

19.	Myopic	unable or unwilling to act prudently; shortsighted
20.	Noisome	offensive or disgusting, as an odor
21.	Occlude	to close, shut, or stop up (a passage, opening)
22.	Paean	any song of praise, joy, or triumph
23.	Panoply	a wide-ranging and impressive array or display
24.	Pellucid	allowing the maximum passage of light, as glass; translucent
25.	Polemic	a controversial argument, as one against some opinion, doctrine
26.	Prosaic	commonplace or dull; matter-of-fact or unimaginative
27.	Puerile	of or relating to a child or to childhood
28.	Pundit	a learned person, expert, or authority
29.	Quiescent	being at rest; quiet; still; inactive or motionless
30.	Quixotic	extravagantly chivalrous or romantic; visionary, impractical, or impracticable
31.	Redress	the setting right of what is wrong
32.	Sanguine	cheerfully optimistic, hopeful, or confident
33.	Soporific	causing or tending to cause sleep
34.	Supine	lying on the back, face or front upward
35.	Tyro	a beginner in learning anything; novice
36.	Upbraid	to find fault with or reproach severely; censure
37.	Verdant	green with vegetation; covered with growing plants or grass
38.	Vitiate	to impair the quality of; make faulty; spoil
39.	Vitriol	something highly caustic or severe in effect, as criticism
40.	Welter	to roll, toss, or heave, as waves or the sea

APPENDIX D

Parallel Analyses for Chapter 4

Parallel Analyses #1: Including First Learning Activity as a Factor

All analyses were conducted using a between-participants Analysis of Variance (ANOVA) with factors of First Learning Activity (Study or Test) and Condition (Individual or Paired) 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.

Learning Phase

Number of Learning Cycles

Participants indicated the number of learning cycles that they had completed in each 20-min learning session. These responses were summed for a total number of learning cycles in the entire learning phase; if participants stated getting through some, but not all, of the words during a particular learning cycle (e.g., "*I got through all of the words once and some of the words twice*"), then that partial cycle was scored by adding 0.50 to the number of complete learning cycles. There was a main effect of Condition such that those learning individually got through significantly more learning cycles ($M = 5.36$, $SD = 1.64$) than those learning in pairs ($M = 4.32$, $SD = 1.44$), $F\ (1, 148) = 17.09$, $p < .001$, $\eta_p^2 = .10$. There was no significant main effect of First Learning Activity, $F\ (1, 148) = 0.09$, $p = .77$, $\eta_p^2 = .001$ and no significant First Learning Activity x Condition interaction, $F\ (1, 148) = 0.32$, $p = .57$, $\eta_p^2 = .002$.

Dropping of Flashcards

Participants reported whether they had dropped flashcards from study, and if so, why they chose to do so. These data were open-ended and were coded by two independent raters blind to condition. The rates demonstrated good interrater reliability (Cohen's $\kappa = .99$ and .85

for whether or not dropped and why, respectively). Two Chi-square tests were performed: one for Condition and one for First Learning Activity. The findings for Condition are reported in the main text of the manuscript. A Chi-square test of First Learning Activity revealed that those who studied first dropped flashcards at a similar rate (65%) as those who tested first (74%), χ^2 (2) = 1.24, p = .54.

Final Cued Recall Test

Overall Performance

As the GRE vocabulary words that participants learned in this experiment were difficult to spell (e.g., *lachrymose*), we used an accuracy threshold of 75% on the final cued recall test. In other words, participants' spelling of a word had to be a 75% or greater match to the actual spelling of the word to be counted as correctly recalled. There was no significant main effect of First Learning Activity, F (1, 148) = 0.05, p = .83, η_p^2 < .001 or Condition, F (1, 148) = 1.65, p = .20, η_p^2 = .01. There was also no significant First Learning Activity x Condition interaction, F (1, 148) = 0.38, p = .54, η_p^2 = .003.

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 JOL and final test performance for both conditions and for both types of First Learning Activity (Study First or Test First; see Figure 2). Participants that initially studied demonstrated moderate-to-large correlations between their JOL and final test performance when learning individually, r (29) = .67, p < .001, and with a partner, r (42) = .75, p < .001, whereas those that initially tested demonstrated moderate JOL-test performance correlations when learning individually, r (30) = .50, p = .004, and with a partner, r (42) = .45, p = .002. Those correlations were lower than for those that tested immediately prior

to making a JOL, but were still somewhat accurate. It thus appears that engaging in retrieval practice immediately prior to predicting future performance enhances the accuracy of metacognitive judgments. Further, although the magnitude of the relationships between JOL 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 JOL data) differ, prompting further analyses of participants' metacognitive calibration.

Metacognitive Calibration. We computed metacognitive calibration by subtracting participants' actual test performance from their JOLs, 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, JOL-test performance and metacognitive calibration scores provide complementary, but distinct, information about learners' metacognitive judgments.

A between-participants analysis of variance (ANOVA) on participants' metacognitive calibration scores with factors of First Learning Activity (Study or Test) and Condition (Individual or Paired) revealed a main effect of Condition, $F\,(1, 147) = 32.20, p < .001, \eta_p^2 = .18$. Individual learners were overconfident ($M = .20, SD = .22$), whereas Paired learners were relatively accurate ($M = .00, SD = .22$). There was no significant main effect of First Learning Activity, $F\,(1, 147) = 0.12, p = .73, \eta_p^2 = .001$, and there was a marginally significant interaction between Condition and First Learning Activity, $F\,(1, 147) = 3.23, p = .07, \eta_p^2 = .02$. This interaction is attributed to Individual learners that initially studied ($M = .18, SD = .21$) being less

overconfident than learners that initially tested (M = .23, SD = .23), with the opposite pattern occurring for the Paired learners (First Learning Activity: Study First, M = .04, SD = .18; Test First, M = -.04, SD = .25). Post hoc t-tests, however, revealed no significant difference in metacognitive calibration between participants that initially studied or tested in the Individual condition, t (1, 61) = 0.94, p = .35, d = 0.24, 95% CI [-.16, .06], or in the Paired condition, t (1, 86) = 1.67, p = .10, d = 0.36, 95% CI [-.01, .17].

Learning Efficiency Analysis

We computed a measure of learning efficiency by dividing the proportion correct on the cued-recall test by the number of learning cycles that participants had completed. This calculation provided a measure of learning gain per cycle. To ease interpretation of the findings, results were multiplied by 100 so that numbers can be interpreted as percentage of definitions recalled as a function of number of learning cycles. We observed no significant main effect of Condition, F (1, 148) = 1.73, p = .19, η_p^2 = .01, or First Learning Activity, F (1, 148) = 0.004, p = .95, η_p^2 < .001, and no significant First Learning Activity x Condition interaction, F (1, 148) = 0.06, p = .80, η_p^2 < .001.

However, as a significantly greater proportion of individual learners dropped flashcards from study relative to the paired condition, a follow-up analysis compared learning efficiency only among participants that did not drop flashcards. There was a marginally significant main effect of Condition, F (1, 99) = 3.35, p = .07, η_p^2 = .03 such that the Paired condition (M = 10.81, SD = 6.23) demonstrated greater learning efficiency than the Individual condition (M = 8.42, SD = 3.81). There was no significant main effect of First Learning Activity, F (1, 99) = 0.10, p = .75, η_p^2 = .001 and no significant First Learning Activity x Condition interaction, F (1, 99) = 0.32, p = .57, η_p^2 = .003.

Positive and Negative Affect

Separate analyses were conducted for the positive affect subscale and negative affect subscale of the PANAS. There was no significant main effect of First Learning Activity, F (1, 148) = 0.09, p = .77, η_p^2 = .001 or Condition, F (1, 148) = 0.03, p = .86, η_p^2 < .001, and no significant First Learning Activity x Condition interaction, F (1, 148) = 0.04, p = .84, η_p^2 < .001 for positive affect. There was a significant main effect of Condition for negative affect, F (1, 148) = 9.23, p = .003, η_p^2 = .06. Those in the Paired condition reported greater negative affect (M = 16.93, SD = 6.54) than those in the Individual condition (M = 14.20, SD = 3.52). There was no main effect of First Learning Activity, F (1, 148) = 0.78, p = .38, η_p^2 = .005 and no First Learning Activity x Condition interaction, F (1, 148) = 1.43, p = .23, η_p^2 = .01.

Attentional Focus

Participants reported the percentage of time they were focused on the experimental tasks during the study. There was no main effect of Condition, F (1, 148) = 0.61, p = .44, η_p^2 = .004 or of First Activity, F (1, 148) = 0.31, p = .58, η_p^2 = .002. There was also no significant First Learning Activity x Condition interaction, F (1, 148) = 1.17, p = .28, η_p^2 = .008.

Parallel Analyses #2: Final Test Performance Scored Using Strict Criterion

In these analyses, test performance was assessed with strict accuracy as opposed to the more lenient threshold of 75% presented in the main text. Strict accuracy means that the participant's response must match the correct response 100% to be marked as correct (i.e. does not allow for spelling errors). Analyses that included First Learning Activity and Condition as factors were conducted using a 2 (First Learning Activity: Study or Test) x 2 (Condition: Individual or Paired) ANOVA. Analyses that included only Condition as a factor (i.e. comparing Individual and Paired conditions) were conducted using an independent samples t-test

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.

Final Cued Recall Test

Overall Performance

Including First Learning Activity and Condition as factors. There was no significant main effect of First Learning Activity, F (1, 148) = 0.07, p = .80, η_p^2 < .001 or Condition, F (1, 148) = 1.07, p = .30, η_p^2 = .007. There was also no significant First Learning Activity x Condition interaction, F (1, 148) = 0.21, p = .65, η_p^2 = .001.

Including only Condition as a factor. Comparing Paired and Individual conditions, there were no significant differences between those who learned individually (M = .36, SD = .22) and those learned with a partner (M = .32, SD = .23), t (150) = 1.04, p = .30, d = 0.17, 95% CI [-.03, .11].

Learning Efficiency Analysis

We computed a measure of learning efficiency by dividing the proportion correct (scored using the strict accuracy criterion) on the cued-recall test by the number of learning cycles that participants had completed. This calculation provided a measure of learning gain per cycle. To ease interpretation of the findings, results were multiplied by 100 so that numbers can be interpreted as percentage of definitions recalled as a function of number of learning cycles.

Including First Learning Activity and Condition as factors. We observed no significant main effect of Condition, F (1, 148) = 1.06, p = .31, η_p^2 = .007, or First Learning Activity, F (1, 148) = 0.06, p = .82, η_p^2 < .001, and no significant First Learning Activity x Condition interaction, F (1, 148) = 0.01, p = .92, η_p^2 < .001.

Including only Condition as a factor. The Individual (M = 7.15, SD = 5.09) and Paired

(M = 8.50, SD = 9.57) conditions demonstrated similar learning efficiency, t (150) = 1.04, p = .30, d = 0.17, 95% CI [-3.96, 1.24].

However, as a significantly greater proportion of individual learners dropped flashcards from study relative to the paired condition, a follow-up analysis compared learning efficiency only among participants that did not drop flashcards.

Including First Learning Activity and Condition as factors. There was no significant main effect of Condition, F (1, 99) = 2.36, p = .13, η_p^2 = .02 and no significant main effect of First Learning Activity, F (1, 99) = 0.05, p = .82, η_p^2 = .001, as well as no significant First Learning Activity x Condition interaction, F (1, 99) = 0.16, p = .69, η_p^2 = .002.

Including only Condition as a factor. Levene's test of equality of variances was significant, F (1, 102) = 8.08, p = .005, so an independent samples t-test with equal variances not assumed was conducted. The Paired condition (M = 7.86, SD = 5.76) demonstrated significantly greater learning efficiency than the Individual condition (M = 6.02, SD = 3.12), t (87.71) = 2.06, p = .04, d = 0.35, 95% CI [-3.60, -0.07].

Metacognitive Judgments

Correlations with Final Test Performance

A series of bivariate correlations related global judgments of learning (JOL) and final test performance (scored using strict criterion) for each of the experimental conditions. Those who studied as their first learning activity (and consequently tested themselves as their second learning activity) demonstrated moderate-to-large correlations between their JOL and final test performance when both learning individually, r (29) = .58, p = .001 and with a partner, r (42) = .73, p < .001. In other words, those who tested themselves immediately prior to judging their learning were fairly accurate in their learning assessment. Those who tested as their first

learning activity (and thus studied as their second learning activity) demonstrated moderate JOL-test performance correlations when learning individually, r (30) = .45, p = .01, and when learning with a partner, r (42) = .40, p = .008. Though conditions that studied immediately prior to making a JOL had lower correlations than those who tested immediately prior to doing so, these moderate correlations still indicate some accuracy in participants' judgments.

Metacognitive Calibration

We computed metacognitive calibration by subtracting participants' actual test performance (using strict scoring criterion) from their JOLs, with positive scores indicating overconfidence and negative scores indicating underconfidence. There was a main effect of Condition, F (1, 147) = 33.89, p < .001, η_p^2 = .19. Those in the Individual conditions were overconfident (M = .33, SD = .23) whereas those in the Paired conditions were less overconfident in their prediction of future test performance (M = .11, SD = .23). There was no significant main effect of First Learning Activity, F (1, 147) = 0.11, p = .74, η_p^2 = .001 or significant interaction between First Learning Activity and Condition, F (1, 147) = 2.27, p = .13, η_p^2 = .02.

REFERENCES

Aramovich, N. P., & Larson, J. R. (2013). Strategic demonstration of problem solutions by groups: The effects of member preferences, confidence, and learning goals. *Organizational Behavior and Human Decision Processes*, *122*(1), 36-52. https://doi.org/10.1016/j.obhdp.2013.04.001

Barber, S. J., Rajaram, S., & Aron, A. (2010). When two is too many: Collaborative encoding impairs memory. *Memory & Cognition*, *38*, 255-264. https://doi.org/10.3758/MC.38.3.255

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.

Biasutti, M., & Frate, S. (2018). Group metacognition in online collaborative learning: Validity and reliability of the group metacognition scale (GMS). *Educational Technology Research and Development*, *66*, 1321-1338. https://doi.org/10.1007/s11423-018-9583-0

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium*, pp. 123–144. Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35-67). Erlbaum.

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

Bloom, D. (2009). Collaborative test taking: Benefits for learning and retention. *College Teaching*, *57*, 216-220.

Blumen, H. M., & Rajaram, S. (2008). Influence of re-exposure and retrieval disruption during group collaboration on later individual recall. *Memory*, *16*(3), 231-244. https://doi.org/10.1080/09658210701804495

Blumen, H. M., & Stern, Y. (2011). Short-term and long-term collaboration benefits on individual recall in younger and older adults. *Memory & Cognition*, *39*, 147-154. https://doi.org/10.3758/s13421-010-0023-6

Brabec, J. A., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2021). True-false testing on trial: Guilty as charged or falsely accused? *Educational Psychology Review*, *33*, 667-692 https://doi.org/10.1007/s10648-020-09546-w

Brown, A. S., Schilling, H. E. H., & Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology, 91*(4), 756–764. https://doi.org/10.1037/0022-0663.91.4.756

Bruffee, K. A. (1984). Collaborative learning and the "conversation of mankind." *College English*, *46*(7), 635-652. https://www.jstor.org/stable/376924

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 41-50. https://doi.org/10.1080/02602930020022273

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323-331. https://doi.org/10.1016/j.jarmac.2018.07.002

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514-527. https://doi.org/10.1080/09541440701326097

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616. https://doi.org/ 10.3758/MC.36.3.604

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268-276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*, 496-511. https://doi.org/10.1038/s44159-022-00089-1

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438-448. https://doi.org/10.3758/MC.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633–642. https://doi.org/10.3758/BF03202713

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Clark, S. E., Hori, A., Putnam, A., & Martin, T. P. (2000). Group collaboration in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1578-1588. https://doi.org/10.1037/0278-7393.26.6.1578

Cocks, A. W. (1929). *The pedagogical value of the true-false examination*. Baltimore: Warwick & York.

Cooke, J. E., Weir, L., & Clarkston, B. (2019). Retention following two-stage collaborative exams depends on timing and student performance. *CBE—Life Sciences Education*, *18*(2), 1-8. https://doi.org/10.1187/cbe.17-07-0137

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., & DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education*, *27*, 102-108. https://doi.org/10.1152/advan.00041.2002

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271-280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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.

Ebel, R. L. (1970). The case for true-false test items. *The School Review*, *78*(3), 373-389. https://www.jstor.org/stable/1084159

Ebel, R. L. (1975). Can teachers write good true-false test items? *Journal of Educational Measurement*, *12*(1), 31-35. https://www.jstor.org/stable/1434372

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & Cognition, 38(8), 1087-1100. https://doi.org/ *10.3758/MC.38.8.1087*

Faiella, F., & Ricciardi, M. (2015). Gamification and learning: A review of issues and research. *Learning and Knowledge Society*, *11*(3), 13-21.

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.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80*(2), 179–183. https://doi.org/10.1037/0022-0663.80.2.179

Forrester, W. R., & Tashchian, A. T. (2010). Effects of personality on attitudes toward academic group work. *American Journal of Business Education*, *3*(3), 39- 46. https://doi.org/10.19030/ajbe.v3i3.397

Garaschuk, K. (2022). Learning benefits of collaborative exams. *Journal for Research and Practice in College Teaching*, *7*(1), 42-55.

Garcia, M. A., & Kornell, N. (2014). Collector [Software]. Available from https://github.com/gikeymarcia/Collector.

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

Gillespie, D., Rosamond, S., & Thomas, E. (2006). Grouped out? Undergraduates' default strategies for participating in multiple small groups. *The Journal of General Education*, *55*(2), 81-102. https://www.jstor.com/stable/27798042

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. https://www.jstor.org/stable/43632038

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392-399. https://doi.org/10.1037/0022-0663.81.3.392

Goos, M., Galbraith, P., & Renshaw, P. (2002). Socially mediated metacognition: Creating collaborative zones of proximal development in small group problem solving. *Educational Studies in Mathematics*, *49*, 193-223. https://doi.org/10.1023/A:1016209010120

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16*, 107-112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hertzberg, O. E., Heilman, J. D., & Leuenberger, H. W. (1932). The value of objective tests as teaching devices in educational psychology classes. *The Journal of Educational Psychology*, *23*(5), 371-380. https://doi.org/10.1037/h0072896

Hevner, K. (1932). A method of correcting for guessing in true-false tests and empirical evidence in support of IT, *The Journal of Social Psychology*, *3*(3), 359-362. https://doi.org/10.1080/00224545.1932.9919159

Hillyard, C., Gillespie, D., & Littig, P. (2010). University students' attitudes about learning in small groups after frequent participation. *Active Learning in Higher Education*, *11*(1), 9-20. https://doi.org/10.1177/1469787409355867

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151-164. https://doi.org/10.1016/j.jml.2013.03.002

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass communication*. Princeton, N.J.: Princeton University Press.

Hyman, I. E., Cardwell, B. A., Roy, R. A. (2013). Multiple causes of collaborative inhibition in memory for categorised word lists. *Memory*, *21*(7), 875-890. https://doi.org/10.1080/09658211.2013.769058

Imundo, M. N., Clark, C. M., Bjork, E. L., & Paquette-Smith, M. (under review). The effects of collaborative practice testing on memory for course content in a college classroom.

Imundo, M. N., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2021). Where and how to learn: The interactive benefits of contextual variation, restudying, and retrieval practice for learning. *Quarterly Journal of Experimental Psychology*, *74*(3), 413-424. https://doi.org/10.1177/1747021820968483

Janssen, J., Erkens, G., & Kirschner, P. A. (2011). Group awareness tools: It's what you do with it that matters. *Computers in Human Behavior*, *27*(3), 1046–1058.

Järvelä, S., Malmberg, J., Sobocinski, M., Kirschner, P.A. (2021). Metacognition in Collaborative Learning. In: Cress, U., Rosé, C., Wise, A.F., Oshima, J. (eds) *International Handbook of Computer-Supported Collaborative Learning*. Springer. https://doi.org/10.1007/978-3-030-65291-3_15

Jersild, A. T. (1929). Examination as an aid to learning. *The Journal of Educational Psychology*, *20*(8), 602-609. https://doi.org/10.1037/h0070530

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, *38*, 365–379. https://doi.org/10.3102/0013189X09339057

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. https://doi.org/10.1080/00091389809602629

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experiment Psychology: Learning, Memory, and Cognition*, *33*(4), 704-719. https://doi.org/10.1037/0278-7393.33.4.704

Kim, M., & Ryu, J. (2013). The development and implementation of a web-based formative peer assessment system for enhancing students' metacognitive awareness and performance in ill-structured tasks. *Educational Technology Research and Development*, *61*(4), 549–561. https://doi.org/10.1007/s11423-012-9266-1

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349-370. https://doi.org/10.1037/0096-3445.126.4.349

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.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85-97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989–998. https://doi.org/10.1037/a0015729

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, *49*(4), 241-253. https://doi.org/10.3102/0013189X20912798

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.

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.

Linnenbrink-Garcia, L., Rogat, T. K., & Koskey, K. L. K. (2011). Affect and engagement during small group instruction. *Contemporary Educational Psychology*, *36*(1), 13-24. https://doi.org/10.1016/j.cedpsych.2010.09.001

Little, J. L. (2011). Optimizing multiple-choice tests as learning events. Dissertation. https://www.proquest.com/openview/9cc9a1a38a2c69242f52dc1fcfa4f8d7/1?pq-origsite=gscholar&cbl=18750

Little, J. L., & Bjork, E. L. (2010). Multiple-choice testing can improve the retention of nontested related information. Proceedings of the Annual Meeting of the Cognitive Science Society, 1535-1540. https://escholarship.org/content/qt6w80f5c3/qt6w80f5c3.pdf

Little, E. L., & Bjork, J. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*, 14-26. https://doi.org/10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. Psychological Science, 23(11), 1337-1344. https://doi.org/10.1177/0956797612443370

Little, J. L., Frickey, E. A., & Fung, A. K. (2019). The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(8), 1473–1485. https://doi.org/10.1037/xlm0000638

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. https://doi.org/10.1037/stl0000041

Lou, Y., Abrami, P. C., & d'Apollonia, S. (2001). Small group and individual learning with technology: A meta-analysis. *Review of Educational Research*, *71*(3), 449-521. https://doi.org/10.3102/00346543071003449

Lou, Y., Abrami, C. A., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, *66*, 423–458.

Lusk, M., & Conklin, L. (2003). Collaborative testing to promote learning. *Journal of Nursing Education*, *42*(3), 121-124. https://doi.org/10.3928/0148-4834-20030301-07

Marquez, E., Imundo, M. N., Denton, V., Brabec, J. A., Ramakrishnan, R., & Bjork, E. L. (2023). The influence of test format on group interactions during collaborative T/F practice testing. Poster. UCLA Psychology Undergraduate Research Conference.

Mashek, D. (2022, June 23). Collaboration is a key skill. So why aren't we teaching it? *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/collaboration-is-a-key-skill-so-why-arent-we-teaching-it/

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.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494-513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 480-499). Cambridge University Press. https://doi.org/10.1017/9781108235631.020

McDaniel, M. A., Roediger, H. L., McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200-206. https://doi.org/10.3758/BF03194052

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.

Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2020). The effect of feedback on metacognition – A randomized experiment using polling technology. *Computers & Education, 152*, 103885. https://doi.org/10.1016/j.compedu.2020.103885

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., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 676-686.

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.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). Cambridge, MA: MIT 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. https://doi.org/10.1080/13546783.2011.642206

Nokes-Malach, T. J., Zepeda, C. D., Richey, J. E., & Gadgil, S. (2019). Collaborative learning: The benefits and costs. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 500–527). Cambridge University Press. https://doi.org/10.1017/9781108235631.021

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. https://doi.org/10.1037/bul0000151

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? *Journal of Educational Psychology*.

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

Paquette-Smith, M., Imundo, M. N., & Clark, C. M. (2023). Encouraging retrieval practice. *Society for the Teaching of Psychology (STP) Psychology Tools eBook*. https://teachpsych.org/ebooks/psytoolbox

Perkins, D. V., & Saris, R. N. (2001). A "Jigsaw classroom" technique for undergraduate statistics courses. *Teaching of Psychology*, *28*(2), 111-113. https://doi.org/10.1207/S15328023TOP2802_09

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.

Rajaram, S. (2011). Collaboration both hurts and helps memory: A cognitive perspective. *Current Directions in Psychological Science*, *20*(2), 76-81. https://doi.org/10.1177/0963721411403251

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, *5*(6), 649-663. https://doi.org/10.1177/1745691610388763

Rapp D. N. (2016). The consequences of reading inaccurate information. *Current Directions in Psychological Science*, 25(4), 281-285. https://doi.org/10.1177/0963721416649347

Rapp, D. N., & Salovich, N. A. (2018). Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences*, *5*(2), 232-239. https://doi.org/10.1177/2372732218785193

Reinig, B. A., Horowitz, I., & Whittenburg, G. E. (2011). A longitudinal analysis of satisfaction with group work. *Group Decision and Negotiation*, *20*, 215-237. https://doi.org/10.1007/s10726-009-9173-y

Remmers H.H., Remmers E.M. (1926). The negative suggestion effect on true-false examination questions. *Journal of Educational Psychology*, 17, 52–56.

Rempel, B. P., Dirks, M. B., & McGinitie, E. G. (2021). Two-stage testing reduces student-perceived exam anxiety in Introductory Chemistry. *Journal of Chemical Education*, *98*, 2527-2535. https://doi.org/10.1021/acs.jchemed.1c00219

Rich, P. R., Donovan, A. M., & Rapp, D. N. (2023). Cause typicality and the continued influence effect. *Journal of Experimental Psychology: Applied*, *29*(2), 221-238.

Roberts, H. M., & Ruch, G. M. (1928). Minor studies on objective examination methods. *The Journal of Educational Research*, *18*(2), 112-116. 10.1080/00220671.1928.10879866

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.

Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Ross M., Spencer S.J., Blatz C.W., Restorick E. (2008). Collaboration reduces the frequency of false memories in older and younger adults. *Psychology and Aging*, *23*, 85–92. https://doi.org/10.1037/0882-7974.23.1.85

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.

Salonen, P., Vauras, M., & Efklides, A. (2005). Social interaction—What can it tell us about metacognition and coregulation in learning? *European Psychologist*, *10*(3), 199-208. https://doi.org/10.1027/1016-9040.10.3.199

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. Psychology Learning & Teaching, 16(2), 175-285. https://doi.org/10.1177/1475725717695149

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

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*(4), 713-724. https://doi.org/10.1086/426605

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592-604. https://doi.org/10.1037/0278-7393.4.6.592

Smith, S. M., & Handy, J. D. (2016). The crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory*, *24*(8), 1134-1141. https://doi.org/10.1080/09658211.2015.1071852

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712-1725. https://doi.org/10.1037/a0033569

Smith, C. M., & Tindale, R. S. (2010). Direct and indirect minority influence in groups. In R. Martin & M. Hewstone (Eds.), *Minority influence and innovation: Antecedents, processes and consequences* (pp. 263-284). New York: Psychology Press.

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

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*(5910), 122-124. https://doi.org/10.1126/science.1165919

Socratous, C., & Ioannou, A., (2022). Evaluating the impact of the curriculum structure on group metacognition during collaborative problem-solving using educational robotics. *TechTrends*, *66*, 771-783. https://doi.org/10.1007/s11528-022-00738-5

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

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications*, *1*(3). https://doi.org/10.1186/s41235-016-0003-x

Standlee, L. S., & Popham, W. J. (1960). Quizzes' contribution to learning. *Journal of Educational Psychology*, *51*(6), 322-325.

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.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true-false examinations. *The Journal of Educational Research*, *83*(2), 119-124. http://www.jstor.com/stable/27540378

Toppino, T. C., & Luipersback, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research*, *86*(6), 357-362. https://doi.org/10.1080/00220671.1993.9941229

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.

Uner, O., Tekin, E., & Roediger, H. L. (2021). True-false tests enhance retention relative to rereading. *Journal of Experimental Psychology: Applied*, *28*(1), 114-129. https://doi.org/10.1037/xap0000363

Vázquez-García, M. (2018). Collaborative-group testing improves learning and knowledge retention of human physiology topics in second-year medical students. Ad Physiol Educ, 42, 232-239. doi:10.1152/advan.00113.2017

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.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*(4), 465-478. https://doi.org/10.1016/S0022-5371(77)80040-6

Wissman, K. T., & Rawson, K. A. (2016). How do students implement collaborative testing in real-world contexts? *Memory*, *24*(2), 223-239. https://doi.org/10.1080/09658211.2014.999792

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

Woody, W. D., Woody, L. K., & Bromley, S. (2008). Anticipated group versus individual examinations: A classroom comparison. *Teaching of Psychology*, *35*, 13-17. https://doi.org/10.1080/00986280701818540

Wosnitza, M., & Volet, S. (2014). Trajectories of change in university students' general views of group work following one single group assignment: Significance of instructional context and multidimensional aspects of experience. *European Journal of Psychology of Education*, *29*, 101-115. https://doi.org/10.1007/s10212-013-0189-y

Yue, C. L., Soderstrom, N. C., & Bjork, E. L. (2015). Partial testing can potentiate learning of tested and untested material from multimedia lessons. *Journal of Educational Psychology*, *107*(4), 991-1005. https://doi.org/10.1037/edu0000031

Zheng, L., Li, X., Zhang, X., & Sun, W. (2019). The effects of group metacognitive scaffolding on group metacognitive behaviors, group performance, and cognitive load in computer-supported collaborative learning. *The Internet and Higher Education*, *42*, 13-24. https://doi.org/10.1016/j.iheduc.2019.03.002

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