Journal of Experimental Psychology: Applied
2018, Vol. 24, No. 3, 347–359

© 2017 American Psychological Association
1076-898X/18/$12.00 http://dx.doi.org/10.1037/xap0000148

The Testing Effect in a Social Setting: Does Retrieval Practice Benefit a Listener?

Magdalena Abel and Henry L. Roediger III
Washington University in St. Louis

Retrieval practice boosts retention relative to other study strategies like restudying, a finding known as the *testing effect*. In 3 experiments, the authors investigated testing in social contexts. Subjects participated in pairs and engaged in restudy and retrieval practice of vocabulary pairs. During retrieval practice, 1 subject acted as speaker (overt practice); the other subject listened and monitored the speaker's responses (covert practice). All experiments showed testing effects, with overt practice by speakers enhancing recall relative to restudy after a 2-day delay. In Experiments 1 and 2, covert practice by listeners did not benefit recall as much as overt practice. Only in Experiment 3, when listeners were asked to monitor their own covert retrieval (instead of the speaker's overt retrieval), did both types of practice convey similar benefits. The results indicate that memory retrieval is not necessarily as beneficial for listeners as for speakers. The practical implication is that the practice of teachers asking questions in class will not yield a positive effect unless special measures are taken to insure students' effortful covert retrieval.

Keywords: retrieval practice, testing effect, social context, delay

Retrieval practice has beneficial effects on both recall and recognition, as demonstrated by a large volume of research on the testing effect. Retrieval practice has repeatedly been shown to boost performance, especially on delayed tests (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Pyc & Rawson, 2010; Roediger & Karpicke, 2006; for a brief review, see Roediger & Butler, 2011). Further research indicates that this benefit not only arises when retrieval practice is compared to passive restudy opportunities, but also when it is compared to elaborative study techniques instead (concept mapping or the keyword-technique; see Karpicke & Blunt, 2011; Karpicke & Smith, 2012). Moreover, testing effects have also been shown to occur in classrooms (e.g., McDaniel, Roediger, & McDermott, 2007; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014), encouraging its use in applied educational contexts to improve learning.

One neglected factor in this body of research is social context. Most studies on the testing effect have examined individual subjects engaging in retrieval practice (or other forms of practice) in isolation. However, retrieval practice often occurs in social settings. Students often study in groups, asking each other questions. In classrooms, teachers often ask questions for students to answer during class or encourage students to work in groups to solve a problem. Previous research on collaborative memory retrieval indicates that working with others during remembering may come with certain costs for recall during collaborative retrieval, but also with certain benefits for later individual memory (see Rajaram, 2011; Rajaram & Pereira-Pasarin, 2010). The question that the current research addresses is whether retrieval practice carried out by an individual in the company of others may benefit not only that individual but also others who are listening to the question and answer. For example, in a classroom discussion a teacher may ask a question and only one student will answer while the rest of class is supposed to "think along" when listening to the student's answer. The hope is that the other students will be retrieving the answer covertly, or trying to. But does being exposed to another person's memory retrieval actually provide similar benefits for listeners? The combined findings of two separate lines of research suggest that this might be the case, and we review them in turn.

When the teacher asks a question in a classroom and one student answers, the other students may be answering the question themselves covertly (or at least trying to). Two sets of experiments indicate that covert retrieval (i.e., thinking of an answer but not saying or writing it) can lead to robust testing effects, at least in individual recall. Smith, Roediger, and Karpicke (2013) asked subjects to engage in retrieval practice and either to type in their answers (overt practice) or to keep on thinking about their answers (covert practice). Across four experiments, performance on a final test was equivalent for such overt and covert forms of retrieval practice. The finding emerged after relatively short delay intervals of 15 min, but it also appeared after 2 days. A direct comparison of forgetting rates across delays also indicated comparable forgetting after overt and covert practice. Consistent with prior work, forgetting after both forms of retrieval practice was attenuated

This article was published Online First November 27, 2017.

Magdalena Abel and Henry L. Roediger III, Department of Psychology, Washington University in St. Louis.

This work was supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD) and by a collaborative activity grant from the James S. McDonnell Foundation. Thanks to Deniz Ariturk and Tanya Nungess for help with data collection, and to Andy DeSoto, Jason Finley, John Nestojko, Adam Putnam, and Victor Sungkhasettee for helpful discussions and comments.

Correspondence concerning this article should be addressed to Magdalena Abel, who is now at Department of Experimental Psychology, Regensburg University, 93040 Regensburg, Germany. E-mail: magdalena.abel@ur.de

compared to a restudy condition. Similar results were reported by Putnam and Roediger (2013). During retrieval practice, subjects in their experiments were asked to say their answers out loud, to write them down, or to think about them. Again, these different forms of retrieval practice were equally beneficial for recall on a final test, also relative to a restudy condition. Taken together, overt and covert practice seem to produce comparable testing effects, at least in individually working subjects (for further findings, see Carpenter et al., 2008; Izawa, 1976; but see also Jönsson, Kubik, Sundqvist, Todorov, & Jonsson, 2014; Tauber et al., in press).

A second line of relevant research is by Hirst and his colleagues, who examined selective retrieval practice in a social context and focused on the resulting negative effect for information that is not practiced (i.e., retrieval-induced forgetting; see Anderson, Bjork, & Bjork, 1994). Cuc, Koppel, and Hirst (2007) tested pairs of subjects and asked one of the two (the speaker) to retrieve a subset of previously studied information out loud while the second subject (the listener) was asked to monitor the speaker's recall for either accuracy or fluidity. On a final test for all initially studied information, Cuc et al. observed retrieval-induced forgetting for the nonpracticed information (relative to a control condition) not only for the speakers, but also for the listeners (for similar findings, see Abel & Bäuml, 2015; Coman, Manier, & Hirst, 2009; Stone, Barnier, Sutton, & Hirst, 2013; for a review, see Hirst & Echterhoff, 2012). Importantly, however, Cuc et al. (2007) only found socially shared retrieval-induced forgetting in listeners when they were asked to monitor the speaker's responses for accuracy. When listeners were instead asked to monitor for fluidity or smoothness of recall, no socially shared forgetting emerged. Based on this finding, Cuc et al. argued that the effects of selective retrieval practice can be socially shared between speakers and listeners, but only when listeners are motivated to engage in covert retrieval practice along with speakers. Coman and Hirst (2015) arrived at the same conclusion in research using a somewhat different version of the paradigm.

Yet, to date, researchers have not directly examined whether benefits from retrieval practice can be socially shared and "transmitted" from speakers to listeners when a typical testing-effect paradigm is used, with a restudy baseline to evaluate the benefits of retrieval practice more closely (for other related work, however, see Congleton & Rajaram, 2011; Wissman & Rawson, 2016). In retrieval-induced forgetting experiments, some material is practiced and those data can be examined, even though they are usually not compared to a restudy baseline and also not of primary interest in this context. Mean performance for practiced items was reported by Cuc et al. (2007), along with analyses showing that practice effects were present (relative to a baseline condition without any practice) for both speakers and listeners. However, even though practice effects occurred in listeners, in two of the three experiments reported by Cuc et al., practiced items were remembered better by speakers than by listeners. Other studies using the same speaker-listener paradigm reported mixed findings as well (e.g., Brown, Kramer, Romano, & Hirst, 2012; Koppel, Wohl, Meksin, & Hirst, 2014; Stone, Barnier, Sutton, & Hirst, 2010). Thus, prior research does not allow firm conclusions on whether retrieval practice affords a testing effect for listeners as well as speakers in social retrieval-induced forgetting experiments.

Because prior research is not definitive, we embarked on three experiments specifically designed to ask if testing effects arise relative to a restudy baseline in social settings and whether they are of the same magnitude for speakers and listeners. All experiments applied a typical testing-effect design but were additionally modeled after the prior work on socially shared retrieval-induced forgetting (e.g., Cuc et al., 2007). Pairs of subjects studied Swahili-English vocabulary pairs and then engaged in restudy of some of the material and in retrieval practice of the rest. In addition, one subject was asked to act as the speaker and to practice retrieval aloud while the other subject was asked to listen and to monitor the first person's responses for either accuracy of the response or its fluidity and smoothness of production. A final test on all word pairs (restudied, overtly retrieved, covertly monitored) was completed after 3 min or 2 days. We expected to observe regular testing effects (i.e., better recall after retrieval practice compared to restudy, especially with longer delays; see Roediger & Karpicke, 2006). In addition, based on the study by Cuc et al. (2007), we expected only accuracy (and not fluidity) monitoring to trigger covert retrieval practice. Based on the findings of comparable testing effects after overt and covert retrieval practice in individuals (Putnam & Roediger, 2013; Smith et al., 2013), monitoring another person's answers for accuracy was predicted to be as beneficial for later retention as overt retrieval practice.

Experiment 1

Method

Participants. Sample sizes in all reported experiments were determined based on prior work on the testing effect (e.g., Roediger & Karpicke, 2006), while at the same time allowing for full counterbalancing of the stimulus materials. One hundred forty-four students at Washington University in St. Louis were recruited for Experiment 1. Subjects participated in pairs and were evenly distributed across one of four conditions (n = 36/condition). Allocation of pairs of subjects to conditions was quasi-random, with alternating allocations to the different delay and monitoring conditions. Mean age was 19.7 years (SD = 1.4 years). Subjects received course credit or $10 for participation. The study protocol was approved by the local institutional review board (IRB).

Material. Thirty Swahili-English word pairs of medium difficulty (e.g., chaza–oyster, hadithi–story, mfupa–bone) were selected from the Nelson and Dunlosky (1994) norms and were divided into three sets of 10 word pairs each, with sets roughly equated for difficulty. Across subjects, each set was equally often used as stimulus material in retrieval-practice, restudy, and monitoring-task conditions, thus counterbalancing materials across conditions.

Design. The experiment employed a 3 × 2 × 2 mixed-factorial design. The first factor was the *practice condition* (retrieval practice, restudy, monitoring) and was manipulated within subjects. After initial study, all participants engaged in retrieval practice for one third of the material (i.e., 10 pairs), in restudy for another third, and in a monitoring task for the last third. The *type of monitoring* factor (accuracy monitoring, fluidity monitoring) was manipulated between subjects. When engaging in the monitoring task, half of all participants judged the accuracy of the other subject's answers during retrieval practice; the other half monitored the other person's retrieval with regard to fluidity and smoothness (which is supposed to cause a more superficial form of monitoring, not involving covert retrieval; for details, see Cuc et al., 2007). Within pairs, the two subjects were always asked to engage in the same type of monitoring task. Finally, *retention interval* (3 min, 2 days) was also manipulated between subjects. A final test on all word pairs was given after 3 min or 2 days.

Procedure.

Study phase. In the initial study phase, the 30 word pairs were presented in a random sequence, for 5 s each, centrally on a computer screen. Two subjects were tested together and sat silently in front of the same computer screen when studying the list; they were asked to try to memorize all word pairs for a final test.

Practice phase. After initial study, subjects were informed that all word pairs would be practiced in three separate blocks (with one third of word pairs practiced in each block). There were three practice cycles on each block, so that each vocabulary pair was repeated three times throughout the course of the respective practice block. Sequencing of word pairs during each practice cycle was random; after all 10 word pairs belonging to a certain practice block had been practiced, a new practice cycle began, with a new random sequence.

Practice blocks differed in the type of practice in which subjects were asked to engage. In the restudy block, word pairs were presented in intact form on the screen for 7 s each; both subjects were asked to silently restudy the word pairs and to make use of the additional study time. In the other two practice blocks, however, one of the two subjects was asked to engage in overt retrieval practice while the other participant was asked to engage in a monitoring task; thus, the block that constituted the active retrieval-practice condition for one subject simultaneously constituted the monitoring condition for the other subject. In these two blocks, the Swahili words were presented on the screen for 5 s each and in random order; one of the subjects was asked to act as the speaker and to try to recall the English meaning out loud. Selection of subjects as speakers and listeners was counterbalanced. For half of all subject pairs, the person sitting on the left side of the screen was asked to be the speaker on the first block that involved retrieval practice (the person on the right was asked to be the listener and to engage in the monitoring task). For the other half of subject pairs, the person sitting on the right was asked to be the speaker on the first block involving retrieval practice (the person on the left acted as the listener). Roles were always switched on the second block with retrieval practice (so that the person who acted as speaker on the first block now acted as listener, and vice versa). Corrective feedback was presented for 2 s after each 5-s retrieval-practice trial, thus equating the overall time of processing with the restudy condition. In particular, corrective feedback was presented in order to avoid social contagion with incorrect answers (e.g., Meade & Roediger, 2002; Roediger, Meade, & Bergman, 2001) and to increase potential testing effects (e.g., Arnold & McDermott, 2013; Kang, McDermott, & Roediger, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005; see also Rowland, 2014). The subject who was not asked to engage in overt retrieval practice was instead asked to listen to the other participant and to monitor his or her retrieval. In the accuracy monitoring condition, listeners were asked to indicate on a 7-point scale, ranging from 1 (*not accurate at all*) to 7 (*very accurate*) how accurate the other subject's answer was, separately for each retrieval-practice trial. In the fluidity monitoring condition, listeners were asked to judge on a similar 7-point scale how fluid and smooth the speaker's retrieval was, ranging from 1 (*not fluid at all*) to 7 (*very fluid*), again separately for each trial. All monitoring judgments were written, with listeners choosing response options on prepared response sheets; speakers were not informed about the listeners' choices. When engaging in monitoring, subjects were asked to make their judgments before the corrective feedback was presented on the screen; judging from the experimenter's perspective, subjects were able to do this on most practice trials. The sequence of practice blocks was counterbalanced across subjects, just as sets of stimulus materials were equally often assigned to each type of practice across subjects. After completing the last practice block, all subjects solved simple arithmetic equations for 3 min as a distractor task.

Final test phase. On the final test, subjects worked on separate computers. Subjects in the short-delay condition completed the test after the 3-min distractor task; subjects in the long-delay condition left the lab and returned to take the same test after 2 days. For the test, the Swahili words of all 30 word pairs were presented in random order for 10 s each on the screen, and subjects were asked to write down the response terms on a piece of paper. After completing the test, subjects were debriefed and thanked for their participation.

Results

Success rates on retrieval-practice cycles. Figure 1 shows mean retrieval success, separately for the three retrieval-practice cycles and as a function of (delay and monitoring) conditions. A 3 × 2 × 2 analysis of variance (ANOVA) with the within-participants factor of retrieval-practice cycle (first, second, third) and the between-participants factors of monitoring task (accuracy monitoring, fluidity monitoring) and delay (3 min, 2 days) revealed a significant main effect for the factor retrieval-practice cycle, $F(2, 280) = 741.60$, $MSE = 160.85$, $p < .001$, $\eta^2 = .84$. Corrective feedback improved recall from the first to the second (16.7% vs. 52.5%), $t(143) = 24.65$, $p < .001$, $d = 2.08$), and from the second to the third retrieval-practice cycle (52.5% vs. 70.7%), $t(143) = 15.14$, $p < .001$, $d = 1.26$. No other main effects or interactions reached significance, all *F*s <1.0, which confirms that success rates on the three retrieval-practice cycles did not differ between conditions (which had yet to be instantiated).

Accuracy monitoring performance. To see how accurate listeners were when monitoring for the other subject's accuracy,

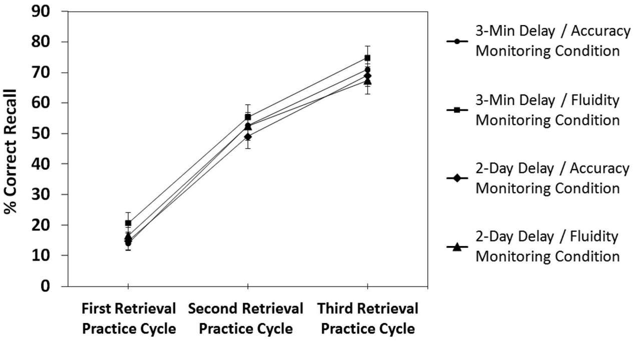

Figure 1. Mean recall performance on the three retrieval-practice cycles in Experiment 1, shown separately for the four conditions (differing in whether the final test was later administered after a 3-min or 2-day delay, and in whether retrieval-practice performance of the other subjects was monitored for accuracy or fluidity).

we analyzed their rating performance (see Table 1 for both monitoring and retrieval-practice performance in the accuracy monitoring condition). We coded the percentage of all 10 trials on each retrieval-practice cycle in which subjects correctly endorsed the other participant's correct answers (i.e., when they rated the other participant's correct response with >4 on the 7-point scale). A 3 × 2 ANOVA with the within-participants factor of retrieval-practice cycle (first, second, third) and the between-participants factor of delay (3 min, 2 days) revealed a significant main effect for the factor retrieval-practice cycle, $F(2, 140) = 261.73$, $MSE = 234.80$, $p < .001$, $\eta^2 = .79$. Similar to accuracy in retrieval practice of speakers, accuracy monitoring by listeners improved from the first to the second (9.3% vs. 41.8%), $t(71) = 13.61$, $p < .001$, $d = 1.60$, and from the second to the third retrieval-practice cycle (41.8% vs. 61.3%), $t(71) = 11.59$, $p < .001$, $d = 1.37$. No other main effects or interactions reached significance, all Fs < 1.0, showing that there were no differences between delay conditions.[1]

Recall on the final test. Figure 2 provides mean recall on the final test, where it can be seen that a different pattern occurred on the immediate and delayed tests. No testing effect occurred on the immediate test but the effect did appear on the delayed test. Items in the monitoring condition were recalled least well on both tests. A 3 × 2 × 2 ANOVA revealed a significant main effect of practice, $F(2, 280) = 53.88$, $MSE = 231.35$, $p < .001$, $\eta^2 = .28$, indicating that recall was differently affected by retrieval practice, restudy, and the monitoring task. In addition, the ANOVA showed a significant main effect of delay, $F(1, 140) = 94.21$, $MSE = 1238.61$, $p < .001$, $\eta^2 = .40$, reflecting time-dependent forgetting. More importantly, we also found a significant Practice × Delay interaction, $F(2, 280) = 8.75$, $MSE = 231.35$, $p < .001$, $\eta^2 = .06$, which indicates that forgetting across the 2-day delay differed in the three practice conditions. Surprisingly, the ANOVA showed that performance on the final test was in no way affected by type of monitoring task (all Fs < 1.0, all ps $\geq .429$). Apparently, it made no difference whether participants engaged in accuracy or fluidity monitoring. Therefore, data sets were collapsed across monitoring conditions for all further analyses.

To consider forgetting in the various practice conditions across the 2-day delay, further 2 × 2 ANOVAs were carried out. A comparison of recall after active retrieval practice and restudy again revealed a significant interaction, $F(1, 142) = 13.65$, $MSE = 224.79$, $p < .001$, $\eta^2 = .09$, reflecting differences in forgetting. Although the two practice types resulted in comparable recall after 3 min (70.7% vs. 71.8%), $t(71) < 1.0$, $p = .652$, $d = 0.05$, a typical testing effect was obtained after two days—retrieval practice led to greater recall than restudy (42.2% vs. 30.3%), $t(71) = 4.69$, $p < .001$, $d = 0.57$. A comparison of recall after restudy and monitoring across delays also revealed a significant interaction, $F(1, 142) = 11.61$, $MSE = 258.58$, $p = .001$, $\eta^2 = .08$. Although recall after the short delay was clearly superior after engaging in restudy compared to monitoring (71.8% vs. 52.6%), $t(71) = 7.00$, $p < .001$, $d = 0.83$, this difference was much smaller after the 2-day delay (30.3% vs. 24.0%), $t(71) = 2.39$, $p = .020$, $d = 0.28$, indicating that the monitoring task also reduced time-dependent forgetting relative to the restudy condition. Critically, a comparison of recall after monitoring and retrieval practice across delays showed no significant interaction and confirmed that time-dependent forgetting was comparable in the two monitoring conditions, $F(1, 142) < 1.0$, $p = .967$. Nevertheless, performance was roughly 18% worse after monitoring as compared to overt retrieval practice—after both the short delay (70.7% vs. 52.6%), $t(71) = 8.05$, $p < .001$, $d = 0.97$, and the long delay (42.2% vs. 24.0%), $t(71) = 7.24$, $p < .001$, $d = 0.86$. Although recall did certainly not benefit from monitoring to the same degree as it did from overt retrieval practice, monitoring triggered at least some degree of covert retrieval practice, which reduced time-dependent forgetting as much as overt retrieval practice did.

Discussion

The results in the retrieval practice and restudy conditions of Experiment 1 replicate prior work on the testing effect; retrieval practice compared to restudy reduced time-dependent forgetting and increased recall after a longer delay (e.g., Roediger & Karpicke, 2006). Thus, overt retrieval practice carried out in front of others boosts retention. In contrast, the results in the monitoring conditions were surprising, at least on two fronts. First, from prior work on socially shared retrieval-induced forgetting (Cuc et al., 2007), we had expected that accuracy monitoring but not fluidity monitoring would trigger covert retrieval practice. In the present data, however, no differences between monitoring tasks occurred. Second, based on previous work showing that testing effects are comparable for overt and covert forms of retrieval practice (see Putnam & Roediger, 2013; Smith et al., 2013), we had expected to observe a similar pattern of results after (overt) retrieval practice and (covert) accuracy monitoring. Yet, the data show that overt practice was much more effective than the accuracy monitoring task on both immediate and delayed tests. Although the same reduction in time-dependent forgetting occurred in the two practice conditions, monitoring another person's retrieval practice did not entail the same mnemonic benefit as engaging in overt retrieval practice. Success rates on the first retrieval practice cycle showed that recall ranged around 17% correct after one study, so all three types of practice increased performance relative to this baseline. However, overt retrieval practice was not just more beneficial for long-term retention than restudy; it also increased recall relative to the monitoring tasks (and thus, covert retrieval).[2]

[1] An additional ANOVA directly comparing retrieval-practice success by speakers and monitoring performance by listeners only revealed a significant main effect, $F(1, 70) = 8.27$, $MSE = 790.36$, $p = .005$, $\eta^2 = .11$, reflecting somewhat better recall by speakers than monitoring by listeners (45.2% vs. 37.5% across all retrieval-practice cycles). Thus listeners failed to recognize about 8% of the correct responses of speakers. Nevertheless, the lack of any interaction effects (all Fs ≤ 1.08, all ps $\geq .333$) indicates that monitoring performance followed the same learning function as retrieval-practice performance. Descriptive statistics for accuracy monitoring conditionalized on the speaker's correct recall show a similar pattern, with listeners rating 63.4% of the speaker's correct answers as accurate on the first retrieval-practice cycle, 77.4% on the second, and 85.6% on the third retrieval-practice cycle. Listeners, similarly to speakers, were better able to monitor accuracy with every additional retrieval-practice cycle. Additional information regarding errors that were committed during retrieval practice and accuracy monitoring can be found in the Appendix.

[2] Studies on social aspects of remembering are often discussed in terms of collective memory and the emergence of shared memories of (larger) groups of people (e.g., Hirst & Manier, 2008; Rajaram, 2011; Roediger & Abel, 2015). Indeed, the present study may speak to this issue, too, by indicating that monitoring a speaker's retrieval can also reinforce the respective contents in a listener's memory (see also Cuc et al., 2007), although potentially not as efficiently as overt retrieval practice.

Table 1

Mean Recall and Monitoring Performance in the Accuracy Monitoring Conditions of Experiment 1 are Displayed for Retrieval-Practice Cycles 1–3 and Separately for Short and Long Delay Conditions

Delay	Retrieval Practice Cycle 1	Retrieval Practice Cycle 2	Retrieval Practice Cycle 3
3-min delay			
Recall performance	14.2% (14.0)	52.8% (21.6)	71.1% (21.5)
Accuracy monitoring	10.3% (11.6)	43.9% (24.9)	64.7% (24.4)
2-day delay			
Recall performance	15.0% (17.5)	49.2% (23.5)	69.2% (22.3)
Accuracy monitoring	8.3% (13.6)	39.7% (26.0)	57.8% (28.0)

Note. Recall performance = mean retrieval success of speakers during the ten trials of each retrieval practice cycle; Accuracy monitoring = the percentage of all ten trials in which the speakers' correct responses were identified as accurate by the listeners. Values in parentheses represent standard deviations of the means.

Although we largely modeled our study after Cuc et al.'s (2007), procedural differences exist between experiments on retrieval-induced forgetting and ours on testing effects. For instance, Cuc et al. used semantically categorized item pairs (e.g., fruit–orange) and recall during retrieval practice was supported by relatively strong retrieval cues (e.g., fruit–or__). Although success rates for retrieval-practice cycles were not reported by Cuc et al., based on other similar studies (e.g., Abel & Bäuml, 2014; Anderson et al., 1994), we may assume that retrieval success was rather high from the first retrieval-practice cycle on, perhaps even close to ceiling. In contrast, in the present study, unrelated vocabulary pairs were used as study materials, and recall of target words was only cued with stimulus words during retrieval practice. Consequently, retrieval success on the first practice cycle was only about 20% correct. As in prior testing effect experiments, we provided corrective feedback, thereby increasing retrieval success to roughly 70% on the last retrieval-practice cycle. The greater difficulty of retrieval in the present experiment due to our more difficult materials together with the presentation of corrective feedback may have made listeners in both monitoring conditions aware that they had no secure grasp of the vocabulary pairs yet and should (covertly) practice them. This conclusion seems plausible though it is post hoc. To the best of our knowledge, the Cuc et al. (2007) study is to date the only one to compare accuracy and fluidity monitoring instructions and to report mnemonic differences caused by the two monitoring tasks. The present study indicates that these differences may be restricted to certain procedures and may not generalize across all learning scenarios.

The similar rates of time-dependent forgetting after the retrieval practice and the monitoring tasks indicate that monitoring may have stimulated at least some degree of covert retrieval practice in listeners, because less forgetting occurred in both these conditions than in the restudy condition. Yet, recall was clearly superior after overt retrieval practice than after such covert practice. Why did the monitoring tasks not benefit memory in the same way as overt retrieval practice, given the evidence that overt and covert forms of retrieval practice lead to largely the same benefits in subjects working alone (see Putnam & Roediger, 2013; Smith et al., 2013; but see too Tauber et al., in press)? One potential explanation is that instructions for accuracy monitoring were not explicit enough. Following Cuc et al. (2007), subjects were asked to judge if the speaker's answers were correct, but they were never explicitly instructed to also try to retrieve the correct answer during monitoring. Experiment 2 was conducted to determine if this same outcome would occur when subjects receive more explicit instructions to engage in covert retrieval practice during monitoring.

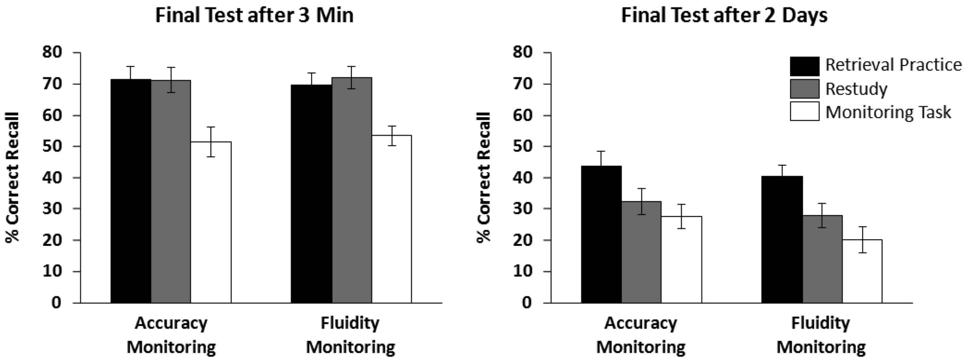

Figure 2. Mean recall performance on the final test in Experiment 1, shown as a function of delay (3-min delay, 2-day delay), type of practice (retrieval practice, restudy, monitoring task), and monitoring conditions (accuracy monitoring, fluidity monitoring). Error bars represent ±1 standard errors of the mean.

Experiment 2

Method

Participants. 72 undergraduates at Washington University in St. Louis participated in the study in pairs and were compensated with course credit or $10. Mean age was 20.4 years (*SD* = 1.9 years), and subjects were evenly distributed across the two delay conditions. The study protocol was approved by the local IRB.

Material. The same study materials were used as in Experiment 1.

Design. Because Experiment 1 showed no differences between accuracy and fluidity monitoring conditions and suggested that both types of monitoring led to a certain degree of covert retrieval, we included only accuracy monitoring in Experiment 2. Apart from this change, the design was identical to Experiment 1, resulting in a 3 × 2 mixed-factorial design with the two factors of *practice* (retrieval practice, restudy, monitoring task) and *delay* (3 min, 2 days).

Procedure. The procedure was identical to that of Experiment 1, with one exception. When instructing participants to engage in the accuracy monitoring task, we explicitly asked them to engage in the same task as the speaker, just silently. Listeners were asked to also make use of this covert form of retrieval practice and to judge the speaker's accuracy on the same 7-point scales as used in Experiment 1.

Results

Success rates on retrieval-practice cycles. Table 2 shows mean retrieval success, separately for retrieval-practice cycles and delay conditions. A 3 × 2 ANOVA revealed a significant main effect for the factor *retrieval-practice cycle*, $F(2, 140) = 432.11$, $MSE = 153.77$, $p < .001$, $\eta^2 = .86$. As in Experiment 1, providing feedback improved performance from the first to the second (20.8% vs. 56.1%), $t(71) = 16.46$, $p < .001$, $d = 1.94$, and from the second to the third retrieval-practice cycle (56.1% vs. 76.1%), $t(71) = 13.63$, $p < .001$, $d = 1.61$. No other main or interaction effects were significant, all *F*s < 1.0, showing that there were no differences between delay conditions.

Accuracy monitoring performance. As in Experiment 1 we also analyzed performance during accuracy monitoring (see Table 2 for mean correct endorsement of accurate answers). A 3 × 2 ANOVA revealed a significant main effect for the factor retrieval-practice cycle, $F(2, 140) = 338.54$, $MSE = 211.28$, $p < .001$, $\eta^2 = .83$. Similar to retrieval-practice performance by speakers, accuracy monitoring performance by listeners improved from the first to the second (11.5% vs. 46.5%), $t(71) = 13.95$, $p < .001$, $d = 1.64$, and from the second to the third retrieval-practice cycle (46.5% vs. 67.6%), $t(71) = 13.18$, $p < .001$, $d = 1.55$. No other main effects or interactions reached significance (all *F*s < 1.0), showing that there were no differences between delay conditions.[3]

Recall on the final test. Figure 3 shows mean recall on the final test for practice conditions (retrieval practice, restudy, monitoring task) and delay conditions (3 min, 2 days). A 3 × 2 ANOVA revealed a significant main effect of practice, $F(2, 140) = 18.39$, $MSE = 257.96$, $p < .001$, $\eta^2 = .21$, indicating that recall was again differently affected by retrieval practice, restudy, and the monitoring task. In addition, the ANOVA showed a significant main effect of delay, $F(1, 70) = 30.45$, $MSE = 1470.46$, $p < .001$, $\eta^2 = .30$, reflecting forgetting over the two day retention interval. More importantly, we also found a significant interaction between the two factors, $F(2, 140) = 6.06$, $MSE = 257.96$, $p = .003$, $\eta^2 = .08$, indicating that forgetting across the 2-day delay was affected by practice format.

Further 2 × 2 ANOVAs contrasted time-dependent forgetting across practice conditions. A comparison of recall after retrieval practice and restudy again showed a significant interaction, $F(1, 70) = 5.74$, $MSE = 267.12$, $p = .019$, $\eta^2 = .08$. Whereas the two practice types caused similar recall levels after 3 min (74.2% vs. 68.9%), $t(35) = 1.17$, $p = .250$, $d = 0.19$, better recall after retrieval practice than after restudy was observed after the 2-day delay (48.1% vs. 29.7%), $t(35) = 6.02$, $p < .001$, $d = 1.00$. An ANOVA comparing recall after restudy and monitoring also revealed a significant interaction, $F(1, 70) = 9.59$, $MSE = 305.85$, $p = .003$, $\eta^2 = .12$. Although recall after the short delay was superior in the restudy compared to the monitoring condition (68.9% vs. 56.1%), $t(35) = 2.71$, $p = .010$, d = 0.45, after the 2-day delay there was a numerical disadvantage after restudy (29.7% vs. 35.0%), $t(35) = 1.54$, $p = .134$, $d = 0.26$. Thus, as in Experiment 1, the monitoring task reduced time-dependent forgetting relative to the restudy condition even though monitoring did not lead to reliably greater recall after a delay. Finally, a comparison of recall after monitoring and retrieval practice confirmed that time-dependent forgetting was again similar in these two practice conditions, $F(1, 70) = 1.12$, $MSE = 200.91$, $p = .294$, $\eta^2 = .02$. Irrespective of the change in instructions and subjects being explicitly asked to engage in covert retrieval practice during monitoring, recall was again clearly inferior after monitoring compared to retrieval practice—after both the short delay (74.2% vs. 56.1%), $t(35) = 5.07$, $p < .001$, $d = 0.85$, and the long delay (48.1% vs. 35.0%), $t(35) = 4.20$, $p < .001$, $d = 0.70$.

Discussion

The results of Experiment 2 replicate those of Experiment 1. Overt retrieval practice compared to restudy resulted in a typical testing effect, evident in better recall after longer delay (Pyc & Rawson, 2010; Roediger & Karpicke, 2006). Also, as in Experiment 1, engaging in accuracy monitoring again triggered at least some degree of covert retrieval practice, since the same reduction in time-dependent forgetting was observed as after overt retrieval practice. Yet even though subjects in Experiment 2 were explicitly asked to silently engage in retrieval practice when monitoring a

[3] Again, an additional ANOVA was run to directly compare retrieval-practice success by speakers and monitoring performance by listeners. As in Experiment 1, the ANOVA only revealed a significant main effect, $F(1, 70) = 9.04$, $MSE = 988.52$, $p = .004$, $\eta^2 = .11$, reflecting better recall performance by speakers than monitoring performance by listeners (51.0% vs. 41.9% across all retrieval-practice cycles). Listeners failed to recognize about 9% of the speakers' correct recalls. The lack of any interaction effects (all *F*s < 1.0) again suggests that monitoring performance generally followed retrieval-practice performance. Descriptive statistics for accuracy monitoring conditionalized on the speaker's correct recall again support this view, with listeners rating 53.0% of the speaker's correct answers as accurate on the first retrieval-practice cycle, 82.9% on the second, and 88.4% on the third retrieval-practice cycle (see the Appendix for additional information on errors committed during retrieval practice and accuracy monitoring).

Table 2
Mean Recall and Accuracy-Monitoring Performance in Experiment 2 are Displayed for Retrieval-Practice Cycles 1–3, Separately for Short and Long Delay Conditions

Delay	Retrieval Practice Cycle 1	Retrieval Practice Cycle 2	Retrieval Practice Cycle 3
3-min delay			
Recall performance	20.8% (17.3)	56.1% (25.0)	77.5% (20.3)
Accuracy monitoring	12.5% (15.0)	46.4% (24.6)	70.0% (23.7)
2-day delay			
Recall performance	20.7% (19.6)	56.1% (27.2)	74.7% (20.8)
Accuracy monitoring	10.6% (14.7)	46.7% (28.4)	65.3% (22.2)

Note. Recall performance = mean retrieval success of speakers during the ten trials of each retrieval practice cycle; Accuracy monitoring = the percentage of all ten trials in which the speakers' correct responses were identified as accurate by the listeners. Values in parentheses represent standard deviations of the means.

speaker's retrieval practice, this more specific instruction did not result in a notable boost in recall. As in Experiment 1, engaging in overt retrieval practice increased performance more than engaging in accuracy monitoring. Again, these results pose a puzzle, because in similar experiments covert retrieval practice has been shown to be as effective as overt retrieval practice in paired associate learning (Putnam & Roediger, 2013; Smith et al., 2013). The question remains: Can benefits from covert retrieval practice in a social setting be enhanced to match those of overt retrieval practice or to at least come close?

Another potential reason for why covert retrieval practice may be less efficient in social groups than in individuals could lie in social loafing (e.g., Karau & Williams, 1993; Latané, Williams, & Harkins, 1979; see also Weldon, Blair, & Huebsch, 2000). Subjects might be less motivated to engage in effortful (covert) retrieval practice when their performance is not directly assessed and when feedback is given after every trial. The monitoring tasks may have diminished personal accountability for listeners. Experiment 3 was conducted to examine this idea by asking listeners to monitor their own (covert) retrieval-practice performance instead of the speaker's performance. If social loafing is the decisive factor underlying the pattern of results observed in Experiments 1 and 2, requiring the listening participants to judge their own retrieval success may enhance personal accountability and therefore decrease the difference in recall between overt retrieval practice by speakers and covert retrieval practice by listeners.

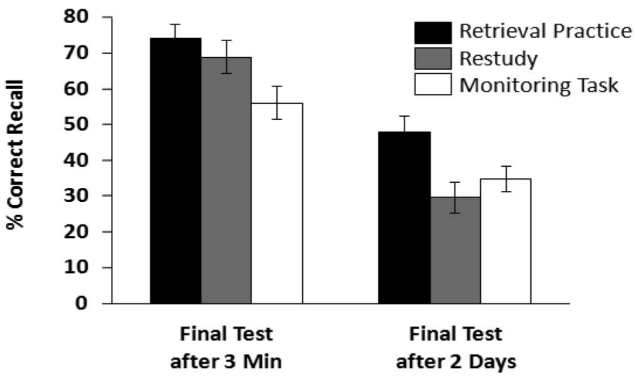

Figure 3. Mean recall performance on the final test in Experiment 2, shown as a function of delay (3-min delay, 2-day delay) and type of practice (retrieval practice, restudy, monitoring task). Error bars represent ±1 standard errors of the mean.

Experiment 3

Method

Participants. Seventy-two students at Regensburg University participated in return for partial course credit. Mean age was 22.4 years (SD = 3.2 years). Subjects were tested in pairs and evenly distributed across the two delay conditions, n = 36 in each. The same ethical standards were used as in Experiments 1 and 2.

Material. Study materials were the same as in Experiments 1 and 2, with vocabulary meanings translated to German.

Design. The experiment again employed a 3 × 2 design with the factors of *practice* (retrieval practice, restudy, monitoring task) and *delay* (3 min, 2 days).

Procedure. The procedure was largely the same as in Experiments 1 and 2, with one exception. When instructing listeners, we now asked them to monitor their own covert retrieval performance instead of the speaker's overt performance. Listeners were asked to silently try to recall the words. They were handed a prepared sheet with two response options (yes or no) for each practice trial and were asked to indicate whether or not they themselves were able to recall the target words. We switched to such dichotomous ratings in Experiment 3 to be able to more directly compare overt retrieval practice by speakers and covert retrieval practice by listeners. Timing of retrieval attempts by listeners was not controlled so as not to make the task more artificial. After cue presentation, listeners may have engaged in covert retrieval practice before, in parallel to, and/or after the speakers provided their overt responses.

Results

Success rates on overt retrieval-practice and covert monitoring cycles. Table 3 shows mean success rates during practice, separately for retrieval-practice cycles, monitoring cycles, and delay conditions. A 3 × 2 × 2 ANOVA revealed a significant main effect for the factor practice cycle, $F(2, 140) = 352.26$, $MSE = 208.45$, $p < .001$, $\eta^2 = .83$. Again, providing feedback improved performance from the first to the second and from the second to the third practice cycle, and this was true for both overt retrieval practice (16.3% vs. 47.1% vs. 61.3%), all $ts(71) \geq 7.28$, *ps* < .001, *ds* ≥ 0.86) and covert monitoring (18.3% vs. 47.5% vs. 61.8%), all *ts*(71) ≥ 7.55, *ps* < .001, *ds* ≥ 0.89). No other main or interaction effects were significant, all *F*s ≤ 1.23, *ps* ≥ .295, showing that performance did not differ between overt and covert practice, or between delay conditions. Of course, we could not check subjects' accuracy in the monitoring condition.

Recall on the final test. Figure 4 shows mean recall performance on the final test, separately for practice conditions (retrieval practice, restudy, monitoring task) and delay conditions (3 min, 2 days). To foreshadow, these results reveal a retrieval practice effect for the monitoring condition on the delayed test. A 3 × 2 ANOVA revealed no significant main effect of practice, $F(2, 140) = 2.61$, $MSE = 236.19$, $p = .077$, $\eta^2 = .04$, but a significant main effect of delay, $F(1, 70) = 18.16$, $MSE = 1{,}468.33$, $p < .001$, $\eta^2 = .21$, reflecting normal forgetting across the 2-day delay. We also found a significant Practice × Delay interaction, $F(2, 140) = 3.46$, $MSE = 236.19$, $p = .034$, $\eta^2 = .05$, suggesting that time-dependent forgetting differed across practice conditions.

A 2 × 2 ANOVA contrasting memory after retrieval practice and restudy again showed a significant interaction, $F(1, 70) = 5.15$, $MSE = 237.86$, $p = .026$, $\eta^2 = .07$. Recall levels were indistinguishable after 3 min (60.8% vs. 60.8%), $t(35) < 1.0$, but after 2 days recall was better after retrieval practice than restudy (42.5% vs. 30.8%), $t(35) = 4.72$, $p < .001$, d = 0.79. An ANOVA comparing recall after restudy and monitoring also revealed a significant interaction, $F(1, 70) = 5.03$, $MSE = 243.57$, $p = .028$, $\eta^2 = .07$. Although there was no difference after the short delay (60.8% vs. 57.5%), $t(35) < 1.0$, $p = .397$, $d = 0.14$, after the 2-day delay recall was better after monitoring than after restudy (39.2% vs. 30.8%), $t(35) = 2.41$, $p = .021$, $d = 0.40$. Finally, a comparison of monitoring and retrieval practice confirmed that time-dependent forgetting was again comparable in these two practice conditions, $F(1, 70) < 1.00$, $MSE = 227.14$, $p = 1.00$, $\eta^2 < .001$. This time, however, there was also no difference in recall levels between the two practice conditions, either after the short delay (60.8% vs. 57.5%), $t(35) < 1.00$, $p = .359$, $d = 0.16$, or after the long delay (42.5% vs. 39.2%), $t(35) < 1.00$, $p = .350$, $d = 0.16$). Thus, we found a roughly comparable retrieval practice effect in the overt and the covert (monitoring) conditions.

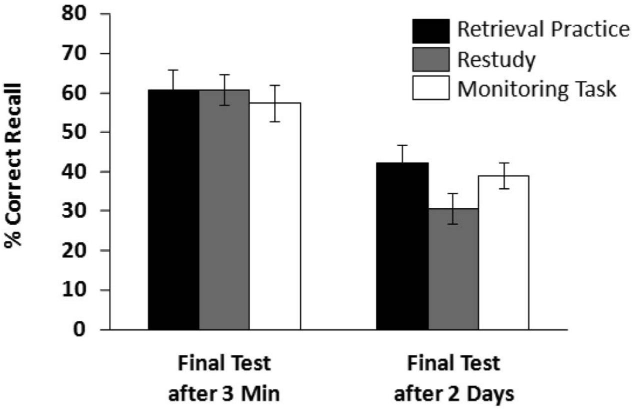

Figure 4. Mean recall performance on the final test in Experiment 3, shown as a function of delay (3-min delay, 2-day delay) and type of practice (retrieval practice, restudy, monitoring task). Error bars represent ±1 standard errors of the mean.

Discussion

Experiment 3 again revealed a typical testing effect, with retrieval practice causing better recall than restudy after 2 days. Importantly, however, Experiment 3 reported that covert monitoring in a social setting can be as effective as overt retrieval practice, at least if one focuses on monitoring one's own retrieval rather than another person's. Under these conditions, overt and covert retrieval practice can be equally useful for retention. Presumably, the changed monitoring task in Experiment 3 increased personal accountability and thus motivated listeners to engage in more effortful retrieval, which has also been suggested to increase testing effects in individual recall (e.g., Karpicke & Roediger, 2007; Maddox & Balota, 2015; Pyc & Rawson, 2009). Alternatively, one could argue that the changed monitoring task also reduced overall task demands and the necessity to divide attention between one's own retrieval attempts and making time-limited monitoring judgments for the speaker's performance. However, because prior work has shown that divided attention decreases the efficiency of restudy, but not of retrieval practice (Gaspelin, Ruthruff, & Pashler, 2013; Mulligan & Picklesimer, 2016), this

Table 3
Mean Recall and Accuracy-Monitoring Performance in Experiment 3, Displayed for Retrieval-Practice Cycles 1–3 and Separately for Short and Long Delay Conditions

Delay	Retrieval Practice Cycle 1	Retrieval Practice Cycle 2	Retrieval Practice Cycle 3
3-min delay			
Recall performance	18.6% (20.9)	48.3% (27.0)	60.8% (25.9)
Monitoring performance	20.8% (21.2)	50.6% (24.3)	65.8% (22.1)
2-day delay			
Recall performance	13.9% (14.0)	45.8% (26.9)	61.7% (28.8)
Monitoring performance	15.8% (14.0)	44.4% (23.0)	57.8% (22.6)

Note. Recall performance = mean retrieval success of speakers during overt retrieval practice; Monitoring performance = mean success of listeners during covert retrieval practice (as indexed by their own judgments). Values in parentheses represent standard deviations of the means.

alternative explanation seems less likely. A different issue that should be kept in mind though is that Experiment 3 was conducted in Germany, whereas Experiments 1 and 2 were conducted in the United States. Although there is no a priori reason to assume that there are general differences between participants from the two countries that could have affected the results in the present experiments, we also cannot rule out this possibility.

By asking listeners to monitor their own instead of the speaker's retrieval, Experiment 3 may have succeeded in creating conditions under which listeners can benefit from covert retrieval, but at the same time these task instructions may also limit the extent to which the experiment captures a representative social setting. In essence, monitoring one's own retrieval for accuracy is the same task that has been shown to entail effective covert retrieval in individual recall (see Putnam & Roediger, 2013; Smith et al., 2013), and the present data show that it may also be effective when one is simultaneously exposed to another person's retrieval. Of course, in other types of social settings more interaction may occur and retrieval dynamics may be very different. For instance, when students study together, they may be more intrinsically motivated to engage in concurrent retrieval practice, for example, to provide each other with feedback (for prior work focusing on such more interactive forms of social practice, see Congleton & Rajaram, 2011; Wissman & Rawson, 2016; for a review on collaborative recall, see Rajaram, 2011). Nevertheless, the present experimental set up has implications for the conditions under which practice questions that teachers often pose to a whole classroom have the potential to benefit all students, not just the few that end up engaging in retrieval practice out loud.

General Discussion

The present experiments investigated the testing effect in a social context, addressing whether listening to and monitoring another person's reports can entail similar benefits for listeners as for speakers. Although two separate lines of prior research had indicated that this might be the case (see Cuc et al., 2007; Putnam & Roediger, 2013; Smith et al., 2013), the present experiments showed that listening to another person's retrieval reports is not equally beneficial as engaging in overt retrieval practice oneself. Monitoring another person's retrieval practice decreased time-dependent forgetting as much as overt retrieval practice (relative to a restudy condition), but overt retrieval practice was still more beneficial for recall than monitoring. In fact, relative to restudy, monitoring did not boost performance at all. Importantly, this outcome emerged irrespective of whether listeners were explicitly asked to engage in covert retrieval practice along with speakers (in Experiment 2) or not (in Experiment 1). We only obtained a significant effect of monitoring relative to restudy in Experiment 3 when we asked listeners to focus on their own covert retrieval performance rather than the speakers' responses, which was essentially the task in Putnam and Roediger (2013) and Smith et al. (2013) on individual recall. Only under this instruction did we observe similar benefits of overt retrieval practice in speakers and covert retrieval practice in listeners. In terms of applied implications, our results suggest that students may only benefit from retrieval practice in social situations if they are actively engaged in attempting to retrieve the material. Asking them to monitor another person's response is not sufficient.

Thus, the beneficial effects of retrieval practice (i.e., testing effects) are not easily shared in a social setting. Although Cuc et al. (2007) argued that retrieval-practice effects may be socially shared when listeners engage in concurrent covert retrieval during accuracy monitoring, we used a similar monitoring task and did not observe a retrieval practice via monitoring effect (relative to restudy) in Experiments 1 and 2. However, in both experiments time-dependent forgetting was reduced after both overt retrieval practice and monitoring, again relative to restudy, indicating that subjects were engaged in some amount of covert retrieval practice during monitoring. Still, because the two monitoring conditions led to similar or worse final recall in Experiments 1 and 2 relative to the restudy condition, the results indicate that more effort in the covert retrieval condition is needed to provide a positive effect.

Clearly, the testing effect in standard retrieval practice experiments may differ from those in retrieval-induced forgetting experiments when only some items are tested. Some studies indicate that retrieval-induced forgetting may not hinge on success during retrieval practice (e.g., Storm, Bjork, Bjork, & Nestojko, 2006; Storm & Nestojko, 2010), but robust testing effects seem to depend on repeated, successful, and effortful retrieval (e.g., Butler & Roediger, 2007; Karpicke, 2009; Karpicke & Roediger, 2007, 2008). Thus, the same covert form of retrieval practice in a social context that can induce socially shared retrieval-induced forgetting may fail to create socially shared testing effects of the same magnitude in listeners. Consistently, for two out of their three experiments on socially shared retrieval-induced forgetting, Cuc et al. (2007) reported that recall for practiced items was worse in listeners than in speakers (no difference was observed in the other experiment); yet, even though listeners were less successful than speakers at recalling the practiced items, they still showed intact (socially shared) retrieval-induced forgetting.

When listeners monitored their own covert retrieval in our Experiment 3 rather than the speaker's reports, recall was enhanced relative to restudying and comparable benefits of covert and overt practice emerged. Presumably, focusing on one's own learning may have motivated subjects to increase their retrieval effort, which in turn increased performance. This finding is consistent with prior work suggesting that listeners must be specifically motivated to engage in effortful retrieval along with speakers (e.g., Cuc et al., 2007; Koppel et al., 2014), and it is also consistent with the retrieval effort hypothesis of the testing effect (Pyc & Rawson, 2009), derived from the desirable difficulties framework (Bjork, 1994, 1999). Although monitoring another person's performance may also trigger some covert retrieval practice in listeners, such monitoring judgments can be made without investing too much effort (e.g., by recognition evaluations alone, and these might be based in part on familiarity of the answer rather than recollection). In contrast, monitoring one's own learning must encourage active retrieval and recollection, and thus, much more effortful processes that have been shown to benefit learning more than recognition judgments (e.g., Carpenter & DeLosh, 2006; Kang et al., 2007; see also Rowland, 2014). In the present study, covert retrieval practice only enhanced recall relative to restudy when such effortful retrieval was required (in Experiment 3). Overall, this outcome indicates that effortful retrieval may not only benefit retention of individuals practicing in isolation, but also when retrieval practice is employed in a social setting. Yet, because participants in Experiment 3 also differed in nationality—they were German—further research should examine the robustness of the finding. Of course, given the results of Putnam and Roediger (2013) and Smith et al. (2013), we strongly doubt that the differing nationalities and languages of university students represents the critical difference in the results of Experiment 3 relative to that of the earlier experiments.

Pashler, Kang, and Harris (2012) have conducted research in both lab and classroom settings asking if teachers' questions to a whole class may induce covert retrieval practice and hence benefit all students rather than only the student who answers the question overtly. The inspiration for their experiment came from a California school that greatly improved students' performance by adopting a procedure that put all students in every class "on the hook" for questions. That is, students each had an assigned number for class and the teacher would ask a question, wait, and then call on one student at random (and then sample with replacement, so the student could be called on again in the same session). Thus, all students were "on the hook" and had to try to think of the answer to the question in preparation for the possibility of being called upon. Of course, the principal who instituted this reform also made other changes, but the teachers and principal attributed a large portion of their success to this universal adoption of the "on the hook" procedure in the classrooms, combined with the teachers asking many questions. Pashler et al. (2012) brought this procedure into the lab and showed that indeed placing students "on the hook" relative to the standard way of answering questions (one student raises his or her hand and answers while the others look on) improved performance. This outcome fits well with conclusions from our experiments: Students may learn best from questions being asked by a teacher (or a fellow student) only if they work to generate an answer themselves, even if covertly, as in our Experiment 3.

We must provide one caveat to our findings. We used paired-associate learning, which is the task that has generally been used both in research on the effects of covert retrieval practice (Putnam & Roediger, 2013; Smith et al., 2013) and in socially shared retrieval-induced forgetting (Cuc et al., 2007). Recently, Tauber et al. (in press) have examined covert retrieval practice using key word definitions of the sort that one finds in textbooks (e.g., the definition of cognitive dissonance). They reported several experiments in which no benefit of covert retrieval practice occurred despite robust effects of overt retrieval practice with these materials. Thus the issue remains open as to whether covert retrieval practice in social situations will produce a positive effect for more natural materials than paired-associates. Pashler et al. (2012) used prose passages in their "on the hook" experiment and so it certainly seems possible that the effect will be established with text materials, but clearly further research is needed.

In sum, the present experiments showed that, in a social context, asking questions that one student answers may not always lead to a benefit for other students who are listening, even if they are monitoring for accuracy of the speaker's response. Of course, in a typical classroom or study group, students who are not called on may not be trying to retrieve the answer or even paying attention to the correct answer when it is given. Yet our findings show that even if the students (the listeners) are monitoring the speaker's reports for accuracy or fluency, no positive effect occurs (relative to restudying, at least). Rather, only when conditions exist that encourage students to covertly retrieve their own response and judge its accuracy did we find a benefit (in Experiment 3). This observation, together with Pashler et al.'s (2012) research, suggests that care must be taken in class or in study groups to place students "on the hook" so that they will engage in effortful covert retrieval. Asking other students in class to "think along silently" and to monitor the response of the student who answers may not produce a beneficial effect relative to restudy, although the process might at least slow time-dependent forgetting. Of course, our experiments used two days as the longest retention interval. Because covert retrieval even in Experiments 1 and 2 slowed forgetting (relative to restudy), we might have observed a positive effect of monitoring if the retention interval had been longer (say, a week). This possibility awaits further research.

Finally, the primary practical implication of our research is that the practice of asking questions in class and then calling on one student to answer—quite common in the classroom—may not be an effective technique for encouraging effortful processing in other students unless the teacher tries to make each student potentially responsible to provide an answer. Students apparently do not naturally exert effort to do so, even when they are asked to monitor another student's answers.

References

Abel, M., & Bäuml, K.-H. T. (2014). The roles of delay and retroactive interference in retrieval-induced forgetting. *Memory & Cognition, 42,* 141–150. http://dx.doi.org/10.3758/s13421-013-0347-0

Abel, M., & Bäuml, K.-H. T. (2015). Selective memory retrieval in social groups: When silence is golden and when it is not. *Cognition, 140,* 40–48. http://dx.doi.org/10.1016/j.cognition.2015.03.009

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. http://dx.doi.org/10.1037/0278-7393.20.5.1063

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

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). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance: XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Brown, A. D., Kramer, M. E., Romano, T. A., & Hirst, W. (2012). Forgetting trauma: Socially shared retrieval-induced forgetting and post-traumatic stress disorder. *Applied Cognitive Psychology, 26,* 24–34. http://dx.doi.org/10.1002/acp.1791

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

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,* 268–276.

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

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

Coman, A., & Hirst, W. (2015). Social identity and socially shared retrieval-induced forgetting: The effects of group membership. *Journal of Experimental Psychology: General, 144,* 717–722. http://dx.doi.org/10.1037/xge0000077

Coman, A., Manier, D., & Hirst, W. (2009). Forgetting the unforgettable through conversation: Socially shared retrieval-induced forgetting of September 11 memories. *Psychological Science, 20,* 627–633. http://dx.doi.org/10.1111/j.1467-9280.2009.02343.x

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. http://dx.doi.org/10.1037/a0024308

Cuc, A., Koppel, J., & Hirst, W. (2007). Silence is not golden: A case for socially shared retrieval-induced forgetting. *Psychological Science, 18,* 727–733. http://dx.doi.org/10.1111/j.1467-9280.2007.01967.x

Gaspelin, N., Ruthruff, E., & Pashler, H. (2013). Divided attention: An undesirable difficulty in memory retention. *Memory & Cognition, 41,* 978–988. http://dx.doi.org/10.3758/s13421-013-0326-5

Hirst, W., & Echterhoff, G. (2012). Remembering in conversations: The social sharing and reshaping of memories. *Annual Review of Psychology, 63,* 55–79. http://dx.doi.org/10.1146/annurev-psych-120710-100340

Hirst, W., & Manier, D. (2008). Towards a psychology of collective memory. *Memory, 16,* 183–200. http://dx.doi.org/10.1080/09658210701811912

Izawa, C. (1976). Vocalized and silent tests in paired-associate learning. *The American Journal of Psychology, 89,* 681–693. http://dx.doi.org/10.2307/1421466

Jönsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research, 78,* 623–633. http://dx.doi.org/10.1007/s00426-013-0522-8

Kang, S. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on memory retention. *European Journal of Cognitive Psychology, 19,* 528–558. http://dx.doi.org/10.1080/09541440601056620

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681–706. http://dx.doi.org/10.1037/0022-3514.65.4.681

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

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. http://dx.doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. http://dx.doi.org/10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L., III (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67,* 17–29. http://dx.doi.org/10.1016/j.jml.2012.02.004

Koppel, J., Wohl, D., Meksin, R., & Hirst, W. (2014). The effect of listening to others remember on subsequent memory: The roles of expertise and trust in socially shared retrieval-induced forgetting and social contagion. *Social Cognition, 32,* 148–180. http://dx.doi.org/10.1521/soco.2014.32.2.148

Latané, B., Williams, K. D., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822–832. http://dx.doi.org/10.1037/0022-3514.37.6.822

Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty. *Memory & Cognition, 43,* 760–774. http://dx.doi.org/10.3758/s13421-014-0499-6

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

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20,* 3–21. http://dx.doi.org/10.1037/xap0000004

Meade, M. L., & Roediger, H. L., III. (2002). Explorations in the social contagion of memory. *Memory & Cognition, 30,* 995–1009. http://dx.doi.org/10.3758/BF03194318

Mulligan, N. W., & Picklesimer, M. (2016). Attention and the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 938–950. http://dx.doi.org/10.1037/xlm0000227

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2,* 325–335. http://dx.doi.org/10.1080/09658219408258951

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8. http://dx.doi.org/10.1037/0278-7393.31.1.3

Pashler, H., Kang, S., & Harris, C. (2012, November). *Testing effects in memory: "On the hook" in simulated classrooms*. Paper presented at the annual meeting of the Psychonomic Society, Minneapolis, MN.

Putnam, A. L., & Roediger, H. L., III. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41,* 36–48. http://dx.doi.org/10.3758/s13421-012-0245-x

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. http://dx.doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. http://dx.doi.org/10.1126/science.1191465

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

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

Roediger, H. L., III, & Abel, M. (2015). Collective memory: A new arena of cognitive study. *Trends in Cognitive Sciences, 19,* 359–361. http://dx.doi.org/10.1016/j.tics.2015.04.003

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. http://dx.doi.org/10.1016/j.tics.2010.09.003

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

Roediger, H. L., III, Meade, M. L., & Bergman, E. T. (2001). Social contagion of memory. *Psychonomic Bulletin & Review, 8,* 365–371. http://dx.doi.org/10.3758/BF03196174

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. http://dx.doi.org/10.1037/a0037559

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

Stone, C. B., Barnier, A. J., Sutton, J., & Hirst, W. (2010). Building consensus about the past: Schema consistency and convergence in socially shared retrieval-induced forgetting. *Memory, 18,* 170–184. http://dx.doi.org/10.1080/09658210903159003

Stone, C. B., Barnier, A. J., Sutton, J., & Hirst, W. (2013). Forgetting our personal past: Socially shared retrieval-induced forgetting of autobiographical memories. *Journal of Experimental Psychology: General, 142,* 1084–1099. http://dx.doi.org/10.1037/a0030739

Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review, 13,* 1023–1027. http://dx.doi.org/10.3758/BF03213919

Storm, B. C., & Nestojko, J. F. (2010). Successful inhibition, unsuccessful retrieval: Manipulating time and success during retrieval practice. *Memory, 18,* 99–114. http://dx.doi.org/10.1080/09658210903107853

Tauber, S. K., Witherby, A. E., Dunlosky, J. D., Rawson, K. A., Putnam, A. L., & Roediger, H. L., III. (in press). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory & Cognition*. Advance online publication. http://dx.doi.org/10.1016/j.jarmac.2016.10.004

Weldon, M. S., Blair, C., & Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1568–1577. http://dx.doi.org/10.1037/0278-7393.26.6.1568

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

Appendix

Errors During Retrieval Practice and Accuracy Monitoring

In the main text, we reported average correct performance of speakers on retrieval practice cycles and the percentage of trials on each cycle for which listeners rated the correct answers given by speakers as accurate. Naturally, however, speakers sometimes made mistakes when trying to recall vocabulary pairs during retrieval practice, and listeners may have made additional mistakes by rating the speakers' incorrect answers as accurate (or their correct answers as inaccurate). For reasons of completeness, we will provide descriptive statistics for the occurrence of such errors in the following section, separately for each experiment. It should be noted, however, that analyzing errors will not enable strong conclusions about covert retrieval-practice performance by listeners (e.g., even for cases in which listeners correctly rejected a speaker's incorrect response as inaccurate, it cannot automatically be assumed that listeners were able to retrieve the correct answer themselves).

Experiment 1

Errors by Speakers

In Experiment 1, speakers committed on average 3.36 errors during retrieval practice (*SD* = 4.02). This corresponds to 11.2% of all 30 retrieval-practice trials. Extralist intrusions from items that were never studied were less common (across all 30 retrieval-practice trials: M = 0.93, *SD* = 1.56) compared to intralist intrusions from items that were studied but incorrectly paired with a different Swahili cue (*M* = 2.47, *SD* = 2.92). The mean number of extralist intrusions decreased from the first retrieval-practice cycle (*M* = 0.68, *SD* = 1.18) to the second retrieval-practice cycle (*M* = 0.17, *SD* = 0.44), *t*(71) = 4.06, $p < .001$, with no major change occurring from the second to the third retrieval-practice cycle (*M* = 0.10, *SD* = 0.38), *t*(71) = 1.15, *p* = .254. In contrast, mean number of intralist intrusions was relatively low on the first retrieval-practice cycle (*M* = 0.43, *SD* = 0.78), increased on the second retrieval-practice cycle (*M* = 1.06, *SD* = 1.31), *t*(71) = 4.91, $p < .001$, and stayed on this level on the third retrieval-practice cycle (*M* = 0.99, *SD* = 1.37), $t(71) < 1.0$, *p* = .638.

Error Detection by Listeners

Out of the 3.36 (*SD* = 4.22) errors that were on average committed by speakers, listeners caught a mean of 1.67 errors (*SD* = 2.38) by correctly rating them as inaccurate. Thus, listeners caught 49.7% of all errors that were committed by speakers, presumably before being exposed to corrective feedback on the screen. Yet, listeners committed on average 0.54 errors of their own (*SD* = 1.14). In particular, listeners rated on average 0.38 (*SD* = 0.96) incorrect answers as accurate and 0.17 (*SD* = 0.41) correct answers as inaccurate.

(*Appendix continues*)

Experiment 2

Errors by Speakers

In Experiment 2, speakers committed on average 2.11 (intralist and extralist) errors (*SD* = 2.17), corresponding to 7% of all 30 retrieval-practice trials. Again, extralist intrusions were less common (across all 30 retrieval-practice trials: *M* = 0.49, *SD* = 0.90) than intralist intrusions (*M* = 1.63, *SD* = 1.67). In parallel to Experiment 1, extralist intrusions decreased from the first retrieval-practice cycle (*M* = 0.32, *SD* = 0.71) to the second retrieval-practice cycle (*M* = 0.10, *SD* = 0.38), *t*(71) = 2.38, *p* = .020, with no major change occurring from the second to the third retrieval-practice cycle (*M* = 0.07, *SD* = 0.26), *t*(71) < 1.0, *p* = .531. In contrast, mean number of intralist intrusions was relatively low on the first retrieval-practice cycle (*M* = 0.39, *SD* = 0.72), increased on the second retrieval-practice cycle (*M* = 0.74, *SD* = 0.96), *t*(71) = 2.67, *p* = .009, and did not decrease substantially on the third retrieval-practice cycle (*M* = 0.50, *SD* = 0.84), *t*(71) = 1.85, *p* = .068.

Error Detection by Listeners

Out of the mean number of 2.11 (*SD* = 2.17) errors committed by speakers, listeners caught on average 1.14 errors (*SD* = 1.89) by correctly rating them as inaccurate. Thus, listeners caught 54.0% of all errors that were committed by speakers, presumably before they were exposed to corrective feedback. Listeners themselves committed on average 0.36 errors (*SD* = 0.72). In particular, listeners rated 0.15 (*SD* = 0.42) incorrect answers as accurate and 0.21 (*SD* = 0.47) correct answers as inaccurate.

Experiment 3

Errors by Speakers

In Experiment 3, speakers committed on average 2.72 (intralist and extralist) errors (*SD* = 2.64), corresponding to 9% of all 30 retrieval-practice trials. Again, extralist intrusions were less common (across all 30 retrieval-practice trials: *M* = 0.35, *SD* = 0.72) than intralist intrusions (*M* = 2.42, *SD* = 2.41). As in both previous experiments, extralist intrusions decreased from the first retrieval-practice cycle (*M* = 0.25, *SD* = 0.55) to the second retrieval-practice cycle (*M* = 0.07, *SD* = 0.26), *t*(71) = 2.71, *p* = .008, but no change occurred from the second to the third retrieval-practice cycle (*M* = 0.04, *SD* = 0.20), *t*(71) < 1.0, *p* = .418. In contrast, mean number of intralist intrusions was lowest on the first retrieval-practice cycle (*M* = 0.44, *SD* = 0.67), increased on the second retrieval-practice cycle (*M* = 0.94, *SD* = 1.20), *t*(71) = 3.38, *p* = .001, and roughly stayed the same on the third retrieval-practice cycle (*M* = 1.00, *SD* = 1.21), *t*(71) < 1.0, *p* = .700.

Received May 17, 2017
Revision received August 9, 2017
Accepted August 19, 2017 ■