Collaborative testing for key-term definitions under representative conditions: Efficiency costs and no learning benefits

Kathryn T. Wissman^{1,2} · Katherine A. Rawson¹

Published online: 28 August 2017
© Psychonomic Society, Inc. 2017

Abstract Students are expected to learn key-term definitions across many different grade levels and academic disciplines. Thus, investigating ways to promote understanding of key-term definitions is of critical importance for applied purposes. A recent survey showed that learners report engaging in collaborative practice testing when learning key-term definitions, with outcomes also shedding light on the way in which learners report engaging in collaborative testing in real-world contexts (Wissman & Rawson, 2016, *Memory*, 24, 223–239). However, no research has directly explored the effectiveness of engaging in collaborative testing under representative conditions. Accordingly, the current research evaluates the costs (with respect to efficiency) and the benefits (with respect to learning) of collaborative testing for key-term definitions under representative conditions. In three experiments ($ns = 94, 74, 95$), learners individually studied key-term definitions and then completed retrieval practice, which occurred either individually or collaboratively (in dyads). Two days later, all learners completed a final individual test. Results from Experiments 1–2 showed a cost (with respect to efficiency) and no benefit (with respect to learning) of engaging in collaborative testing for key-term definitions. Experiment 3 evaluated a theoretical explanation for why collaborative benefits do not emerge under representative conditions. Collectively, outcomes indicate that collaborative testing versus individual testing is less effective and less efficient when learning key-term definitions under representative conditions.

Keywords Collaborative testing · Key-term definitions · Retrieval practice

Key-term definitions are one of the most common kinds of information that students are expected to learn across many different grade levels and academic disciplines. Textbooks in many academic subjects are heavily populated by key terms, and classroom instruction is often directed at teaching students about these concepts. Thus, investigating ways to promote understanding of key-term definitions is of critical importance for applied purposes. Importantly, a recent survey showed that learners report engaging in collaborative practice testing when learning key-term definitions (Wissman & Rawson, 2016). However, no research has directly explored the effectiveness of collaborative testing when learning key terms. Furthermore, various aspects of the methods used in prior research on collaborative testing do not reflect how learners typically report implementing collaborative testing in a real-world collaborative setting. Accordingly, the current research evaluates the costs (with respect to efficiency) and the benefits (with respect to learning) of collaborative testing for key-term definitions under representative conditions.

Evaluating the effectiveness of collaborative testing for key-term definitions has important implications for applied purposes. For example, Wissman and Rawson (2016) asked learners about the extent to which they use collaborative testing and how they implement collaborative testing in real-world contexts. Importantly, 86% of learners reported that when studying in a group, some or most of the time was spent engaging in testing. Furthermore, of the learners that reported engaging in collaborative testing when studying in a group environment, 87% of learners reported using key-term definitions (Wissman & Rawson, 2016).

Although no study has investigated how collaborative testing affects the learning of key-term definitions in particular, recent

✉ Kathryn T. Wissman
kwissman@skidmore.edu

¹ Kent State University, Kent, OH, USA

² Present address: Department of Psychology, Tisch Learning Center, Skidmore College, Room #139, 815 N Broadway, Saratoga Springs, NY 12866, USA

collaborative memory research has shown that collaborative testing has learning benefits for other types of material (e.g., word lists). More specifically, research has shown that engaging in collaborative testing during practice enhances subsequent individual memory, referred to as *postcollaborative benefits* (Blumen & Rajaram, 2008, 2009; Blumen & Stern, 2011; Congleton & Rajaram, 2011; Wissman & Rawson, 2016). For example, Blumen and Stern (2011) presented learners with a list of 40 nouns, and all learners studied these individually. Following initial study, learners completed retrieval practice that occurred either individually or collaboratively. Learners then completed a final individual test immediately and then 1 week later. Results showed that learners who engaged in collaborative versus individual retrieval practice recalled significantly more words on the immediate test. The effect also emerged on the delayed test 1 week later, suggesting that postcollaborative benefits are robust for simple materials. Although these results are suggestive for more complex material, the extent to which postcollaborative benefits generalize to key-term definitions is unclear.

In addition to a potential learning benefit of engaging in collaborative testing for key-term definitions, it is also important to consider a potential cost—namely, the amount of time it takes learners to engage in collaborative testing under representative conditions. Evaluating how much time it takes students to learn information is important for applied purposes. More specifically, students are expected to learn and remember a substantial amount of information both within and across classes, and they only have a limited amount of time to do so in a real-world setting. Thus, achieving long-term learning while also considering efficiency is difficult (Rawson & Dunlosky, 2011). Although efficiency is an important factor from a student's point of view, efficiency is typically an overlooked outcome from an empirical standpoint. Important for current purposes, even if engaging in collaborative testing during practice leads to greater success on a final test, using collaborative testing for key-term definitions may not be worth a student's time if the learning technique takes substantially longer.

A final consideration motivating the current research concerns the conditions that are representative of those that students report using when engaging in collaborative testing for key-term definitions. When students in the survey study by Wissman and Rawson (2016) were asked about how they implement collaborative testing, the modal response (39%) was that they typically studied with one other person. Additionally, only 5% reported attempting to answer questions together; rather, the majority reported taking turns being the asker versus responder (with similar proportions reporting alternating after each trial or switching at the end of an item set). When asked about what occurs after an incorrect response during collaborative testing, 96% of students reported that the responder is provided with the correct answer as feedback. Finally, 66% of students reported engaging in collaborative testing proximal to the exam. These particular outcomes motivated the methods used in the current research, to reflect conditions under which learners report using collaborative

testing in real-world contexts. Specifically, learners studied key-term definitions and engaged in collaborative testing with one other person. During collaborative testing, learners implemented a turn-taking approach, with the correct answer being provided as feedback on each trial. Learners completed the final test 2 days after practice. Although this methodology differs from previous research, the current research was intended to examine the costs and benefits of engaging in collaborative testing for key-term definitions under representative conditions, which has important implications for student learning.

The goal of the current study was to evaluate the effectiveness of engaging in collaborative testing for key-term definitions under representative conditions. In three experiments, learners individually studied key-term definitions and then completed retrieval practice, which occurred either individually or collaboratively. Two days later, all learners completed a final individual test. All three experiments evaluated the effectiveness of collaborative testing in terms of both costs and benefits. Concerning the benefits of collaborative testing, performance on the final individual test was of primary interest in all three experiments. Concerning the costs of collaborative testing, in Experiment 1, we controlled the total amount of practice time and operationalized efficiency as the number of trials learners were able to complete during practice. In Experiment 2, we controlled the number of trials per key-term concept and operationalized efficiency as the amount of time it took learners to complete the practice trials. Thus, these two experiments involved converging methods (rather than direct replication) to examine the potential learning benefits and efficiency costs of collaborative testing for key terms under representative conditions. To foreshadow, Experiments 1–2 both showed significant costs of collaborative testing and no evidence of postcollaborative benefits on the final individual tests. Accordingly, Experiment 3 was designed to evaluate a theoretical explanation for why postcollaborative benefits did not emerge under the representative conditions implemented in the first two experiments.

Experiment 1

Method

Participants and design Undergraduates who participated for course credit ($n = 94$) were randomly assigned to one of two groups: collaborative or individual. Two participants were excluded due to noncompliance, three participants were excluded due to attrition, and eight participants were excluded due to computer errors. The final sample included $n = 44$ in the collaborative group and $n = 37$ in the individual group. The a priori targeted sample size for Experiment 1 was $n = 80$, based on sample sizes we have used in similar prior research. For the outcomes of greatest interest (paired comparisons of individual performance in the two groups on final test measures), with a power of .80 and a two-tailed $\alpha =$

.05, this sample size provided sufficient sensitivity to detect effects of $d = .63$ or larger based on a sensitivity power analysis using G*Power 3.1.5 (Faul, Erdfelder, Buchner, & Lang, 2009).

Materials Materials included a short passage on operant conditioning. The passage included 10 key terms and their definitions (see the [Appendix](#)).

Procedure *Phase 1.* Phase 1 was the same for all participants, and all tasks in this phase were completed individually. Participants studied the text containing the 10 key-term definitions on operant conditioning for 6 minutes. After initial study, participants restudied each key-term definition in isolation for 15 seconds. All participants then completed an initial practice cued recall test. On each trial, participants were shown the key term and given 30 seconds to recall the definition. After the initial cued recall test, all participants engaged in one last restudy trial.

Phase 2. After Phase 1, the research assistant told participants that the next phase of the experiment involved recall of the previously studied key-term definitions. Participants were told to imagine that they had an exam in 2 days and that they were sitting down to study the 10 key-term definitions for 30 minutes. Participants were also told that recall would be completed either collaboratively or individually. Dyads were randomly created by combining participants based on the day of the month on which they were born. More specifically, participants with the lowest days of the month were assigned to the dyad, with the lowest day designated as the scribe who would type in the responses during collaborative recall (e.g., for participants with birthdates of January 3, March 12, August 25, and December 1, dyad assignment was based on the numbers 1 and 3, with the participant with the birthdate on the 1st designated as the scribe). The remaining participants completed Phase 2 individually.

Participants in the individual group were shown the key term and asked to recall the definition on each test trial. After recalling the definition, participants were shown the correct answer along with their response and asked to make a monitoring decision about the accuracy of their response. Participants entered a decision about whether their response was right, wrong, or in between. Participants were then asked to make a control decision about whether they wanted to drop the key term or practice the key term again later. If participants elected to drop all items before 30 minutes had elapsed, participants were instructed to practice recalling all 10 key terms again.

Participants in the collaborative group sat at two computers and alternated asking versus answering on each recall trial. For example, Participant A asked their partner to recall “positive reinforcement” on the first recall trial, whereas Participant B asked their partner to recall “negative reinforcement” on the next recall trial. Concerning the monitoring decision, participants were instructed to work together however they saw fit, with the correct answer provided to the participant who asked their partner to recall the definition. Concerning the control decision about

whether to drop the key term or practice the key term again later, participants were instructed to work together however they saw fit but that regardless of how the decision was made, both learners needed to enter the same decision into each of their computers. If participants elected to drop all items before 30 minutes had elapsed, participants were instructed to practice all 10 key terms again but switched which key terms they asked versus answered (e.g., Participant A now asked their partner to recall “negative reinforcement,” whereas Participant B now asked their partner to recall “positive reinforcement”).

Phase 3. Two days later, all participants individually completed two final tests. For the cued recall test, learners were shown the key terms one at a time and asked to recall the definition for each key term. Following the cued recall test, participants completed a four-alternative multiple-choice test consisting of 20 comprehension questions, with two questions for each of the 10 key terms. Multiple-choice questions required inferencing and application of the key-term definitions. Split-half reliability was acceptable for the cued recall test ($\alpha = .93$) and the multiple-choice test ($\alpha = .76$).

Scoring All scoring of cued recall responses was completed by trained research assistants. Research assistants were blind to group assignment. Responses were scored based on the proportion of idea units contained in the learner’s response. Performance was calculated as proportion scores for each key term and then averaged across all 10 key terms. Thus, credit was given to fully correct and partially correct answers. Credit was given for verbatim responses or correct paraphrases of the idea units.

Results and discussion

To facilitate examination of the costs (with respect to efficiency) and benefits (with respect to learning), outcomes of primary interest are shown in Fig. 1. As shown in the left panel of Fig. 1, final cued recall test performance was significantly lower for the collaborative group versus the individual group, $t(79) = 2.44$, $p = .017$, $d = .54$. In addition, performance on the final multiple-choice test was also numerically lower for the collaborative group versus the individual group, $t(79) = 1.32$, $p = .190$, $d = .29$ (middle panel of Fig. 1). In contrast to prior research involving simpler verbal materials, the current results suggest that postcollaborative benefits do not emerge for key-term definitions. Most important for current purposes, these outcomes suggest that when students are given a fixed amount of time to engage in retrieval practice for key-term definitions, working individually versus collaboratively leads to a greater level of learning.

To diagnose potential costs of engaging in collaborative testing, we examined what occurred during Phase 2 retrieval practice.¹ Outcomes showed a substantial efficiency

¹ There were no differences on the initial cued recall test during Phase 1 for the collaborative group ($M = .32$, $SE = .02$) versus the individual group ($M = .35$, $SE = .03$), $t(79) = 0.74$, $p = .462$.

Fig. 1 Performance on outcomes of primary interest in Experiment 1, for learners who completed Phase 2 collaboratively in a dyad or working individually. Error bars represent standard errors

difference in the number of trials completed during practice in the individual group versus the collaborative group. As shown in the right panel of Fig. 1, learners in the individual group were able to complete 37 trials during practice, whereas learners in the collaborative group were only able to complete 28 trials in total during practice (which translates to 14 retrieval practice opportunities for each of the two learners). Thus, learners in the collaborative group versus the individual group were able to complete significantly fewer practice trials, $t(79) = 4.67$, $p < .001$, $d = 1.04$. The large efficiency difference in the of number of trials completed resulted in fewer correct responses during retrieval practice for the collaborative group versus the individual group (3 versus 12), $t(79) = 6.17$, $p < .001$, $d = 1.37$. Accordingly, one plausible explanation for why engaging in collaborative testing did not produce learning benefits for key-term definitions is an efficiency difference during practice—namely, learners working collaboratively accomplished fewer trials within the allotted amount of time.

In addition to examining the efficiency costs and learning benefits as separate components of engaging in collaborative testing for key-term definitions, we calculated a combined measure to examine the incremental gain in learning for each unit of time spent during retrieval practice (for a similar approach, see Rawson & Dunlosky, 2013; Zamary & Rawson, 2017). For each participant, we divided their cued recall score on the final test by the total number of minutes spent during the retrieval practice phase to create a gain per minute (GPM) measure. For this aggregate measure of costs and benefits, GPM was greater for the individual group versus the collaborative group, $t(79) = 2.44$, $p = .017$, $d = .54$ (see Table 1).

Although the primary interest of the current research was to evaluate efficiency costs and learning benefits of collaborative testing for key-term definitions, we also examined potential costs

and benefits in metacognitive monitoring and control for exploratory purposes. Table 2 provides information concerning metacognitive outcomes that occurred during practice for learners who engaged in collaborative versus individual testing. These outcomes provide novel evidence to the collaborative memory literature and to the broader literature on self-regulated learning, in that the current research is the first study to empirically examine metacognitive components during collaborative testing. Of note, for trials on which learners judged an item as correct, accuracy was 65% for the collaborative group and 76% for the individual group, $t(78) = 2.54$, $p < .013$, $d = .57$. This outcome suggests that learners working collaboratively versus individually were less accurate in judging the correctness of their responses. Thus, another potential cost to collaborative testing for key-term definitions is suboptimal metacognitive monitoring.

Experiment 2

Experiment 1 showed that when given a fixed amount of time, students achieve a higher level of learning when engaging in retrieval practice for key-term definitions individually versus collaboratively. Furthermore, this learning benefit primarily resulted from more efficient practice (i.e., the opportunity to complete a greater number of trials and to experience a higher level of retrieval success during practice) for learners working individually versus collaboratively. The primary purpose of Experiment 2 was to evaluate the effectiveness of engaging in collaborative testing for key-term definitions when learners are given a fixed number of practice trials. Experiment 2 was the same as Experiment 1 except that instead of Phase 2 being 30 minutes, learners were given a total of 30 trials to complete during Phase 2.

Table 1 Gain per minute outcomes

	Individual	Collaborative
Experiment 1	2.1 (.1)	1.6 (.1)
Experiment 2	2.9 (.3)	2.0 (.2)
Experiment 3	3.5 (.3)	2.2 (.2)

Note. *Individual* refers to learners who completed Phase 2 individually and *Collaborative* refers to learners who completed Phase 2 in a dyad. Mean values reflect percentage points gained on final cued recall test per minute spent during the retrieval practice phase. Standard errors of the mean are reported in parentheses

Method

Participants and design Undergraduates who participated for course credit ($n = 74$) were assigned to one of two groups: collaborative or individual. One participant was excluded due to noncompliance, three participants were excluded due to attrition, and three participants were excluded due to computer errors. The final sample for the collaborative and the individual group was $n = 28$ and $n = 39$, respectively. The a priori targeted sample size for Experiment 2 was $n = 80$, as in Experiment 1. A majority of the sample were freshman (40%) and female (67%), with a mean grade-point average of 3.2 and an average age of 20 years.

Materials and procedure Materials were the same as in Experiment 1. The procedure was identical to Experiment 1, with three exceptions. First, all learners now completed a total of 30 trials during Phase 2. Learners in the individual group were tested on each of the 10 key terms three times. Learners in the collaborative group also completed a combined total of 30 trials, with each of the two learners being tested by their partner on five of the key terms three times (e.g., Participant A recalled “positive reinforcement” three times and Participant B recalled “negative reinforcement” three times). This design equated the number of test trials for a subset of items for learners in both the individual group

Table 2 Metacognitive outcomes of secondary interest for Phase 2 in Experiments 1–2

	Individual	Collaborative
Experiment 1		
Percent of trials judged correct	62%	54%
Percent of trials correct	32%	21%
Mean score on trials judged correct	76%	65%
Mean score on drop trials	70%	63%
Experiment 2		
Percent of trials judged correct	57%	66%
Percent of trials correct	20%	19%
Mean score on trials judged correct	61%	58%

Note. *Individual* refers to learners who completed Phase 2 individually and *Collaborative* refers to learners who completed Phase 2 in a dyad

and the collaborative group (i.e., three test trials for five of the key terms). Second, given that the number of trials was held constant, Phase 2 was now self-paced (instead of 30 minutes). Third, given that the number of trials was equated, participants no longer made a control decision about whether they wanted to drop the key term from practice.

Split-half reliability was acceptable for the cued recall test ($\alpha = .93$) and the multiple choice test ($\alpha = .72$).

Results and discussion

To facilitate examination of the costs (with respect to efficiency) and benefits (with respect to learning), outcomes of primary interest are shown in Fig. 2. For the collaborative group, outcomes on the final tests are reported separately for tested items and nontested items. More specifically, tested refers to the items that collaborative learners were asked to recall, whereas nontested refers to the items that collaborative learners asked their partner to recall. Importantly, performance on the tested items for the collaborative group is directly comparable to the individual group given that each of these items received three test trials during Phase 2. As shown in the left panel of Fig. 2, final cued recall performance was the same for the collaborative group versus the individual group for both tested items and nontested items, $t(65) = 0.07$, $p = .949$, and $t(65) = 0.09$, $p = .931$, respectively. In addition, performance on the multiple-choice test was also the same for the collaborative group versus the individual group for both tested items and nontested items, $t(65) = 0.47$, $p = .638$, and $t(65) = 0.32$, $p = .749$, respectively (see the middle panel of Fig. 2). As in Experiment 1, the current results suggest that postcollaborative benefits do not emerge for key-term definitions when students are given a fixed number of practice trials, working collaboratively versus individually results in similar levels of performance.²

To diagnose potential costs of engaging in collaborative testing, we again examined what occurred during Phase 2 retrieval practice.³ As shown in the right panel of Fig. 2, the groups differed substantially in the amount of time it took learners to complete the 30 practice trials: learners in the individual group completed the 30 practice trials in 19 minutes, whereas learners in the collaborative group completed the 30 practice trials in 28 minutes, $t(65) = 5.00$, $p < .001$, $d = 1.24$. Thus, although learners in both groups showed similar levels of learning on the final test, it took learners working collaboratively significantly longer to achieve that level of performance. Replicating Experiment 1 for

² Although not of interest for current purposes, the similar recall of tested versus not-tested items for learners in the collaborative group is surprising given the robustness effects of retrieval practice (for reviews of testing effects, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rawson & Dunlosky, 2011; Roediger & Butler, 2011; Rowland, 2014). This was an unanticipated finding, which needs replication and further evaluation in future research.

³ There were no differences on the initial cued recall test during Phase 1 for the collaborative group ($M = .30$, $SE = .03$) versus the individual group ($M = .28$, $SE = .04$), $t(65) = 0.42$, $p = .679$.

Fig. 2 Performance on outcomes of primary interest in Experiment 2, for learners who completed Phase 2 collaboratively in a dyad or working individually. *Tested items* refers to the subset of items that each learner in the dyad was tested on during Phase 2. *Nontested items* refers to the

subset of items that each learner was not tested on during Phase 2 (i.e., the items that they tested their partner on). *Error bars* represent standard errors

the aggregate measure of costs and benefits, GPM was greater for the individual group versus the collaborative group, $t(65) = 2.38$, $p = .020$, $d = .59$ (see Table 1).

Of secondary interest, Table 2 provides information concerning metacognitive outcomes that occurred during practice for learners who engaged in collaborative versus individual testing. For trials on which learners judged an item as correct, accuracy was at 58% for the individual group and 61% for the collaborative group, $t(65) = .53$, $p = .598$.

Experiment 3

Experiment 2 showed that even when equating the number of practice trials, students do not show a postcollaborative benefit from engaging in retrieval practice for key-term definitions collaboratively versus individually. Experiment 2 also showed an efficiency cost such that it took learners working collaboratively significantly longer to complete the practice trials. Taken together, Experiments 1–2 suggest that collaborative testing of key terms does not afford learning benefits and comes with a concomitant efficiency cost under representative conditions.

Why might postcollaborative benefits not extend to key-term learning under these conditions? One theoretical account of postcollaborative benefits is *reexposure*, which refers to when one group member recalls an item that another group member may have otherwise forgotten, which in turn provides an opportunity for additional restudy and increases learning (Blumen & Rajaram, 2008; Weldon & Bellinger, 1997). In contrast to a majority of previous collaborative memory research, the current research included feedback during retrieval practice (i.e., Phase 2) given that 96% of students reported that feedback is provided during collaborative testing (Wissman & Rawson, 2016). More specifically, after recalling each key term, learners were

presented with the correct definition as feedback. Consequently, the opportunity to engage in restudy may supplant the benefits of reexposure in a real-world collaborative setting. To examine the extent to which benefits of collaborative testing may emerge in the absence of feedback, Experiment 3 eliminated feedback during retrieval practice. Eliminating feedback will afford evaluation of a theoretical explanation concerning why postcollaborative benefits may not extend to more representative conditions—namely, that feedback supplants the benefits of reexposure.

Method

Participants and design Undergraduates who participated for course credit ($n = 95$) were assigned to one of two groups: collaborative or individual. One participant was excluded due to noncompliance and 11 participants were excluded due to attrition. The final sample for the collaborative and the individual group was, $n = 43$ and $n = 40$, respectively. The a priori targeted sample size for Experiment 3 was $n = 80$, as in Experiments 1–2. A majority of the sample were freshman (43%) and female (75%), with a mean grade point average of 3.1 and an average age of 20 years.

Materials and procedures Materials were the same as in Experiments 1–2. The procedure was identical to Experiment 2 with three exceptions. First, participants were no longer provided with the correct answer during Phase 2. Second, given that feedback was eliminated, participants no longer made a monitoring decision about the accuracy of their response. Third, participants in the collaborative group had the opportunity to modify their answers on test trials. After typing in their response the computer screen advanced to a new screen that had a text field with their original response along with the instructions to discuss and

modify their response however they saw fit. Allowing collaborative participants to modify their responses afforded examination of any change in recall after an opportunity to discuss an answer with a partner.⁴

Split-half reliability was acceptable for the cued recall test ($\alpha = .94$) and the multiple choice test ($\alpha = .71$).

Results and discussion

Following from the results of Experiments 1–2, outcomes of primary interests concerning costs and benefits of collaborative testing are shown in Fig. 3. Similar to Experiment 2, outcomes on the final tests are reported separately for the tested items and nontested items. In contrast to Experiments 1–2, outcomes on the initial cued recall test during Phase 1 showed differences for the collaborative group ($M = .43$, $SE = .03$) versus the individual group ($M = .34$, $SE = .02$), $t(81) = 2.14$, $p = .036$. Accordingly, we conducted ANCOVAs, with initial cued recall as a covariate for final test measures, and report the corresponding adjusted mean values below.⁵

As shown in the left panel of Fig. 3, final cued recall performance was numerically greater for the collaborative group versus the individual group for both tested items and nontested items, $F(1, 80) = 7.03$, $MSE = .020$, $p = .010$, $\eta_p^2 = .081$, and $F(1, 80) = 1.80$, $MSE = .020$, $p = .184$, $\eta_p^2 = .022$, respectively. As shown in the middle panel of Fig. 3, performance on the multiple-choice test was the same for the collaborative group versus the individual group for tested items, $F(1, 80) = 0.04$, $MSE = .016$, $p = .950$, $\eta_p^2 < .001$, and greater for the collaborative group versus the individual group for nontested items, $F(1, 80) = 11.92$, $MSE = .015$, $p < .001$, $\eta_p^2 = .130$. These outcomes suggest a theoretical explanation for why postcollaborative benefits were not observed in Experiments 1–2—namely, that feedback may have supplanted the benefits of reexposure.

To diagnose costs of engaging in collaborative testing, we again examined the time it took learners to complete the 30 practice trials during Phase 2 retrieval practice. Replicating Experiment 2, the collaborative group required significantly more time than the individual group, $t(81) = 6.48$, $p < .001$, $d = 1.42$ (see right panel of Fig. 3). Replicating Experiments 1–2, GPM was greater for the individual group versus the collaborative group, $F(1, 80) = 23.32$, $MSE = .000$, $p < .001$, $\eta_p^2 = .226$ (see Table 1). Collectively, results

suggest that collaborative testing for key-term definitions can lead to learning benefits when feedback is eliminated but is still associated with an efficiency cost with respect to time.

General discussion

The current research evaluated the effectiveness of engaging in collaborative testing for key-term definitions under representative conditions, in terms of both efficiency costs and learning benefits. Experiment 1 controlled the amount of time learners had to complete practice to examine efficiency differences in the number of trials learners were able to complete during practice, and learners working collaboratively versus individually completed fewer trials. Experiment 2 controlled the number of trials learners completed during practice to examine efficiency differences in the amount of time required to complete practice, and learners working collaboratively versus individually took substantially longer. Furthermore, neither experiment demonstrated postcollaborative benefits for learning. In sum, results indicate that collaborative testing (vs. individual testing) is less effective overall when learning key-term definitions under conditions that are representative of what students report implementing, given that the efficiency costs of collaboration are not offset by concomitant benefits to learning.

Although the primary goal of the current research was to evaluate costs and benefits of collaborative testing under representative conditions, Experiment 3 suggests that how collaborative testing is implemented may influence the extent to which postcollaborative benefits are observed. Indeed, when feedback was eliminated (which aligns with the methodology typically used in prior research), postcollaborative benefits emerged. As previously discussed, reexposure refers to when one group member recalls an item that another group member may have otherwise forgotten (Blumen & Rajaram, 2008; Weldon & Bellinger, 1997). Consistent with empirical demonstrations of reexposure, survey research has shown that 67% of students report engaging in discussion during collaborative testing (Wissman & Rawson, 2016). Thus, one plausible explanation for why postcollaborative benefits emerged in Experiment 3 is that eliminating feedback provided learners working collaboratively with the opportunity to benefit from reexposure.

Given that the goal of the current research was to examine the costs and benefits of collaborative testing under representative conditions, several of the methodological dimensions instantiated here extend beyond those used in prior collaborative research. For example, learners in the current research alternated being the asker versus responder (instead of recalling key terms together) because a vast majority of students report implementing a turn-taking approach (Wissman & Rawson, 2016). In addition, learners here were provided with feedback during retrieval practice because survey outcomes indicated that 96% of students reported that the correct answer was provided as feedback during retrieval practice (Wissman & Rawson, 2016). Some of the

⁴ Outcomes showed no differences in recall from the first test trial ($M = .61$, $SE = .04$) to the modified test trial ($M = .62$, $SE = .03$) for the collaborative group, $t(42) = 1.15$, $p = .256$. Thus, these outcomes are not discussed further.

⁵ Although outcomes from Experiments 1–2 showed no differences on initial cued recall for the collaborative versus individual group, employing a parallel analytic approach affords a clearer comparison across experiments. Notably, the same pattern of results emerged when ANCOVAs with initial cued recall as a covariate were conducted for cued recall and multiple test performance in Experiment 1, $F(1, 78) = 5.88$, $MSE = .039$, $p = .018$, $\eta_p^2 = .070$, and $F(1, 78) = 1.18$, $MSE = .020$, $p = .282$, $\eta_p^2 = .015$; and Experiment 2, tested items: $F(1, 64) = 0.12$, $MSE = .033$, $p = .735$, $\eta_p^2 = .002$, and $F(1, 64) = 0.08$, $MSE = .022$, $p = .775$, $\eta_p^2 = .001$; nontested items: $F(1, 64) = 0.29$, $MSE = .028$, $p = .590$, $\eta_p^2 = .005$, and $F(1, 64) = 0.03$, $MSE = .023$, $p = .870$, $\eta_p^2 < .001$.

Fig. 3 Performance on outcomes of primary interest in Experiment 3, for learners who completed Phase 2 collaboratively in a dyad or working individually. *Tested items* refers to the subset of items that each learner in the dyad was tested on during Phase 2. *Nontested items* refers to the

subset of items that each learner was not tested on during Phase 2 (i.e., the items that they tested their partner on). *Error bars* represent standard errors

methodological differences used here, compared to previous research, may affect subsequent recall, whereas some may not. For example, Experiment 3 demonstrates that feedback likely influences the memorial consequences of collaborative testing, in that postcollaborative benefits emerged when feedback during retrieval practice was eliminated. In contrast, recent research conducted in our laboratory suggests that having learners recall key-term definitions together (vs. turn taking) may not impact subsequent recall. When learners jointly recalled key terms during retrieval practice, postcollaborative benefits did not emerge on final tests (Wissman, manuscript in preparation). Interestingly, Experiments 1–2 reported here suggest the way in which learners implement collaborative testing in a real-world setting may be disadvantageous, with Experiment 3 further implicating feedback as a potential boundary condition to postcollaborative benefits. Further investigation of factors that affect whether and when postcollaborative benefits obtain will further inform both theory and application.

In addition to investigating underlying mechanisms of postcollaborative benefits more specifically, future research should also evaluate the efficiency of engaging in collaborative testing more generally. Research investigating different learning techniques often focuses on learning benefits (with respect to effectiveness), whereas potential costs (with respect to efficiency) are often overlooked. Outcomes from the current research suggest that engaging in collaborative testing for key-term definitions yields an efficiency cost in terms of the number of trials learners are able to complete in a set amount of time (Experiment 1) and in the amount of time it takes learners to complete a set amount of trials (Experiment 2), with no concomitant benefit to learning. Given that students only have a limited amount of time to learn a substantial amount of information, evaluating the efficiency of learning techniques is important for applied purposes.

Although not of primary interest for present purposes, the current research is the first study to empirically examine metacognitive monitoring and control decisions during

collaborative testing. Current theories of self-regulated learning assume that learners' control decisions are informed by their monitoring judgments (for a review, see Dunlosky & Ariel, 2011), with prior research showing that learners make suboptimal monitoring decisions that leads to poor control decisions when working individually (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Dunlosky & Lipko, 2007). Importantly, results from the current research established that suboptimal metacognitive accuracy and control decisions also occur for learners working collaboratively. These outcomes provide novel evidence to the collaborative memory literature and to the broader literature on self-regulated learning, suggesting that two heads are not better than one when monitoring and controlling retrieval practice.

To conclude, the current research demonstrated that engaging in collaborative testing for key-term definitions is less efficient during practice and does not yield memorial benefits on a subsequent individual test under representative conditions. Given that the recent research has emphasized the importance of independent replication for novel findings (e.g., Braver, Thoenes, & Rosenthal, 2014; Cumming, 2008; Ferguson & Heene, 2012; Francis, 2012; LeBel & Peters, 2011; Maner, 2014; Pashler & Harris, 2012; Roediger, 2012; Schimmack, 2012; Schmidt, 2009; Simmons, Nelson, & Simonsohn, 2011; Simons, 2014), future research should replicate and further explore the effects observed here. The current work provides an important foundation for further evaluating the costs and benefits of collaborative testing for key-term definitions in real-world contexts, which will have important practical implications and will provide outcomes that further guide theory development in the burgeoning literature on collaborative testing.

Acknowledgements The research reported here was supported by a Collaborative Award from the James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior.

Appendix. Material set used in Experiments 1–3

Key terms and definitions

Positive reinforcement: When something pleasurable is added or experienced after a behavior occurs, the behavior is more likely to occur again in the future

Negative reinforcement: When something unpleasant is removed or escape or avoidance is allowed after a behavior occurs, the behavior is more likely to occur again in the future

Positive punishment: When an aversive event or stimulus is presented after a behavior occurs, the behavior is less likely to occur again in the future

Negative punishment: When something pleasurable is diminished or taken away after a behavior occurs, the behavior is less likely to occur again in the future

Interval schedule: Reinforcement only occurs if responses are made after a certain amount of time has elapsed

Ratio schedule: Reinforcement depends only on the number of responses

Continuous schedule: Reinforcement takes place after each and every correct response and leads to rapid learning

Satiation: When a reinforcer loses effectiveness from repeated use and leads to a decrease in the behavior

Extinction: When a previously reinforced behavior no longer receives reinforcement, the likelihood of the behavior eventually decreases

Partial schedule: Responses are reinforced sometimes but not every time, which leads to behaviors that are very resistant to extinction

References

- Blumen, H. M., & Rajaram, S. (2008). Influence of re-exposure and retrieval disruption during group collaboration on later individual recall. *Memory, 16*, 231–244.
- Blumen, H. M., & Rajaram, S. (2009). Effects of repeated collaborative retrieval on individual memory vary as a function of recall versus recognition tasks. *Memory, 17*, 840–846.
- 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.
- Braver, S. L., Thoenes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342.
- Congleton, A. R., & Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General, 140*, 535–551.
- Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*, 286–300.
- Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed.), *Psychology of Learning and Motivation, 54*, 103–140.
- Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology, 64*, 467–484.
- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228–232.
- 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*, 4–58.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561.
- Francis, G. (2012). The psychology of replication and replication of psychology. *Perspectives on Psychological Science, 7*, 585–594.
- LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*, 371–379.
- Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*, 343–351.
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.
- Rawson, K. A., & Dunlosky, J. D. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*, 283–302.
- Rawson, K. A., & Dunlosky, J. D. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General, 142*, 1113–1129.
- Roediger, H. L., III (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer, 25*, 27–29.
- Roediger, H. L., III, & Butler, A. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463.
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566.
- Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social science. *Review of General Psychology, 13*, 90–100.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9, 76–80.
- Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1160–1175.
- Wissman, K. T., & Rawson, K. A. (2016). How students implement collaborative testing in real-world contexts. *Memory*, 24, 223–239.
- Zamary, A., & Rawson, K.A. (2017). Which technique is most effective for learning declarative concepts-provided examples, generated examples, or both? *Educational Psychology Review*.