Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Does Covert Retrieval Benefit Learning of Key-Term Definitions?

Sarah K. Tauber*, Amber E. Witherby
Texas Christian University, United States

John Dunlosky, Katherine A. Rawson
Kent State University, United States

Adam L. Putnam
Carleton College, United States

Henry L. Roediger III
Washington University in St Louis, United States

Even though retrieval practice typically has a robust, positive influence on memory, response format (overt vs. covert retrieval) may moderate its effect when students learn complex material. Overt retrieval is likely to promote exhaustive retrieval, whereas covert retrieval may not be exhaustive for familiar key terms. In two experiments, students were instructed to study key-term definitions and were asked to practice retrieval overtly, to practice retrieval covertly, or to restudy the definitions. Students also made metacognitive judgments. A final criterion test was administered two days later. Students' final recall was greater after overt retrieval practice than after covert retrieval practice or restudy, with a continuously cumulating meta-analysis establishing the effect as moderate in size (pooled $d = 0.43$). Thus, response format does matter for learning definitions of key terms, supporting the recommendation that students use overt retrieval when using retrieval practice as a strategy to learn complex materials.

Keywords: Covert retrieval, Retrieval practice, Key-term definitions, Monitoring of learning, Metacognition

General Audience Summary

One strategy that typically improves students' memory is to test themselves on information that they need to learn. Students may do so by speaking their answers out loud, by writing or typing their answers, or by mentally answering each question. For instance, a student studying in a library may mentally answer questions to avoid distracting others. By contrast, a student studying with a group may offer answers out loud as a part of the group discussion. Our interest was to evaluate whether these different types of responses (typed recall vs. mental recall) influence how effective self-testing is for improving students' memory when

Authors Note

Sarah "Uma" Tauber, Department of Psychology, Texas Christian University; Amber E. Witherby, Department of Psychology, Texas Christian University; John Dunlosky, Department of Psychological Sciences, Kent State University; Katherine A. Rawson, Department of Psychological Sciences, Kent State University; Adam L. Putnam, Department of Psychology, Carleton College; Henry L. Roediger III, Department of Psychological and Brain Sciences, Washington University in St. Louis.

We would like to thank Mariah Beltran and Melissa Bishop for their assistance with data scoring.

This research was supported by the James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award.

Please note that this paper was handled by the current editorial team of JARMAC.

* Correspondence concerning this article should be addressed to Sarah "Uma" Tauber Department of Psychology, Texas Christian University, TCU Box 298920, 2800 S. University Dr., Fort Worth, TX 76129, United States. Contact: uma.tauber@tcu.edu.

Please cite this article in press as: Tauber, S. K., et al. Does Covert Retrieval Benefit Learning of Key-Term Definitions? *Journal of Applied Research in Memory and Cognition* (2016), http://dx.doi.org/10.1016/j.jarmac.2016.10.004

they learn key-term definitions. In two experiments, students studied key terms (e.g., *self-serving bias*) and the corresponding definition for each (*When explaining one's own behavior it is the tendency to attribute good behaviors to one's disposition and to attribute bad behaviors to the situation*). Students then restudied the key terms and definitions, or tested themselves on them. Students who tested themselves typed the definition for each term, or were instructed to mentally recall the definition for each. Students in all three groups also made judgments about their memory and returned two days later to complete a final memory test. In a first experiment, students' memory on the final test was greater after typing the recalled definitions than after mentally recalling the definitions, or after restudying the definitions. In a second experiment, the same patterns were evident, although the memory benefit after typing the recalled definitions was smaller. These results suggest that *how* students test themselves is important when they are learning conceptual definitions. Thus, our recommendation is that students type out recalled answers during self-testing when they are learning relatively complex materials.

Retrieval practice typically benefits learning and memory, and its benefits have been referred to as *test-enhanced learning* (for a review see Rowland, 2014). This robust benefit has been demonstrated across a wide range of materials, learners, outcomes, and settings (for reviews, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011). Although retrieval practice has a robust effect on people's retention, some factors moderate its benefits (e.g., Kornell, Hays, & Bjork, 2009; Roediger & Karpicke, 2006). Thus, fully understanding the benefits of retrieval practice will involve discovering its moderators, which is the main goal of the present research. In particular, we evaluated whether the benefits of retrieval practice are moderated by response format (i.e., overt versus covert responding). Investigating response format is important because students are likely to adopt different formats in different contexts. For instance, a student using retrieval practice in a library may covertly retrieve answers (i.e., mental retrieval) to avoid distracting others. By contrast, a student studying with a group may overtly retrieve answers as a part of the group discussion. As such, it is critical to estimate the effectiveness of each response format, which leads us to the key question of this research: Does response format influence the magnitude of final recall performance when students attempt to learn key-term definitions?

Previous research suggests that overt and covert retrieval have similar effects on learning. Putnam and Roediger (2013) explored the influence of response format during retrieval practice when students learned paired associates (e.g., *airplane-trip*). After initial study, participants were instructed to overtly practice retrieving the target for some pairs, to covertly practice retrieving the target for some pairs, and to restudy some word pairs. On a final cued-recall test 2 days later, participants were shown each cue word (e.g., *airplane-?*) and were asked to type the target (i.e., *trip*). Performance on the final recall test was superior after retrieval practice (overt or covert) as compared to no retrieval practice (i.e., restudied word pairs), and it was similar for overt and covert practice. Smith, Roediger, and Karpicke (2013) analyzed the results from 10 experiments comparing the effect of overt (vs. covert) retrieval on memory, and their analysis yielded an effect size close to zero ($d = -.0027$). By contrast, comparing retrieval practice (either covert or overt) to no retrieval practice yielded a large effect size ($d = 1.1$) in favor of retrieval practice.

In other studies, however, response format for retrieval practice has recently been shown to have a minor influence on paired associate recall. Jönsson, Kubrik, Sundqvist, Todorov, and Jonsson (2014) had participants study Swahili–Swedish translations, and participants were instructed to practice overt retrieval or covert retrieval, or to restudy the pairs in preparation for immediate or delayed tests. Although overt retrieval practice was superior to covert retrieval practice in one experiment using a within-participant design, it was a small effect ($d = 0.21$ for a long retention interval). It was also not robust; overt retrieval did not statistically differ from covert retrieval in another experiment (or after a short retention interval). Thus, across all the available studies, retrieval practice appears to be effective with covert responding, at least with paired associates.

Response format for retrieval practice may not matter much for learning paired associates because presenting the cue alone triggers a retrieval attempt (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996); so, regardless of whether responses are covert or overt, people are expected to initiate retrieval of the single-word response. The situation may be different for longer and more complicated material. In such cases, covert retrieval may not benefit recall as much as overt retrieval because students may not undergo exhaustive retrieval. For example, consider students learning key-term definitions, such as the definition of *confirmation bias* (answer: *The tendency to only seek out or attend to information that confirms one's belief and to ignore counterevidence*). For these materials, the retrieval demands are presumably higher because students need to retrieve multiple units of information to accurately represent the response. And, if students feel they are familiar with the concept, this familiarity may short-circuit a retrieval attempt (e.g., by responding, "Oh, I already know that one"). In the case of unfamiliar terms, students may not even try to retrieve the answer. In either case, if students do not attempt to exhaustively retrieve the definition when they covertly practice retrieval, then no benefit would be expected. By contrast, during overt practice, students may be more likely to fully retrieve the definition simply because they are being asked to type (or say aloud) as much of it as possible. Thus, when compared with covert retrieval

Please cite this article in press as: Tauber, S. K., et al. Does Covert Retrieval Benefit Learning of Key-Term Definitions? *Journal of Applied Research in Memory and Cognition* (2016), http://dx.doi.org/10.1016/j.jarmac.2016.10.004

practice, overt retrieval practice may be more beneficial for students' retention of larger units of information.

Only one study has investigated response format using lengthy materials (800-word prose passages; Orlando & Hayward, 1978). In the reread group, participants read through the passage and were then instructed to reread it (i.e., restudy group). In the mental review group, after reading each paragraph, participants looked away from the passage and attempted to mentally recall it, and then examined it again for feedback (i.e., covert retrieval practice). The notetaking group was identical to the mental review group, except that participants were required to write the information recalled after each paragraph (i.e., overt retrieval practice). Participants' memory was evaluated on immediate and delayed tests. On the immediate test, memory performance was enhanced for both the mental review group and the notetaking group relative to the reread group. On the delayed test, the same trends were evident but not significant. Thus, when students learned lengthy prose passages, covert retrieval benefitted their learning as much as overt retrieval. However, aspects of the method limit the relevance of this research. In particular, their effects did not reach conventional levels of statistical significance following a long retention interval, which may be attributable to low power ($n = 16$ per group). Retrieval-practice tasks were also interpolated throughout reading the passage, so it is likely that they were based (at least in part) on access to information active in working memory. By contrast, our primary interest was in the influence of response formats on the benefits of retrieval practice when retrieval occurs from long-term memory, which is a standard procedure for demonstrating the benefits of retrieval practice (cf. Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008).

For the current experiments, our primary goal was to investigate whether covert retrieval practice benefits students' learning and memory as much as overt retrieval practice when students learn definitions of key terms. We selected key-term definitions because students are often required to learn them, especially in introductory courses that rely heavily on learning foundational terms to a discipline. In addition, prior research with overt retrieval practice has shown healthy testing effects (for a review see Rowland, 2014). Based on the aforementioned rationale, we predicted that final recall performance would be lower after covert than overt retrieval practice. We evaluated this possibility in two experiments in which students learned definitions to key terms and were then instructed to practice overt retrieval (with feedback), covert retrieval (with feedback), or to restudy the key-term definitions. Both experiments included this basic design, which was inspired by recent emphasis on the importance of replication and recommendations to base conclusions on cumulative outcomes involving multiple estimates of effect sizes (e.g., Braver, Thoemmes, & Rosenthal, 2014; Lishner, 2015; Maner, 2014; Pashler & Harris, 2012; Simons, 2014). Our secondary focus was on the influence of response format on students' monitoring of learning, which we explored by including metacognitive judgments as discussed below.

Experiment 1

In Experiment 1, students studied key-term definitions from social and cognitive psychology from an introductory psychology textbook, with each consisting of a key term (e.g., *self-serving bias*) and the corresponding definition (*When explaining one's own behavior it is the tendency to attribute good behaviors to one's disposition and to attribute bad behaviors to the situation*). We selected key-term definitions from two fields in psychology to evaluate covert retrieval practice when students learn different samples of items to establish whether effects evident with one item set can be replicated with another set (cf. Westfall, Judd, & Kenny, 2015).

Students then received three trials in which they either restudied the key terms and definitions or engaged in retrieval practice. Students in the restudy group were given the key terms and definitions and were asked to type the definition for each. Students in the overt retrieval group were given only the key terms and were asked to recall and type the definition, after which they received feedback (i.e., they were presented with the correct definition). Students in the covert retrieval group were given only the key terms and were then instructed to covertly recall the definitions, after which they received feedback. Following restudy or retrieval practice, students in all groups made item-by-item judgments assessing their level of knowledge for the key terms. Finally, students in all groups returned two days later to complete a final criterion test.

Method

Design and participants. Practice group (Restudy, Covert Retrieval, Overt Retrieval) was a between-participants manipulation. Ninety-five students from Kent State University participated in exchange for course credit and were randomly assigned to practice group. Of those students, 5 failed to return for the second session and were excluded from analyses. Thus, data from 90 students are reported ($n = 29$ in the restudy group, $n = 31$ in the covert retrieval group, $n = 30$ in the overt retrieval group).

Materials and procedure. Materials included 16 key terms (8 social psychology terms and 8 cognitive psychology terms) and their associated definitions (from Rawson & Dunlosky, 2011). Social psychology terms (e.g., *attribution*, *correspondence bias*) comprised List 1, and cognitive psychology terms (e.g., *encoding*, *sensory memory*) comprised List 2. Forty-five participants studied List 1 ($n = 14$ in the restudy group, $n = 15$ in the covert retrieval group, $n = 16$ in the overt retrieval group) and 45 participants studied List 2 ($n = 15$ in the restudy group, $n = 16$ in the covert retrieval group, $n = 14$ in the overt retrieval group).

In Session 1, students completed self-paced initial study trials during which they were instructed to study terms and their definitions so that on a future memory test they would be able to remember the definition when given the term. Students were presented with List 1 or List 2 twice, which was manipulated between-participants. Thus, students studied a set of 8 key-term definitions. We selected this set size to obtain multiple observations from each student while keeping performance on the final criterion test off the floor (as might occur with a larger set size). The order of presentation was randomized anew for each student, with the caveat that students studied the entire list once before the key terms were repeated.

Next, students were instructed to practice overt or covert retrieval, or to restudy the key terms and definitions. Students in the covert retrieval group saw each key term one-at-a-time (randomized anew for each student), and they were instructed to silently retrieve the definition to each term. Further, they were instructed to try their best to retrieve each definition as completely and accurately as possible. Students were given unlimited time to silently retrieve the definition for each and were instructed to click a button once retrieval was complete. Immediately after covert retrieval, students made a self-paced judgment of knowledge by responding to the prompt, "How well did you know the definition to this term?" on a scale from 0 (not at all) to 4 (perfectly). Following the judgment of knowledge, the key term and definition were re-presented to provide students with feedback (self-paced). The procedure for the overt retrieval group was identical to the covert retrieval group except that students typed their responses into a text field on the screen. The procedure for the restudy group was also the same except that they did not practice retrieval. Instead, with the key term and definition present on the screen, they were asked to type a copy of the definition into a text field on the screen. Students in all practice groups completed one block of practice retrieval or restudy, judgments of knowledge, and feedback for the 8 key-term definitions. Then, they repeated that procedure 2 additional times (i.e., 3 blocks total).

Two days later, students in all practice groups returned to complete the final criterion test. Students were presented with each key term that they had previously studied (randomized anew per student), and they were given unlimited time to type the recalled definition for each.

Data scoring. Students' initial responses during the restudy or retrieval-practice phase, as well as responses on the final criterion test, were hand scored by two independent raters who were blind to students' practice-group assignment. Each student's response was scored by identifying the number of idea units correctly recalled for each key-term definition. Definitions contained between three and five idea units. For instance, *procedural memory* contained three idea units: (1) *memory for* (2) *how to perform actions* (3) *that cannot be stated verbally*. An idea unit was counted as being present if it was a verbatim copy of the idea unit or if the idea was correctly paraphrased (Rawson & Dunlosky, 2007). The proportion of idea units correctly recalled for each definition during the retrieval-practice phase and on the final criterion test was then calculated by dividing the total number of idea units earned by the total number of idea units possible for that key term. This average was calculated separately for each rater's scores, and yielded scores that ranged from no credit to full credit for each term. Thus, each rater produced a mean proportion of correct recall for each definition for each participant. Two averages including all 8 key terms were then calculated per participant, one for each rater. The same scoring procedures were used for scoring participants' transcripts for the restudy group. Given that the analyses focused on the continuous variable at the level of participants' proportion of correct recall, we assessed inter-rater reliability using Pearson's r correlations across participants. That is, correlations were computed for raters' proportion of correct recall across all participants; a correlation was computed for each test. For all tests, correlations were significant ($ps < .001$) and high ($r = .82$ to $r = .90$), indicating that agreement between the two raters was high. Further, analyses conducted separately with each rater's scores revealed similar effects and supported the same conclusions. Accordingly, scores from the two raters were averaged for the reported analyses.

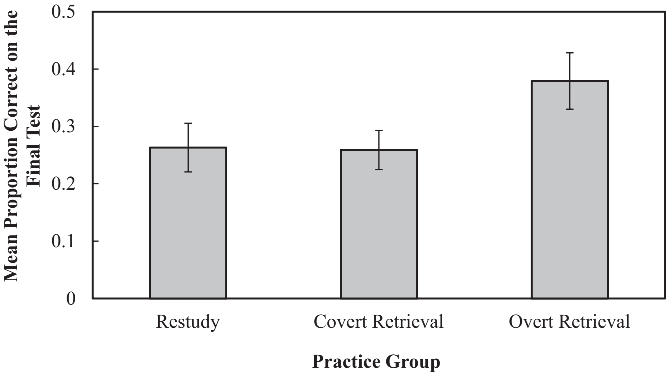

Figure 1. Mean proportion correct on the final criterion test for the three practice groups (restudy, covert retrieval, and overt retrieval) in Experiment 1. Error bars are standard errors of the mean.

Results and Discussion

Our primary interest was in recall on the final criterion test; thus, outcomes for this measure are presented first. Afterwards, we present analyses of judgments of knowledge followed by assessments of the accuracy of retrieval practice from the overt retrieval group and of restudy responses from the restudy group.

Final recall performance. Final recall performance was significantly greater for List 1 (cognitive terms; $M = .43$, $SE = .03$) than for List 2 (social terms; $M = .30$, $SE = .18$), $t(88) = 3.13$, $p = .002$, $d = .66$. Even so, list did not interact with practice group ($F < 1$), and it did not impact any other measure. Thus, lists are collapsed in all reported analyses.

As is evident from Figure 1, students' recall somewhat differed between the practice groups, $F(2, 89) = 2.85$, $p = .06$, $\eta_p^2 = .06$. Specifically, students' recall was significantly greater in the overt retrieval group than in the covert retrieval group, $t(59) = 2.19$, $p = .03$, $d = .56$. Recall was also significantly greater in the overt retrieval group than in the restudy group, $t(57) = 1.89$, $p = .03$, $d = .49$ (via a one-tailed test, as per planned comparisons relevant to standard test-enhanced learning). The covert retrieval and restudy groups did not significantly differ, $t < 1$. Thus, retention of the key-term definitions was greater when students practiced retrieval overtly relative to when they practiced retrieval covertly or restudied the terms.

Judgments of knowledge. The magnitude of students' judgments of knowledge was lower in the overt retrieval group than

Table 1
Magnitude of Judgments of Knowledge from Experiment 1 and Experiment 2

	Trial 1	Trial 2	Trial 3	Mean across Trials
Experiment 1				
Restudy	2.7 (.1)	3.1 (.1)	3.6 (.1)	3.1 (.2)
Covert retrieval	2.4 (.1)	2.9 (.1)	3.1 (.1)	2.8 (.1)
Overt retrieval	2.0 (.1)	2.2 (.2)	2.5 (.2)	2.2 (.2)
Experiment 2				
Restudy	2.6 (.1)	2.9 (.1)	3.3 (.1)	3.0 (.1)
Covert retrieval	2.6 (.1)	2.9 (.1)	3.3 (.1)	2.9 (.1)
Overt retrieval	2.0 (.1)	2.3 (.1)	2.5 (.1)	2.3 (.1)
Enhanced covert retrieval	2.5 (.1)	2.9 (.1)	3.2 (.1)	2.9 (.1)

Note. Students were asked, "How well did you know the definition to this term?" with a scale from 0 (not at all) – 4 (perfectly). Mean judgments are provided with standard errors in parentheses.

in the restudy group and the covert retrieval group (Table 1). Additionally, the magnitude of judgments in all practice groups increased across trials. These observations were supported by a 3 (Practice Group) × 3 (Trial: 1, 2, 3) mixed-factor analysis of variance (ANOVA). The main effect of practice group was significant, $F(2, 87) = 12.26$, $p < .001$, $\eta_p^2 = .22$, because judgment magnitude was lower for the overt retrieval group than either for the restudy group, $t(57) = 4.61$, $p < .001$, $d = 1.2$, or for the covert retrieval group, $t(59) = 3.08$, $p = .003$, $d = 0.79$. Judgment magnitude was also higher for the restudy group than for the covert retrieval group, $t(58) = 1.89$, $p = .06$, $d = 0.49$. Judgment magnitude also significantly increased across trials: Trial 1, $M = 2.4$, $SE = .08$; Trial 2, $M = 2.7$, $SE = .08$; Trial 3, $M = 3.1$, $SE = .08$; $F(2, 174) = 93.80$, $p < .001$, $\eta_p^2 = .52$. Finally, the interaction between practice group and trial approached significance, $F(4, 174) = 2.39$, $p = .05$, $\eta_p^2 = .05$, suggesting that increases in judgment magnitude were somewhat greater for the restudy group. However, this interaction is not meaningful (Loftus, 1978) and did not replicate in Experiment 2, so we do not discuss it further.

The magnitude of judgments of knowledge increased across trials, which is similar to effects established with judgments of learning on multi-trial learning (e.g., Koriat, Sheffer, & Ma'ayan, 2002; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Tauber & Rhodes, 2012). The increase across trials likely reflects students' beliefs (at least in part) that learning increases with additional study opportunities (Ariel, Hines, & Hertzog, 2014). More interesting, the magnitude judgments of knowledge differed among the three groups. Students engaged in overt retrieval (effortful and fraught with error) made lower judgments than those in the covert retrieval group in which students may not have engaged in exhaustive search and/or simply assumed they knew the concept due to familiarity. The students who restudied material showed the illusion of mastery seen in other research; they judged themselves to have the most knowledge, although their recall on the later test was worse than that of students who had overtly practiced retrieval and no different from that of students who covertly practiced.

Overt retrieval practice. Retrieval practice benefits later memory most when practice results in correct retrieval (Pashler, Cepeda, Wixted, & Rohrer, 2005). Thus, we assessed the accuracy of responses for the overt retrieval group during the practice trials. The accuracy of overt practice retrieval increased from Trial 1 ($M = .35$, $SE = .04$) to Trial 2 ($M = .41$, $SE = .04$), $t(29) = 4.13$, $p < .001$, $d = .29$, and from Trial 2 to Trial 3 ($M = .46$, $SE = .04$), $t(29) = 2.96$, $p = .006$, $d = .19$.

Restudy performance. To evaluate whether students in the restudy group were engaged with the task, we assessed the accuracy of restudy responses during study, which should be consistently high if they copied the definitions as directed. As expected, the accuracy of restudy responses was near ceiling for all trials: Trial 1, $M = .97$, $SE = .02$; Trial 2, $M = .97$, $SE = .02$; Trial 3, $M = .96$, $SE = .02$. Thus, in terms of sheer re-exposure to material during the initial test, the restudy group had an advantage to the retrieval-practice groups, but nonetheless retrieval practice was equally or more effective, depending on the group comparison.

Experiment 2

In Experiment 1, engaging in overt retrieval practice benefitted retention more than restudying the key terms and definitions, which is consistent with prior research establishing test-enhanced learning (e.g., Rowland, 2014). More important, covert retrieval practice did not benefit retention as much as did overt retrieval practice, which is inconsistent with previous research using paired associates (e.g., Smith, Roediger, & Karpicke, 2013). One possibility is that with complex materials such as key-term definitions, students in the covert retrieval group did not engage in full covert retrieval. As discussed in the introduction, when students are quite familiar with the term, this familiarity may short-circuit their retrieval attempt. High familiarity may lead students to believe that key terms have been learned well enough to be retrievable without actually engaging in a full-blown retrieval attempt to access the definition. This assumption was indirectly corroborated by students' judgments of knowledge in Experiment 1. Specifically, the magnitude of judgments of knowledge was higher in the covert retrieval group than in the overt retrieval group.

The major goals of Experiment 2 were to replicate the outcomes from Experiment 1 and to extend them by ensuring that students understood how to fully engage in covert retrieval. For the latter goal, a second covert retrieval group was added that received detailed instructions about how to engage in exhaustive covert retrieval. Students in this *enhanced covert-retrieval* group were instructed that they should not rely on their level of familiarity with each key term. Instead, they should attempt to retrieve the *entire* definition during each covert retrieval attempt. They also received practice trials with both covert and overt retrieval prior to beginning the experiment. By doing so, students were able to contrast the two types of retrieval and were instructed that their covert retrieval should be just like overt retrieval (except that they would not type the retrieved definitions).

Additionally, students in the enhanced covert retrieval group made a new judgment for each key term wherein they indicated how much of each definition was retrieved. Students in the overt retrieval group also made these retrieval judgments. These judgments were included both to assess the completeness of students' retrieval and to encourage them to engage in exhaustive retrieval.

Method

Design and participants. Practice group (Restudy, Covert Retrieval, Overt Retrieval, Enhanced Covert Retrieval) was a between-participants manipulation. One-hundred forty-five students from Kent State University participated in exchange for course credit and were randomly assigned to practice group. Of those students, 9 failed to return for the second session and were excluded from analyses. Thus, data from 136 students are reported ($n = 32$ in the restudy group, $n = 35$ in the covert retrieval group, $n = 34$ in the overt retrieval group, $n = 35$ in the enhanced covert retrieval group).

Materials and procedure. The materials were identical to Experiment 1, with one exception: only the cognitive psychology terms were used. We omitted the social psychology terms to simplify the experiment and because in Experiment 1 final recall performance was significantly greater for cognitive terms than for social psychology terms but still not close to ceiling level performance. The procedure for the enhanced covert retrieval group was identical to that of the covert retrieval group in Experiment 1, with a few exceptions. Prior to covert retrieval, students in the enhanced covert retrieval group were given the following instructions:

> **IMPORTANT**: During silent retrieval, **you should do your best to recall as much of the definition as you can**.
> Think about retrieval practice as if you and a friend were quizzing each other on terms for an upcoming exam in one of your classes—if your friend gave you a term, you'd try to come up with the entire definition. Or think about it like the short answer questions on a course exam, in which you are asked to write down the whole definition for a term. In both of these cases, you would try to retrieve the entire definition for each term. That is what we'd like you to do on each trial here, only silently in your head instead of out loud or on paper.
> The important thing is to really try to recall the definition. On each trial, you should not just read the key term and assume that you know it because it seems familiar. Rather, use the term as a cue to try to silently recall the definition from memory.

These instructions (i.e., both paragraphs including the bolded and underlined text) were provided only to participants in the enhanced covert retrieval group and were not provided to participants in any of the other groups. Participants in the enhanced covert retrieval group practiced both covert and overt retrieval with an example item prior to engaging in covert retrieval during study. To do so, students studied a key term and definition that would not appear on the final test (i.e., *spinal reflexes*). They were then presented with the key term alone and silently retrieved the entire definition for it. Next, they typed the definition that they silently retrieved. Only students in the enhanced covert retrieval group were then given the following instructions:

> Now that you have tried retrieval practice by silently retrieving the definition for this practice item, and by typing out what you retrieved, you probably have a better sense about how to go about retrieval practice. You may have noticed that it is hard to retrieve the definition, and it can be particularly hard when silently practicing. In order to be effective while silently retrieving you should try to retrieve as much as you would if you were required to type out everything you can remember. For the next part of the experiment you are going to practice retrieval silently. Again, **be sure to silently retrieve the entire definition for each term!**

Students in the enhanced covert retrieval group also made retrieval judgments immediately following judgments of knowledge (but prior to feedback). They responded to the prompt, "For the term below, how much of the definition did you come up with during retrieval practice?" Judgments were made on a 0%–100% scale. Retrieval judgments were self-paced.

The procedure for the restudy group and covert retrieval group were identical to Experiment 1. The procedure for the overt retrieval group was identical to Experiment 1 except they also made retrieval judgments as in the enhanced covert retrieval group.

Data scoring. Students' responses during the restudy or retrieval-practice phase and on the final criterion test were scored as in Experiment 1. Inter-rater reliability was assessed for each measure via Pearson's r correlations. In all cases correlations were significant ($ps < .001$) and high ($r = .83$ to $r = .94$). As in Experiment 1, agreement between the two raters was high and effects were maintained when analyses were conducted separately with each rater's scores; thus, scores from the two raters were averaged for the reported analyses.

Results and Discussion

Final recall performance. As is evident from Figure 2, students' recall was numerically greater in the overt retrieval group relative to the other groups, although the practice groups did not

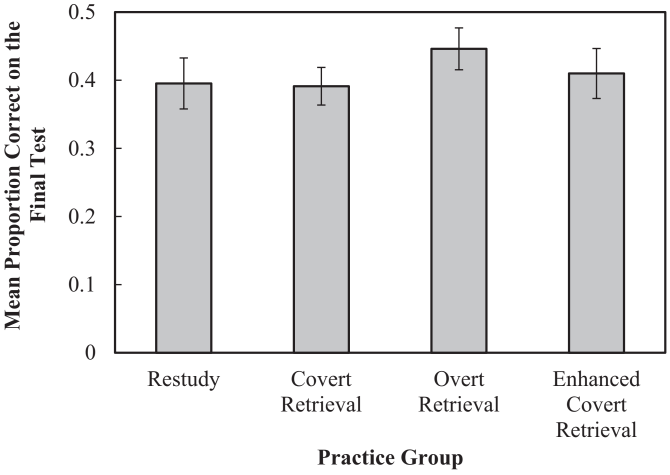

Figure 2. Mean proportion correct on the final criterion test for the four practice groups (restudy, covert retrieval, overt retrieval, and enhanced covert retrieval) in Experiment 2. Error bars are standard errors of the mean.

Table 2
Continuously Cumulating Meta-Analyses (CCMAs) for Experiment 1 and Experiment 2

Experiment	Mean Diff	S_{pooled}	t	p	Cohen's d	Z
CCMA for covert retrieval and overt retrieval groups						
Experiment 1	.11	.20	2.19	.033	.56	2.14
Experiment 2	.06	.17	1.36	.180	.33	1.34
CCMA results				.015	.43	2.46
CCMA for restudy and overt retrieval groups						
Experiment 1	.10	.21	1.89	.064	.49	1.85
Experiment 2	.05	.19	1.07	.287	.26	1.07
CCMA results				.040	.37	2.06

Note. The homogeneity test was nonsignificant for the CCMA for covert retrieval and overt retrieval groups (top of table), $Q(1) = .43$, $p = .51$, $I^2 = 0.00$. The homogeneity test was nonsignificant for the CCMA for restudy and overt retrieval groups (bottom of table), $Q(1) = .38$, $p = .54$, $I^2 = 0.00$.

differ, $F < 1$. Thus, in Experiment 2 students' recall did not significantly differ between the overt retrieval group and the covert retrieval group, $t(67) = 1.35$, $p = .09$, one-tailed, $d = .33$.

To provide the best estimate of the magnitude of the overt retrieval-practice effect, we conducted a *continuously cumulating meta-analysis* (CCMA) as recommended by Braver et al. (2014). Specifically, a CCMA was conducted comparing final recall performance for the overt retrieval-practice groups versus covert retrieval-practice groups across the two experiments (see top portion of Table 2). Final recall performance was higher following overt retrieval versus covert retrieval (Experiment 1, $M_{\text{diff}} = .11$, $S_{\text{pooled}} = .20$; Experiment 2, $M_{\text{diff}} = .06$, $S_{\text{pooled}} = .17$; pooled $d = 0.43$, 95% CI [0.09, 0.78]). Thus, results from the CCMA support the conclusion that retention is enhanced more by overt retrieval practice than by covert retrieval practice (or by restudying).

Also, in Experiment 2 students' recall did not significantly differ between the overt retrieval group and the restudy group, $t(64) = 1.07$, $p = .14$, $d = .26$, and no other practice group differences were significant, $ts < 1$. Even so, a CCMA comparing final recall performance for the overt retrieval-practice groups versus the restudy groups across the two experiments revealed that final recall performance was greater following overt retrieval versus restudy (see bottom portion of Table 2; Experiment 1, $M_{\text{diff}} = .10$, $S_{\text{pooled}} = .21$; Experiment 2, $M_{\text{diff}} = .05$, $S_{\text{pooled}} = .19$; pooled $d = 0.37$, 95% CI [0.02, 0.72]). Thus, the overt retrieval groups outperformed the restudy groups on the final criterion test (replicating the modal finding in research on test-enhanced learning; for reviews, see Dunlosky et al., 2013; Roediger & Butler, 2011; Rowland, 2014).

Judgments of knowledge. As is evident from Table 1, the magnitude of students' judgments of knowledge was lower in the overt retrieval group than in the other three practice groups. As in Experiment 1, judgment magnitude in all practice groups increased across trials. These observations were supported by a 3 (Practice group) × 3 (Trial) mixed-factor ANOVA. The main effect of practice group was significant, $F(3, 131) = 11.43$, $p < .001$, $\eta_p^2 = .21$. Follow-up tests revealed that judgment magnitude was significantly lower for the overt retrieval group than for the restudy group, $t(63) = 4.57$, $p < .001$, $d = 1.1$, the covert retrieval group, $t(66) = 4.52$, $p < .001$, $d = 1.1$, and the enhanced covert retrieval group, $t(66) = 4.08$, $p < .001$, $d = 0.99$. The magnitude of students' judgments in the restudy group, covert retrieval group, and enhanced covert retrieval groups did not differ, $ts < 1$. The main effect of trial was also significant (Trial 1, $M = 2.4$, $SE = .05$; Trial 2, $M = 2.8$, $SE = .06$; Trial 3, $M = 3.1$, $SE = .06$), $F(2, 262) = 128.7$, $p < .001$, $\eta_p^2 = .50$. The interaction between practice group and trial was not significant, $F < 1$.

As in Experiment 1 the magnitude of judgments increased across trials, which parallels a similar effect with judgments of learning (e.g., Koriat et al., 2006; Koriat et al., 2002; Tauber & Rhodes, 2012), and is likely attributable to students' beliefs that more study opportunities benefit learning (Ariel et al., 2014). Further, overt retrieval practice was associated with lower judgments of knowledge than was covert retrieval practice or restudy. Relative to the other groups, students who practiced overt retrieval may have been more likely to engage in exhaustive retrieval, and when such attempts failed or were incomplete they may have reduced their judgments.

Retrieval judgments. The magnitude of judgments about the amount of retrieval during practice increased across trials, as expected. More importantly, they were lower in the overt retrieval group than in the enhanced covert retrieval group. These observations were supported by a 2 (Practice group: overt retrieval, enhanced covert retrieval) × 3 (Trial) mixed-factor ANOVA. The main effect of practice group was significant, $F(1, 66) = 14.20$, $p < .001$, $\eta_p^2 = .18$, with judgment magnitude being lower for the overt retrieval group ($M = 52.9$, $SE = 3.4$) relative to the enhanced covert retrieval group ($M = 70.7$, $SE = 3.3$). The main effect of trial was also significant (Trial 1, $M = 53.4$, $SE = 2.4$; Trial 2, $M = 61.3$, $SE = 2.6$; Trial 3, $M = 70.8$, $SE = 2.6$), $F(2, 132) = 62.30$, $p < .001$, $\eta_p^2 = .49$, and the interaction between practice group and trial was not significant, $F(2, 132) = 2.74$, $p = .07$, $\eta_p^2 = .04$. Thus, relative to a standard overt retrieval group, students in a covert retrieval group who were provided with instructions about practicing covert retrieval and with practice items prior to retrieval practice judged that their retrieval attempts were more complete.

Overt retrieval practice. As in Experiment 1, the accuracy of overt retrieval practice significantly increased from Trial 1 ($M = .42$, $SE = .03$) to Trial 2 ($M = .47$, $SE = .03$), $t(32) = 3.68$, $p = .001$, $d = 0.33$, and from Trial 2 to Trial 3 ($M = .53$, $SE = .04$), $t(32) = 3.89$, $p < .001$, $d = 0.31$. Thus, students' overt retrieval improved with trial experience.

Please cite this article in press as: Tauber, S. K., et al. Does Covert Retrieval Benefit Learning of Key-Term Definitions? *Journal of Applied Research in Memory and Cognition* (2016), http://dx.doi.org/10.1016/j.jarmac.2016.10.004

Restudy performance. As in Experiment 1, the accuracy of restudy responses was near ceiling for all trials (Trial 1, $M = .99$, $SE = .004$; Trial 2, $M = .99$, $SE = .008$; Trial 3, $M = .99$, $SE = .004$). Therefore, students in the restudy group were engaged in the restudy task.

General Discussion

The current experiments demonstrated that the response format of retrieval practice on an initial test does influence final recall performance when students learn definitions of key terms. Specifically, retention is enhanced more by overt than covert retrieval practice. This effect was statistically significant in Experiment 1 and the same trend was evident in Experiment 2. Such variability would be expected (e.g., Stanley & Spence, 2014); thus, outcomes were evaluated with a CCMA analysis (Braver et al., 2014), which established a moderate effect size (pooled $d = 0.43$). Why was learning enhanced more by overt than covert retrieval practice? Why did covert retrieval practice not help at all relative to a restudy control? Overt retrieval practice involves an explicit written, typed, or oral response, which encourages students to engage in exhaustive retrieval so they have an answer to provide. By contrast, covert retrieval practice does not involve an explicit response, and with complex material students may sometimes avoid an exhaustive retrieval attempt especially for terms students judge to be already known well.

The suggestion that students may avoid exhaustive retrieval with covert retrieval practice is indirectly supported by students' metacognitive judgments. Learners frequently use experiences during retrieval as a basis for their metacognitive judgments (e.g., Benjamin, Bjork, & Schwartz, 1998). Thus, if covert and overt retrieval practice support similar engagement in retrieval, then the magnitude of students' judgments of knowledge should not differ because they would be using equivalent retrieval experiences as a basis for their judgments. Contrary to this prediction, the overt retrieval group provided significantly lower judgments of knowledge relative to the other groups. This reduced confidence may have been due to failed retrieval or only partial retrieval of some definitions. Such retrieval failures may occur less often with covert retrieval practice because students may use their relatively high familiarity with a portion of key terms to circumvent retrieval practice, thus reflecting a metacognitive illusion that those items were well learned (e.g., Son & Metcalfe, 2005; for a review see Finn & Tauber, 2015).

Differences in the dynamics of retrieval due to response mode suggest that covert retrieval practice will benefit learning in some contexts. When retrieval demands are high because students need to retrieve multiple units of information, learning is more likely to be enhanced by overt than covert retrieval practice. However, when retrieval demands are low such as when students only need to retrieve a single word response (such as with word pairs), presentation of the cue presumably triggers a retrieval of the response (Craik et al., 1996), so that both covert and overt practice would tend to yield exhaustive retrieval attempts. In such cases, learning would be expected to be equally enhanced with overt and covert retrieval (e.g., Putnam & Roediger, 2013). As well, differences among students may play an important role in the effectiveness of covert retrieval. For instance, more conscientious students may engage in more exhaustive covert retrieval practice relative to less conscientious students. Exploring the role of such individual differences will be an important direction for future research on the efficacy of covert retrieval practice for benefitting learning.

Given that at least some students report using retrieval practice to study (e.g., Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007), an important direction for researchers will be to focus on methods to increase the effectiveness of covert retrieval practice. Unfortunately, outcomes from Experiment 2 revealed that final recall was equivalent between a standard covert retrieval group and an enhanced covert retrieval group who was provided with (a) detailed instructions on the importance of retrieving the entire definition for each term, (b) a warning to avoid relying on their level of familiarity of each term, and (c) practice with overt and covert retrieval. Further, recall was lower for both covert retrieval-practice groups than the overt retrieval-practice group. This pattern suggests that as often as possible students should engage in overt retrieval practice.

In Experiment 2, recall on the final test did not statistically differ between the overt retrieval group and the restudy group. Even though a CCMA revealed that final recall performance was greater following overt retrieval than restudy (pooled $d = 0.37$; see bottom portion of Table 2), the nonsignificant effect in Experiment 2 was surprising because a wealth of research has established robust test-enhanced effects on retention (for a review see, Rowland, 2014). One possibility is that the restudy group spent more time during the practice phase relative to the overt retrieval group, with the additional time obscuring the test benefits. To explore this possibility, we evaluated self-paced reaction times during restudy and overt retrieval practice for Experiment 2. Reaction times were measured by recording the number of seconds from the onset of each stimulus (i.e., key term and definition for the restudy group and the key term alone for the overt retrieval group) to when participants clicked a button indicating that they were finished. Reaction times were then averaged per trial, and results demonstrated that they tended to decreased across them, $F(2, 126) = 9.7$, $p < .001$, $\eta_p^2 = .13$. Specifically, mean reaction times significantly decreased from Trial 1 ($M = 33.0$ s, $SE = 1.5$) to Trial 2 ($M = 29.7$ s, $SE = 1.6$; $p = .01$), and marginally decreased from Trial 2 to Trial 3, $M = 28.1$ s, $SE = 1.4$; $t(64) = 1.9$, $p = .06$, $d = .13$. More important, they did not significantly differ between the restudy group ($M = 28.1$ s, $SE = 2.0$) and the overt retrieval group ($M = 32.5$ s, $SE = 1.9$; $F(1, 63) = 2.6$, $p = .11$, $\eta_p^2 = .04$, and trial did not interact with group ($F < 1$). Thus, the lack of test-enhanced learning in Experiment 2 remains mysterious, and unknown sources of variance may be responsible for it, though variability in the magnitude of any effect is to be expected.

To conclude, students' learning is typically enhanced by practicing retrieval during study relative to restudying the information, but it can matter *how* students practice retrieval. Students' learning of large units of information is enhanced more by overt than covert retrieval practice. In contrast with covert retrieval, overt retrieval practice is more likely to support exhaustive retrieval attempts and to discourage reliance on a general level of familiarity with the to-be-learned information. Thus, until effective interventions have been established that increase the likelihood of exhaustive covert retrieval, we recommend that students use overt retrieval practice when learning key-term definitions.

Author Contributions

All authors developed the study concept and contributed to the study design. Testing and data collection were performed by Tauber. Data scoring was overseen by Witherby. Tauber and Witherby conducted analyses and interpreted them. Tauber drafted the manuscript and all authors provided feedback and revisions. All authors approved the final version of the manuscript for submission.

Conflict of Interest Statement

The authors declared no conflicts of interest with respect to the authorship or the publication of this article.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876. http://dx.doi.org/10.1002/acp.1391

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. *Journal of Memory and Language*, *75*, 181–198. http://dx.doi.org/10.1016/j.jml.2014.06.003

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

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333–342. http://dx.doi.org/10.1177/1745691614529796

Craik, F. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*, 159–180. http://dx.doi.org/10.1037/0096-3445.125.2.159

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. http://dx.doi.org/10.1177/1529100612453266

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*, 567–586. http://dx.doi.org/10.1007/s10648-015-9313-7

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134. http://dx.doi.org/10.3758/s13423-011-0181-y

Jönsson, F. U., Kubrik, 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

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 595–608. http://dx.doi.org/10.1037/0278-7393.32.3.595

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*, 147–162. http://dx.doi.org/10.1037/0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224.

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

Lishner, D. A. (2015). A concise set of core recommendations to promote the dependability of psychological research. *Review of General Psychology*, *19*, 52–68. http://dx.doi.org/10.1037/gpr0000028

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*, 312–319. http://dx.doi.org/10.3758/BF03197461

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. http://dx.doi.org/10.1177/1745691614528215

Orlando, V. P., & Hayward, K. G. (1978). A comparison of the effectiveness of three study techniques for college students. In P. D. Peterson, & J. Hansen (Eds.), *Reading: Disciplined inquiry in process and practice* (pp. 242–245). Clemson, SC: National Reading Conference.

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.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. http://dx.doi.org/10.1177/17456916124634

Putnam, A. L., & Roediger, H. L., III. (2013). Does the 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

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*, 283–302. http://dx.doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*, 559–579. http://dx.doi.org/10.1080/09541440701326022

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

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

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. http://dx.doi.org/10.1177/1745691613514755

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

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, *33*, 1116–1129. http://dx.doi.org/10.3758/BF03193217

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305–318. http://dx.doi.org/10.1177/1745691614528518

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

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, *10*, 390–399. http://dx.doi.org/10.1177/1745691614564879

Received 11 July 2016;
accepted 18 October 2016
Available online xxx