Educational Psychology Review (2023) 35:98
https://doi.org/10.1007/s10648-023-09809-2

INTERVENTION STUDY

# Spaced Retrieval Practice: Can Restudying Trump Retrieval?

**Philip A. Higham[1] · Greta M. Fastrich[1] · Rosalind Potts[2] · Kou Murayama[3,4] · Jade S. Pickering[1] · Julie A. Hadwin[5]**

Accepted: 31 August 2023


## Abstract

We investigated spaced retrieval and restudying in 3 preregistered, online experiments. In all experiments, participants studied 40 Swahili–English word pair translations during an initial study phase, restudied intact pairs or attempted to retrieve the English words to Swahili cues twice in three spaced practice sessions, and then completed a final cued-recall test. All 5 sessions were separated by 2 days. In Experiment 1, we manipulated the response format during retrieval (covert vs. overt) and the test list structure (blocked vs. intermixed covert/overt retrieval trials). A memory rating was required on all trials (retrieval: "Was your answer correct?"; restudy: "Would you have remembered the correct translation?"). Response format had no effect on recall, but surprisingly, final test performance for restudied items exceeded both the overt and covert retrieval conditions. In Experiment 2, we manipulated the requirement to make a memory rating. If a memory rating was required, final test restudy performance exceeded retrieval performance, replicating Experiment 1. However, the pattern was descriptively reversed if no rating was required. In Experiment 3, the memory rating was removed altogether, and we examined recall performance for items restudied versus retrieved once, twice, or thrice. Performance improved with practice, and retrieval performance exceeded restudy performance in all conditions. The reversal of the typical retrieval practice effect observed in Experiments 1 and 2 is discussed in terms of theories of reactivity of memory judgments.

**Keywords** Spaced retrieval practice · Testing effect · Spacing effect · Distributed learning · Reactivity · Successive relearning · Spaced restudying

This article is part of the Topical Collection on Test-Enhanced Learning and Testing in Education: Contemporary Perspectives and Insights.

✉ Philip A. Higham
higham@soton.ac.uk

[1] School of Psychology, University of Southampton, Highfield, Southampton SO17 1BJ, UK

[2] Division of Psychology and Language Sciences, University College London, London, England

[3] School of Psychology, University of Reading, Reading, England

[4] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

[5] School of Education, University of Birmingham, Birmingham, England

Published online: 26 September 2023

## Introduction

A common goal of education is to promote long-term retention of learned material. One learning technique that helps learners achieve this goal is *spaced practice* (spreading learning over multiple, temporally spaced sessions). When compared to *massed practice* (i.e., learning which is concentrated in a single session), long-term retention is typically much better if it is spaced, a finding referred to as the *spacing effect* (see Cepeda et al., 2006 for a review). A second learning technique that is also highly effective for long-term retention is *retrieval practice* (self-testing or taking practice tests on learned material). Indeed, in a review of ten learning techniques, Dunlosky et al. (2013) concluded that these two techniques had the highest utility.

While there have been many studies spanning decades demonstrating the efficacy of spaced (vs. massed) practice in psychological science, the type of practice used in many of these studies has been repeated, spaced presentations occurring within a single learning session. That is, information is typically presented for restudy across spaced learning sessions with no retrieval requirement and the spacing gaps between successive presentations are only minutes long. However, several studies have shown that retrieval practice can be combined with much longer spacing gaps to produce a particularly effective learning technique (see Carpenter et al., 2022 and Latimier et al., 2021, for reviews). For example, Larsen et al. (2013) asked medical students to repeatedly practice retrieving content from a teaching session once a week for four weeks. Compared to restudying the same content weekly over the same period, students performed much better on a free-recall clinical application test of the material administered six months later. Thus, although spaced presentations produce better retention than massed ones, spaced retrieval practice confers even greater benefits, particularly if the spacing gaps are long.

## Successive Relearning

One type of spaced retrieval practice that has incorporated long spacing gaps and retention intervals like those used in Larsen et al.'s (2013) research is referred to as successive relearning (Bahrick, 1979; Bahrick et al., 1993; Bahrick & Hall, 2005; Higham et al., 2022; Janes et al., 2020; Rawson & Dunlosky, 2011, 2013). In typical successive relearning experiments, participants start by learning some material in an initial study session, for example, English translations of foreign vocabulary. Following an inter-session interval (ISI), which is typically two days or more, they complete a learning practice session (LPS) in which they attempt to retrieve the English translations when cued with the foreign words. The participants' goal during the session is to correctly retrieve each English translation to a preset *criterion level*, which may be one successful retrieval or more. Once the criterion is reached for a given item, it is dropped from further practice in that session. Translations not yet retrieved to the criterion level are presented again later within the same session,

and if the retrieval attempt is incorrect, corrective feedback is provided.[1] This process of a retrieval attempt followed by feedback is repeated until all items reach the criterion and are dropped out, thereby ending the LPS (i.e., mastery learning). Participants repeat the whole LPS process several more times at spaced intervals (e.g., two days) before taking a final memory test. Thus, the main characteristics that separate successive relearning from other research on spaced retrieval practice is the requirement to achieve mastery during the LPSs, which is facilitated with dropout methodology.

Like spaced retrieval practice, the literature on successive relearning has shown that it is highly effective for long-term learning (e.g., see Rawson & Dunlosky, 2022, for a review; although also see Rawson et al., 2020). For example, Rawson et al. (2018) found that after three LPSs, each separated by one week, participants remembered 80% of the to-be-remembered material one week after the last session. After a fourth LPS, participants still remembered 77% of the material three weeks later, suggesting successive relearning mitigated forgetting. Moreover, Higham et al. (2022) showed that successive relearning also had beneficial effects on other learning outcomes. For example, multiple successive relearning sessions were associated with higher self-reported subjective feelings of mastery and attentional control and lower anxious affect.

Because learners achieve mastery at different rates (e.g., some learners may require more trials to reach criterion than other learners), the need to achieve mastery in successive relearning studies has methodological implications. That is, unlike the Larsen et al. (2013) study on spaced retrieval practice described earlier, successive relearning research usually does not include a restudy control condition (although see below). One reason for omitting the restudy control condition is that the different learning rates across participants make it difficult to decide how long or how many times information must be restudied to match exposure times between the relearning and restudy conditions. Instead, researchers have typically compared successive relearning performance against a "business as usual" baseline, or between successive relearning conditions that varied the criterion level, number of LPSs, and/or number of ISIs. The omission of a restudy control condition is potentially problematic in that there is no control for the level of exposure to the to-be-remembered material. For example, if final recall following successive relearning is higher after three LPSs rather than two, is that difference attributable to more spaced retrieval or more spaced exposure?

Higham et al. (2022) addressed this question in a university classroom intervention with first-year psychology students. They compared memory performance between a successive relearning condition and a restudy control condition matched for exposure time. The results showed that recall performance was greater in the successive relearning condition than the restudy condition on a final recall test. However, some items in the final recall test were only learned once during an initial study episode (lecture) and were not practiced during the subsequent LPSs. Recall

[1] In some cases, corrective feedback is provided even when the target is correctly retrieved to maintain consistency across the trials (e.g., Higham et al., 2022).

performance for relearned and restudied items was far superior to these studied-once items, suggesting that spaced exposure in the form of restudying was also beneficial for learning, consistent with typical demonstrations of the spacing effect. Unfortunately, the applied nature of the study did not allow for full item counterbalancing, compromising between-item comparisons. Thus, the benefit of relearned items and/or the restudied items over unpracticed items could be partly attributed to the unpracticed items being harder to learn. Hence, there is a need for further research into how restudied material is learned in a spaced retrieval context with multiple LPSs, long ISIs, and long retention intervals.

## Overt Versus Covert Retrieval

In addition to the question about the effect of exposure duration, another question that has received very little attention outside of single-session retrieval practice research is whether the response format during retrieval practice is important. Specifically, is it necessary for participants to make an overt response during retrieval practice to experience a boost to later test performance or is covert retrieval sufficient? This question is important from both a theoretical and practical perspective. In terms of theory, some research has shown that *producing* items during encoding by saying, writing, or typing them yields better memory performance compared to silent reading, the so-called *production effect* (see MacLeod & Bodner, 2017). According to the one dominant account of this effect, producing items during encoding enhances their distinctiveness in memory relative to unproduced items. Although the production effect pertains to encoding differences, the process of retrieving items during learning practice doubles as both a retrieval event and an encoding event. Hence, it is conceivable that production in the form of overt retrieval during practice would enhance the distinctiveness of the produced items and lead to better memory performance than covert retrieval. On the other hand, response format was explicitly considered in Tulving's (1983) *general abstract processing system* (GAPS) framework. He argued that in terms of retrieval from episodic memory, "…'thinking about' or reviewing the event in one's mind—produces consequences comparable to those resulting from responses to explicit questions" (p. 47). Thus, under the assumptions of the GAPS framework, response format should not matter.

In terms of practical considerations, students may attempt to engage in retrieval practice in a setting where they must be silent (e.g., covertly retrieving answers to an instructor's questions in a classroom setting or using flashcards in a library). Additionally, some smartphone apps designed to facilitate learning through retrieval practice (e.g., *RememberMore* and *MosaLingua*) do not require users to make overt responses, only to think of them. Also, apps such as *Anki* allow users to create their own flashcards or download sets from others. With these apps, users simply indicate when they want to see the flashcard again by selecting a particular option (e.g., Again, Hard, Good, and Easy), with each option determining when to show the card again. Therefore, it is important to establish whether these students will reap the benefits of retrieval practice to the same extent as students who are retrieving information in a setting where overt production is more likely.

While covert retrieval has been known to be an effective learning tool for some time (e.g., Gates, 1917), modern research comparing overt versus covert retrieval formats during retrieval practice has produced inconsistent results. Some studies have found that overt retrieval produces larger testing effects than covert retrieval (e.g., Jönsson et al., 2014; Putnam & Roediger, 2013, Experiment 2; Sundqvist et al., 2017, Experiments 3 & 4; Tauber et al., 2018), some have found the opposite (e.g., Smith et al., 2013, Experiment 4), and others have found no difference (Putnam & Roediger, 2013, Experiments 1 & 3; Smith et al., 2013, Experiments 1-3; Sundqvist et al., 2017, Experiments 1 & 2). Sundqvist et al. (2017) conducted a meta-analysis on 13 published experiments that directly compared the effects of overt and covert retrieval practice. They concluded that there was a slight advantage for overt retrieval, but the effect size was negligible (Cohen's $d < 0.20$).

One factor that has been shown to moderate the covert/overt difference is the extent to which participants engage in complete covert retrieval when instructed to do so. Only the participants themselves can assess the completeness of the covert retrieval attempt, or indeed, whether a covert attempt was made at all. Consequently, some cases where an overt advantage was obtained might be due to differences in the rate or completeness of retrieval attempts in each condition rather than differences in response format per se. Researchers have used different strategies to encourage complete retrieval attempts in covert retrieval conditions. For example, Tauber et al. (2018) emphasized the importance of complete retrieval using enhanced instructions. Others have encouraged retrieval on covert trials by making it impossible to predict during a retrieval time window whether the target item needs to be reported when the time has elapsed (e.g., Putnam & Roediger, 2013; Sundqvist et al., 2017). For example, for each trial during the retrieval practice phase in Putnam and Roediger (2013, Experiment 3), participants were first shown a cue word with question marks next to it for 4 s and instructed to bring the target to mind. On overt retrieval trials, participants saw "Recall!" at the end of the interval, and they were required to say the retrieved target word aloud (or say nothing if they could not retrieve the item). Conversely, on covert trials, participants saw "Did you remember?", and they were required to respond "yes" or "no" but not actually say the target word aloud. Critically, the overt and covert retrieval trials were randomly intermixed such that participants could not predict during the 4 s window whether they would be required to provide an overt response. Thus, participants were likely to retrieve targets on every trial just in case an overt response was necessary. Under these circumstances, overt and covert retrieval practice yielded substantial and comparable testing effects.

Although the effect of response format has proven to be inconsistent in prior research, it is noteworthy that none of this research has involved multiple, spaced LPSs, which is the hallmark of spaced retrieval practice research. Potentially, retrieval practice of the items over these LPSs could moderate the effect of response format. For example, participants might become better at complete covert retrieval with practice. If covert retrieval is initially incomplete, memory performance may suffer and not measure up to performance with overtly retrieved items. In response, participants may adopt a different (and improved) covert retrieval strategy during later LPSs that more closely matches their behavior on overt retrieval trials, thereby closing the gap in performance. Thus, there is a need for research on response format

in spaced retrieval practice paradigms to determine the relative effectiveness of overt versus covert retrieval practice in learning and memory.

## Overview of the Experiments

The overarching aims of the current research were twofold. First, we aimed to determine the effectiveness of *spaced restudying* in promoting durable learning, and to compare it to the efficacy of *spaced retrieval practice*, in an experimental context where the relevant comparisons are not compromised by item differences. Second, we aimed to determine whether response format (i.e., overt versus covert responses) influences memory in a spaced retrieval practice context where there are multiple, spaced LPSs, and a delayed final test.

In three preregistered experiments, online participants first studied a list of Swahili-English word pairs. Two days later, participants returned for the first of three LPSs during which they restudied some intact pairs or attempted to retrieve the English words to Swahili cue words for other pairs. On retrieval trials, the retrieval attempt was followed by feedback in the form of presentation of the intact pair. The ISI for the LPSs in all cases was two days. Finally, two days after the last LPS, all participants completed a cued-recall test that consisted of items that were studied once (during the initial study phase), retrieved during the LPSs (overtly or covertly), or restudied during the LPSs. For this final test, the Swahili words were shown individually, and participants were required to recall the English translations.

Following the feedback in the LPSs, participants in some conditions of the three experiments we report were asked to assess the accuracy of their own responses (see later for details). Specifically, they were asked "Was your answer correct?" (covert and overt retrieval conditions) or "Would you have remembered the correct translation?" (restudy condition) after the intact pair was presented. A 0-100 slider scale was provided to make the rating. These ratings ensured that participants in all three conditions made some form of overt response, but only participants in the overt retrieval condition responded overtly with the to-be-remembered information. However, as will become clear, the ratings unexpectedly affected memory performance in interesting ways, which became our focus in the later experiments.

Our research design is like that used in successive relearning research in that participants were given more than one opportunity to retrieve the items within each LPS. However, it differs from successive relearning research in that we did not drop correctly recalled items from the practice sessions once they had been correctly recalled to a specific criterion. Indeed, we did not set a criterion that had to be met during the LPSs and so within-LPS mastery was not guaranteed. Instead, each item was presented twice in every LPS for a fixed amount of time in both the restudy and retrieval conditions. We adopted this design primarily so that we could carefully control exposure between the restudy and retrieval conditions. Hence, our research design is a spaced retrieval practice design with repeated opportunities to practice the items during the LPSs.

Our preregistered hypotheses pertaining to memory performance on the final cued-recall test in all experiments are shown in Table 1. In Experiment 1, we investigated response format (covert vs. overt retrieval) by manipulating the list

**Table 1** Preregistered hypotheses and associated analyses for Experiments 1–3

| Hypothesis | Analysis |
|---|---|
| Experiment 1 | |
| *Hypothesis 1a*: Overt > covert > restudy > study once for final cued-recall accuracy | 4 × 3 mixed ANOVA on final cued-recall accuracy with a within-subjects factor of condition (overt, covert, restudy, study once) and a between-subjects factor of group (blocked 50%, random 25%, random 75%) with follow-up *t*-tests as needed |
| *Hypothesis 1b*: Overt vs. covert difference in final cued-recall accuracy greater in the blocked group than the random groups | 2 (condition: overt, covert) x 2 (group: blocked 50%, collapsed random) mixed ANOVA on final cued-recall accuracy with follow-up *t*-tests as needed |
| *Hypothesis 1c*: Overt vs. covert difference in final cued-recall accuracy smaller in the random 75% than the random 25% group. | 2 (condition: overt, covert) x 2 (group: random 25%, random 75%) mixed ANOVA on final cued-recall accuracy with follow-up *t*-tests as needed. |
| Experiment 2 | |
| *Hypothesis 1*: Restudy with memory rating > overt retrieval with memory rating & overt retrieval without memory rating > restudy without memory rating for final cued-recall accuracy. | 2 x 2 repeated measures ANOVA on final cued-recall accuracy with the factors of practice type (overt retrieval, restudy) and of memory rating during feedback (present, absent) with follow-up *t*-tests as needed. |
| Experiment 3 | |
| *Hypothesis 1*: Both restudy and retrieval performance on the final cued-recall test will improve with greater practice frequency, but there will be no difference between the restudy and retrieval conditions for items practiced three times (i.e., STTT = SRRR). | 3 (practice frequency: 1, 2, 3) X 2 (condition: restudy, retrieval) mixed ANOVA on recall accuracy with follow-up *t*-tests as needed. |

*S* = initial study, *R* = restudy, *T* = test, *ANOVA* = Analysis of Variance. Follow-up *t*-tests were conducted as needed if the preregistered ANOVA yielded a significant interaction and/or a main effect of a factor with more than two levels.

structure of the LPSs (e.g., Sundqvist et al., 2017). Specifically, there were three types of items in the LPSs – restudy, overt retrieval, and covert retrieval – that occurred in different proportions between groups (see Sundqvist et al., 2017). We also varied the sequencing of the items such that presentation of the item types was blocked for one group but randomized for two others. We anticipated that participants who engaged in learning practice when there was blocked sequencing would be least likely to engage in complete covert retrieval. Under these circumstances, participants would know that an overt response was not required during the covert retrieval block. Conversely, participants who practiced learning items in a list with random sequencing and a high proportion of overt retrieval trials would be the most likely to engage in complete covert retrieval.

To foreshadow the results, no differences were observed between the overt and covert retrieval conditions in any of the groups in Experiment 1. However, surprisingly, we found that the restudied items were remembered better on the final

recall test than the items from any other condition, including the retrieval conditions, and this restudy advantage occurred in all three groups.

In Experiments 2 and 3, we therefore focused on exploring this pattern of results. Specifically, we hypothesized that the memory rating was producing *reactivity* akin to Judgment of Learning (JOL) reactivity (see Double et al., 2018 for a meta-analysis). JOL reactivity is sometimes positive, boosting later memory (e.g., Soderstrom et al., 2015), sometime negative, reducing later memory (e.g., Mitchum et al., 2016), and sometimes has no effect on memory (e.g., Ariel et al., 2021; Janes et al., 2018; Tauber & Witherby, 2019). Typically, if positive reactivity is observed in this research, it occurs with related English word pairs. It is seldom observed with unrelated word pairs (although see Rivers et al., 2021). However, to our knowledge, no one has investigated reactivity with English translation pairs. Therefore, in Experiment 2, we tested the hypothesis that the post-feedback rating during the LPSs may have interfered with retrieval (negative reactivity) and/or boosted restudy performance (positive reactivity). This hypothesis was tested in Experiment 2 by manipulating the requirement to make these ratings. The results suggested that both interference on retrieval trials and facilitation on restudy trials contributed to the superiority of the restudy condition in Experiment 1. Finally, in Experiment 3, we removed the memory rating altogether and compared restudy and retrieval performance when items were restudied or retrieved once, twice, or thrice across the three LPSs. The results showed that retrieval was superior to restudying regardless of the number of times the items were encountered across the LPSs.

## Experiment 1

Experiment 1 included three groups to manipulate the sequencing (blocked vs random order) and the proportion of overt retrieval trials in the LPSs. Three preregistered hypotheses pertaining to memory performance on the final cued-recall test are shown in Table 1 (Hypotheses 1a, 1b, and 1c). We expected that complete retrieval may not occur in some covert conditions, thereby limiting the retrieval practice benefit. Hypothesis 1a, therefore, was that overall accuracy on a final cued-recall test would be highest for overtly retrieved items followed by covertly retrieved items, then restudied items, and finally study-once items. We further anticipated that covert retrieval would be most limited in the blocked condition. Therefore, Hypothesis 1b was that the difference in accuracy for overtly versus covertly retrieved items on the final cued-recall test would be larger if items were retrieved in blocked versus random order during the LPSs. Finally, we expected that, when the proportion of overt trials is high, then the likelihood that participants will engage in complete retrieval on covert trials will increase. Consequently, both overtly and covertly retrieved items will benefit equally from retrieval practice during the LPSs. Therefore, Hypothesis 1c was that the benefit of overt versus covert retrieval with randomly intermixed pairs would decrease as the proportion of overt pairs in the list increased.[2]

[2] We preregistered a second hypothesis for Experiment 1 regarding the individual difference measures for a manuscript in preparation. Before learning, participants completed several questionnaires measuring academic efficacy, attentional control, generalized anxiety, and intolerance of uncertainty. During

## Method

### Participants

All data were collected online using Prolific (https://www.prolific.co/) and Gorilla (https://gorilla.sc/; Anwyl-Irvine et al., 2020) and participants were reimbursed £5 per hour for all sessions they completed. Participants were initially recruited using a prescreening survey on Prolific which determined their willingness to participate in a series of study sessions on language learning. The prescreening survey also assessed participants against the inclusion/exclusion criteria for the study. Specifically, participants must have been 18–60-year-old native English speakers with no knowledge of Swahili or Arabic (which overlaps significantly with the Swahili language). On Gorilla, we used additional screening to ensure that participation was on a laptop or desktop computer. Participants who met these criteria were invited into Session 1 and those that completed each subsequent session were invited to follow up sessions (five sessions in total: initial study, three LPSs, and one final memory test).

Our total sample size was constrained by pragmatic concerns (e.g., budget), but we aimed to collect data from 90 participants in each of three groups ($n = 270$ participants in total). To compensate for expected attrition, we over-recruited participants at Session 1 and stopped when we reached our desired sample size (i.e., at least 270 participants) at Session 5, keeping all over-recruited participants in the dataset. Three-hundred-and-eighty-nine participants took part in Session 1. However, 53, 32, 10, and 10 participants did not return for Sessions 2-5, respectively (total attrition = 105). Of the remaining 284 participants, 25 had their data excluded (see *Data Exclusions* section). Sixty-seven percent of participants ($n = 259$; 154 female) remained in the sample with a mean age of 35.89 ($SD = 10.88$). Participant demographics and characteristics are shown in Table S1 in the Supplementary Materials.

### Design and Materials

An overview of the experimental design is shown in Fig. 1. Participants were randomly assigned to one of three experimental groups by the Gorilla Experiment Builder. For all three groups, one-third of all trials were restudy trials, and two-thirds were retrieval trials. The blocked 50% group participants saw 33% restudy trials, 33% overt retrieval trials, and 33% covert retrieval trials. Note that this list structure meant that participants made overt retrieval responses on 50% of the retrieval trials, hence the group name. Participants in the random 25% group saw 33% restudy trials, 17% overt retrieval trials, and 50% covert retrieval trials,

Footnote 2 (continued)
the LPSs, they reported their learning experience using a Visual Analogue Scale (VAS) to measure the learning experience. The VASs administered in the LPSs and final test were adapted from Higham et al. (2022) who measured anxiety, attentional control, and mastery. Here, we also included a measure of intrinsic motivation. Each construct consisted of three items which were each measured on a scale of 0-100.

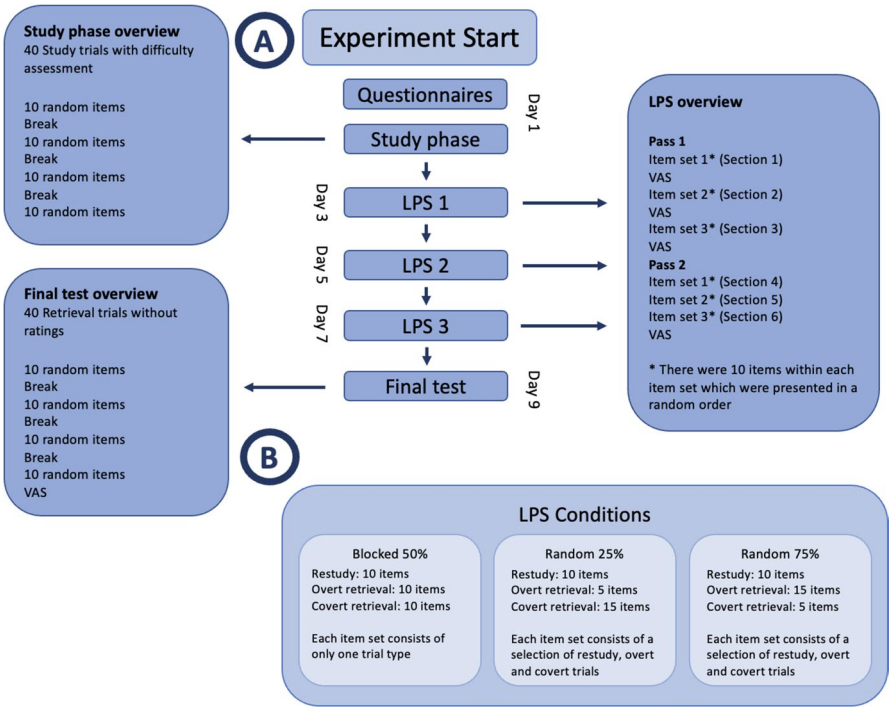



**Fig. 1** Overview of the design in Experiment 1. (**A**) An overview of the study phase, LPSs, and final test. (**B**) A description of the different LPS conditions. LPS = learning practice session; VAS = visual analog scale

whereas those in the random 75% group saw 33% restudy, 50% overt retrieval, and 17% covert retrieval trials. Thus, 25% versus 75% of the retrieval trials required an overt response in the random 25% versus random 75% groups, respectively.

Stimuli for the experimental study were 40 Swahili–English word pairs (e.g., *bustani–garden*) taken from Nelson and Dunlosky (1994). The 40 pairs are listed in the Supplementary Materials. The English translations were all between three and six letters in length. All 40 pairs were shown intact (Swahili word on the left; English word on the right) in the initial study phase. Ten items were randomly assigned to each of the four within-subjects conditions (covert, overt, restudy, and study once). To ensure that each condition was represented by a variety of items, this process of random assignment was done three times per group, with approximately one-third of participants encountering each assignment format. The ten items assigned to the study-once condition were shown in the initial study phase and were not encountered again until the final recall test. The order of presentation of the pairs in the initial study session was freshly randomized for each participant.

Thirty pairs were practiced during the LPSs (40 original pairs minus 10 that were only studied once). Participants practiced the full set of 30 pairs in one

pass and then practiced the full set again in a second pass (60 trials in total). Each pass through the 30 pairs was divided into three sections of 10 pairs. Following each section in the first pass, participants completed a visual analogue scale (VAS). For the second pass, the VASs were completed again after practicing all 30 items (four VAS completions per LPS).

In the blocked 50% condition, the items in each section of each LPS were blocked such that participants restudied, covertly retrieved, or overtly retrieved groups of 10 items in each LPS section for both passes. The order of the three tasks was counterbalanced across participants but was the same on each pass. Each counterbalance version had a different random assignment of pairs to the covert, overt, and restudy conditions.

Participants in the random 25% and random 75% groups practiced all trial types in an intermixed order. Again, there were two passes through the set of 30 pairs (60 trials in total). In the first, second, and third section of both passes in each LPS in these two groups, there were 3, 3, and 4 restudy trials, respectively. However, the proportion of overt versus covert trials in each section depended on the group. In the random 25% group, the first, second, and third section of each pass contained 2, 2, and 1 overt trials, and 5, 5, and 5 covert trials, respectively. This structure was reversed for the random 75% group (i.e., 2, 2, and 1 covert trials, and 5, 5, and 5 overt trials, for the first, second, and third section of each pass, respectively). Even though there were no blocked sets of trials to counterbalance for order in either of the random groups, there were still three different random item assignments to conditions across participants within each group, ensuring that each condition was represented by a variety of items. As in the blocked 50% group, participants completed the VASs every ten trials during the first pass of 30 trials, but only once at the end of the second pass, for a total of four VAS completions per LPS. The assignment of items to the overt, covert, and restudy conditions remained the same across the three LPSs for any given participant in all conditions.

In all groups, the order of presentation of the pairs assigned to each section of each LPS was freshly randomized for each participant. Per participant, the order was freshly randomized between passes of the same LPS with the constraint that the same pairs were assigned to both passes of each section. In other words, the items assigned to both passes through Item Set 1 (i.e., trials 1-10 and 31-40; see Fig. 1) were the same, but presented in a different order, and the same was true of both passes through Item Set 2 (trials 11-20 and 41-50) and Item Set 3 (trials 21-30 and 51-60). This process of fresh randomization per section for each participant while maintaining the same item assignment in each section was repeated for the second and third LPS.

Two days following the third LPS, participants returned for final cued-recall test. All 40 cues were presented one at a time in a unique random order for each participant, with 5-15 items each for the overt, convert, restudy, and study once conditions (exact number depended on group; see Fig. 1), and a prompt to recall the English translation. The final VAS was completed after the final test.

## Procedure

The procedure consisted of five sessions: a study phase, three LPSs, and final test (Fig. 1).

***Session 1: Study Phase.*** Session 1 was available on Prolific starting at approximately 9:00-10:00 BST for a period of approximately 24 hours. Participants provided their date of birth (month and year only), sex, gender identity, and highest level of completed education. They then completed the Student Self-Efficacy Scale (Rowbotham & Schmitz, 2013), Attentional Control Scale (Derryberry & Reed, 2002), Intolerance of Uncertainty (Carleton et al., 2007), and the GAD-7 (Spitzer et al., 2006) questionnaires. These data are reported in a separate paper along with the VAS results from the later sessions and will not be discussed further here.

Participants then studied all 40 to-be-learned word-pairs in a unique random order for each participant on a trial-by-trial basis, with a self-paced break after every 10 trials. Each trial started with a 250 ms centrally presented fixation cross preceded and followed by a 100 ms blank screen. Then, a singular word-pair in the format "*bustani* – *garden*" was presented for 5 s during which participants were instructed to study the word pair. After a 150 ms blank screen, participants had 10 s to answer the question "How difficult is this pair to learn?" with a slider scale from 0-100 (the default starting position was 50) and the label "Very easy, I'll remember it" at 0 and "Very hard, I won't remember it" at 100. This rating was included solely to encourage participants to pay attention whilst studying the items, and the data were not analyzed.[3] The trial was timed out if no response was provided after 10 s. A 500ms blank screen preceded the next trial. At the end of the study phase, participants completed a short exit survey about data quality and any technical issues.

***Sessions 2-4: LPSs.*** Forty-eight hours after Session 1 was initially available on Prolific, all eligible participants had approximately 24 hours to return for Session 2. Participants completed the first pass through the three sections of 10 trials (in the blocked 50% group each block consisted of only one trial type), completing VAS ratings of anxiety, attentional control, mastery, and intrinsic motivation. This design meant that there were three evenly spaced VAS ratings during the first pass through the 30 pairs. One final VAS completion occurred at the end of the second pass through the 30 pairs (Fig. 1).

The procedure on each trial during the LPSs is shown in Fig. 2. Each trial began with a centrally presented fixation cross for 250 ms (preceded and followed by a 100 ms blank screen) and then a 1 s instruction to either "study" (on restudy trials) or "recall" (on overt and covert retrieval trials). On restudy trials, the intact word pair

[3] These data (for both this experiment and the later experiments) are available on OSF in case readers wish to analyze them.

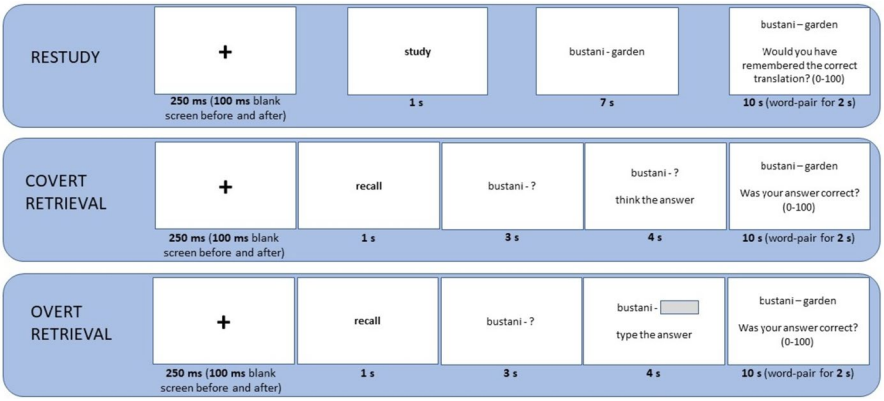



**Fig. 2** Schematic diagram of the learning practice activities on each trial in the restudy, covert retrieval, and overt retrieval conditions of Experiment 1

(e.g., "*bustani – garden"*) was presented for 7 s followed by a retrospective question: "Would you have remembered the correct translation?" and a 0-100 slider scale with the label "Definitely not" at the far-left and "Definitely yes" at the far-right. For the first 2 s of this screen, the word pair continued to be displayed so that it was visible for a total of 9 s. However, the question and slider remained on screen for a further 8 s, so participants had a total of 10 s (if they needed it) to respond and move on to the next word pair. The trial was timed out if no response was provided after 10 s. The trials were separated with a 500 ms blank screen.

Following the fixation cross and instruction to "recall" on overt and covert retrieval trials, only the Swahili word was presented with a question mark next to it (e.g., *"bustani - ?"*) for 3 s. After 3 s, on overt retrieval trials, an instruction appeared below the Swahili word to "type the answer" and the question mark next to the Swahili word was replaced by a text box in which participants could type a response (or indicate that they did not know the answer). The instruction and text box remained on the screen for 4 s. After 3 s of viewing the Swahili word and question mark on covert retrieval trials, the instruction to "think the answer" appeared below the Swahili word, which remained on the screen for 4 s. Finally, participants saw correct-answer feedback in the form of the intact word-pair (e.g., *"bustani – garden"*) for 2 s with the question "Was your answer correct?" below the pair. Participants provided answers on a 0-100 slider scale with the label "Definitely wrong" at the far-left and "Definitely correct" at the far-right. The trial ended after participants responded and advanced to the next trial. The question and slider were available for an additional 8 s after the correct-answer feedback disappeared. If participants took longer than that, the trial was timed out. The trials were separated with a 500-ms blank screen.

Participants then completed the second pass through the 30 pairs which were again divided into three sections of 10 pairs. Participants were not required to complete VASs between sections in the second pass. Finally, participants completed a fourth VAS at the end of the second pass followed by an exit survey about data

quality and any technical issues (Fig. 1). The LPSs in Sessions 3 and 4 were identical to Session 2, with a 2-day ISI, except for the presentation order of the items within the LPS sections.

***Session 5: Final Test.*** Two days after Session 4 was initially available on Prolific, participants were invited back with a 24-hour time window to complete the final test in Session 5 (Day 9 of the study). All 40 word-pairs were tested. Test trials started with a 250 ms fixation cross preceded and followed by a 100 ms blank screen after which a Swahili word with a question mark next to it appeared (e.g., "*bustani - ?*"). Participants typed in the English translation in a text box (or indicated that they did not know by typing *don't know* or *dk*) and pressed the enter key to continue to the next trial. Each trial was self-paced with unlimited time to respond and was followed by a 150 ms blank screen, 2 s of correct-answer feedback in the form of the intact word pair, and a 500 ms blank screen. Self-paced breaks were permitted every 10 trials and the order of the trials was freshly randomized for each participant. At the end of the final test, participants completed a final (thirteenth) VAS and a final exit survey on data quality and technical issues.

## Analysis

### Data Exclusions

Four participants were excluded because they met one or more of our preregistered exclusion criteria. Participants' data were excluded if (1) they timed out on more than 10 out of 40 encoding trials (Session 1) by not interacting with the slider (one participant excluded); (2) they timed out on more than eight out of 30 LPS trials (Session 2, 3, or 4) by not interacting with the slider (two participants excluded)[4]; or if they did not submit a response on more than 10 out of 40 trials at final test (one participant excluded).

We also excluded participants (as per the preregistration) if they answered affirmatively to any of the following on the exit survey: writing word-pairs down during Sessions 1–4 (no participants), giving poor quality data (one participant), or significant technical issues (one participant).

Following our preregistration, an intra-class correlation coefficient was calculated using a two-way mixed effects model with participants and accuracy (researcher-rated vs. participant-rated) as crossed-random effects (Shrout & Fleiss, 1979). This correlation coefficient assessed the agreement between researcher rated accuracy and participants' rating of their own accuracy on overt trials during the LPSs (which were on a scale of 0–100). Seventeen additional participants' data were removed for having an intra-class correlation coefficient less than 0.75. These participants were excluded because we reasoned that if participants could not evaluate whether their answer was correct or not on a significant number of items after being provided with

[4] We originally preregistered that we would exclude participants who timed out on 10 out of 40 trials, but the total number of trials was an error (see "Deviations from Preregistered Protocol").

explicit feedback, they were likely not paying attention or were disengaged with the task. Two additional participants' data were excluded for miscellaneous technical issues in the data that were discovered during analysis.

### Hypotheses and Analysis Plans

Our preregistered hypotheses are shown in Table 1. For Hypothesis 1a, we predicted a main effect of condition in a 4 × 3 mixed Analysis of Variance (ANOVA) with a within-subjects factor of condition (overtly retrieved, covertly retrieved, restudied, and study once) and a between-subjects factor of group (blocked 50%, random 25%, and random 75%). We preregistered that this analysis would be followed up with three post hoc comparisons using two-tailed *t*-tests with sequentially rejective Bonferroni corrections (Holm, 1979) to test three comparisons making up Hypothesis 1a: overt versus covert, covert versus restudied, and restudied versus study-once.[5] For Hypothesis 1b, we collapsed the random 25% and random 75% groups together and conducted a 2 × 2 mixed ANOVA with a within-subjects factor of condition (overtly retrieved and covertly retrieved) and a between-subjects factor of group (collapsed random and blocked 50%). We predicted a significant condition by group interaction where the difference in accuracy for overtly versus covertly retrieved items on the final cued-recall test would be larger when items were retrieved in a blocked versus random order. For Hypothesis 1c, we were interested only in the two groups who saw the word-pairs randomly intermixed. Using a 2 × 2 mixed ANOVA with a within-subjects factor of condition (overtly retrieved and covertly retrieved) and a between-subjects factor of group (random 25% and random 75%), we predicted a significant condition by group interaction where the benefit of overt versus covert retrieval would decrease as the proportion of overt pairs in the list increased.

We conducted a sensitivity analysis using G*Power (Faul et al., 2009). With 259 participants, an alpha level of 0.05 and power of 0.90, we would be able to detect an effect size of Cohen's $f = 0.09$ for the main effect of condition (Hypothesis 1a). For the interaction effects for Hypotheses 1b and 1c, the same parameters allowed us to detect an effect size of Cohen's $f = 0.10$. Thus, we had enough power to detect a small effect ($f = 0.10$).

[5] Throughout the experiments we report, we used the standard Bonferroni correction (i.e., alpha divided by the number of comparisons) if the follow-up comparisons were orthogonal. However, we corrected the alpha level for follow-up *t*-tests using sequentially rejective Bonferroni corrections if the comparison were non-orthogonal, which was the intended application of the Holm (1979) procedure. With this method, if there are $k$ pairwise comparisons, researchers first test the null hypothesis for the pair with the smallest $p$ value among all the pairwise comparisons, using alpha = $0.05/k$ as the adjusted alpha level. If this comparison is statistically significant, then researchers test the null hypothesis of the pair with the second smallest $p$ value, using alpha = $0.05/(k - 1)$. This process continues until the null hypothesis is retained.

**Table 2** Recall accuracy for each pass of the three learning practice sessions (LPSs) in Experiments 1 and 2

| LPS & pass | Experiment | | | | | |
|---|---|---|---|---|---|---|
| | Experiment 1 | | Experiment 2 | | | |
| | Overt retrieval | | Rating present | | Rating absent | |
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| LPS 1 | | | | | | |
| Pass 1 | 0.08 | 0.14 | 0.09 | 0.14 | 0.09 | 0.17 |
| Pass 2 | 0.27 | 0.23 | 0.25 | 0.22 | 0.30 | 0.23 |
| LPS 2 | | | | | | |
| Pass 1 | 0.31 | 0.23 | 0.33 | 0.26 | 0.34 | 0.25 |
| Pass 2 | 0.51 | 0.28 | 0.47 | 0.29 | 0.54 | 0.28 |
| LPS 3 | | | | | | |
| Pass 1 | 0.53 | 0.27 | 0.51 | 0.29 | 0.58 | 0.27 |
| Pass 2 | 0.65 | 0.26 | 0.65 | 0.30 | 0.71 | 0.25 |

*LPS* = learning practice session

## Deviations from Preregistered Protocol

We initially stated we would exclude participants if they had 10 or more missed responses out of a total of 40 for self-rated accuracy across all trial types in each LPS. However, the total number of trials was an error, as there are only 30 trials per LPS. Consequently, we instead adopted an exclusion rule of eight or more missed responses. We also added an exclusion rule based on missed responses at final test. Although the final test was self-paced, participants could still miss trials by pressing enter without submitting a typed response. The instructions under the box clearly asked participants to type a response to indicate that they did not know the answer, so we excluded one participant who did not provide a response to 10 or more out of 40 final test trials.

In the preregistration, we indicated that we would test Hypothesis 1a with a 3 (group: blocked 50%, random 25%, and random 75%) × 4 (condition: overtly retrieved, covertly retrieved, restudied, and study once) ANOVA. However, items that were studied once showed very low final-test accuracy on the final test ($M = 0.03$, $SD = 0.10$). We were concerned that including these items in the main analysis could potentially distort the results because the variance for studied-once items was considerably lower than for the other conditions. One reason for this poor performance could have been that the retention interval was longer (seven days) than for the other items (two days). Consequently, we also examined recall performance for the overtly retrieved items on the first pass in the first LPS so that the retention interval would be the same as the other conditions (i.e., two days). However, performance was not much better (see Table 2). Consequently, the study-once data were excluded from the main analysis. Therefore, Hypothesis 1a was tested with a 3 (group: blocked 50%, random 25%, random 75%) × 3 (condition: overtly retrieved, covertly retrieved, and restudied) ANOVA.

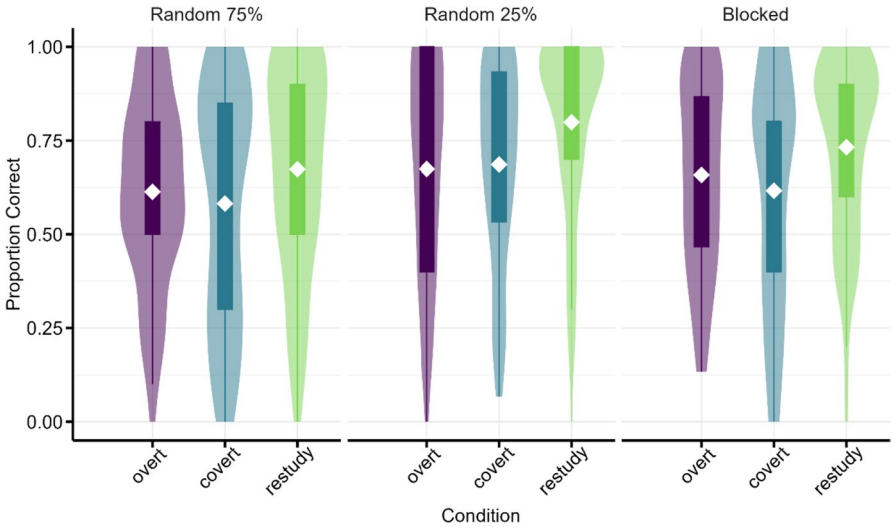



**Fig. 3** Mean proportion correct on the final test as a function of group (random 75%, random 25%, and blocked 50%) and condition (overt, covert, and restudy)

Finally, we preregistered that we would conduct a sensitivity analysis with power = 0.80. However, to be consistent with later experiments, we increased power in that analysis to 0.90.

## Results

Although we intended both the ISIs and retention intervals for all groups to be 48 hr, participants had a 24-hr time window to start the LPSs and the final test. Thus, there was a risk that the intervals were not 48 hr as intended. To ensure that there was not excessive deviation from 48 hr for either the ISIs or the retention intervals, we computed mean intervals for each group. The results showed that both mean interval types were close to 48 hr (lowest mean interval = 46 hr, 39 min; $SD$ = 6 hr, 21 min; highest mean interval = 50 hr, 49 min; $SD$ = 6 hr, 40 min).

Mean recall accuracy on the final cued-recall test is shown in Fig. 3, and mean recall accuracy in the overt retrieval condition for both passes through the items during the LPSs is shown in Table 2. Bayes factors were calculated with the BayesFactor package for R (Morey & Rouder, 2022) using the default prior. Following Olejnik and Algina (2003), we report generalized eta squared ($\eta^2_G$) as a measure of effect size for ANOVA.

### Hypothesis 1a (Overt > Covert > Restudy > Study Once for Final Cued–Recall Accuracy)

A 3 (group: blocked 50%, random 25%, and random 75%) × 3 (condition: overtly retrieved, covertly retrieved, and restudied) mixed ANOVA on accuracy at final test showed a significant main effect of group, $F(2, 256) = 3.52$, $p = 0.03$, $\eta^2_G = 0.02$,

$BF_{10} = 1.46$, a significant main effect of condition, $F(2, 512) = 34.86$, $p < 0.001$, $\eta^2_G = 0.03$, $BF_{10} = 1.02e + 12$, but no significant interaction, $F(4, 512) = 1.30$, $p = 0.27$, $\eta^2_G < 0.01$, $BF_{10} = 0.04$. As we found a significant main effect of condition, we conducted two planned paired samples *t*-tests with sequentially rejective Bonferroni corrections (Holm, 1979).

The first *t*-test showed that pairs overtly retrieved ($M = 0.65$, $SD = 0.26$) during the practice sessions were not significantly better recalled at final test compared to pairs that were covertly retrieved ($M = 0.63$, $SD = 0.30$), $t(258) = 1.42$, $p = 0.157$ $d_z = 0.09$, $BF_{10} = 0.19$, $a = 0.05$. Nor were covertly retrieved pairs better recalled than restudied pairs ($M = 0.74$, $SD = 0.26$). In fact, a *t*-test showed that there was a significant difference in the opposite direction, $t(258) = 7.86$, $p < 0.001$, $d_z = 0.49$, $BF_{10} = 57{,}545{,}696{,}947$; $a = 0.0167$. An additional exploratory sequentially rejective Bonferroni corrected *t*-test showed that restudied pairs were also better recalled than overtly retrieved pairs, $t(258) = 6.61$, $p < 0.001$, $d_z = 0.41$, $BF_{10} = 32{,}331{,}252$; $a = 0.025$. Thus, no comparison was consistent with Hypothesis 1a.

Although not preregistered, we conducted three additional exploratory *t*-tests with sequentially rejective Bonferroni corrections to break down the main effect of group. These tests showed that the blocked 50% group ($M = 0.59$, $SD = 0.24$) produced worse recall compared to the random 25% group ($M = 0.69$, $SD = 0.23$), $t(170) = 2.62$, $p < 0.010$, $d = 0.40$, $BF_{10} = 3.83$; $a = 0.017$, but not the random 75% group ($M = 0.64$, $SD = 0.22$), $t(168) = 1.40$, $p = 0.16$, $d = 0.22$, $BF_{10} = 0.41$; $a = 0.025$. Final test recall for the random groups did not differ significantly, $t(174) = 1.30$, $p = 0.19$, $d = 0.20$, $BF_{10} = 0.36$; $a = 0.05$.

#### Hypothesis 1b (Overt vs. Covert Difference in Final Cued–Recall Accuracy Greater in the Blocked Group than the Random Groups)

We collapsed the random 25% and random 75% groups and performed a 2 (group: blocked 50%, collapsed random) × 2 (condition: overtly retrieved and covertly retrieved) mixed ANOVA on accuracy on the final test. There was no significant main effect of group, $F(1, 257) = 3.23$, $p = 0.07$, $\eta^2_G = 0.01$, $BF_{10} = 0.91$, nor a significant main effect of condition, $F(1, 257) = 2.01$, $p = 0.16$, $\eta^2_G < 0.01$, $BF_{10} = 0.26$, nor a significant interaction, $F(1, 257) = 0.30$, $p = 0.59$, $\eta^2_G < 0.01$, $BF_{10} = 0.17$. Because we did not find the significant group by condition interaction that we predicted, we did not conduct any follow up *t*-tests.

#### Hypothesis 1c (Overt vs. Covert Difference in Final Cued–Recall Accuracy Smaller in the Random 75% than the Random 25% Group)

A 2 (group: random 25% and random 75%) × 2 (condition: overtly retrieved and covertly retrieved) mixed ANOVA on accuracy on the final test indicated that neither the main effect of group, $F(1, 174) = 1.24$, $p = 0.27$, $\eta^2_G < 0.01$, $BF_{10} = 0.43$, nor condition, $F(1, 174) = 0.82$, $p = 0.37$, $\eta^2_G < 0.01$, $BF_{10} = 0.17$, was significant. The interaction was also not significant, $F(1, 174) = 2.74$, $p = 0.10$, $\eta^2_G < 0.01$, $BF_{10} = 0.61$. Because we did not find the predicted group by condition interaction, we did not conduct any follow up *t*-tests.

### Additional Exploratory Analyses

The unexpected superiority of the restudy condition led us to conduct an exploratory analysis on reaction times (RTs). Our design ensured that exposure to each item was balanced up to the point that the memory rating prompt appeared. Specifically, on restudy trials, the intact pair was shown for 7 s. On retrieval trials, the cue was shown on its own for a 4 s retrieval period, and then, there was an additional 3 s to provide a response (7 s total).[6] However, after the memory rating prompt appeared, participants had up to 10 s to submit an answer which would effectively end the trial (see Fig. 2). Hence, there could be some variability in exposure to the material after the memory rating prompt appeared that favored the restudy condition leading to a memory advantage. For example, participants in the restudy condition might have found the hypothetical nature of the memory rating difficult to answer such that they processed the intact pair for longer than in either of the retrieval conditions.

To address this possibility, we compared mean RTs between the conditions in a one-way, between-subjects ANOVA. The RT was measured starting at the point that the memory rating prompt appeared to the point that participants entered a memory rating (Fig. 2). This analysis revealed a significant effect, $F(2, 516) = 34.78$, $p < 0.001$, $\eta^2_G = 0.03$, $BF_{10} = 9.60e{+}245$. Follow-up two-tailed *t*-tests with sequentially rejective Bonferroni correction indicated that participants spent *less* time completing the memory rating in the restudy condition ($M = 3067$ ms, $SD = 862$ ms) than in either the overt ($M = 3657$ ms, $SD = 1002$ ms) or covert ($M = 3314$ ms, $SD = 916$ ms) retrieval conditions, $t(258) = 18.20$, $p < 0.001$, $d_z = 1.13$, $BF_{10} = 5.45e + 44$; $a = 0.017$, and $t(258) = 10.86$, $p < 0.001$, $d_z = 0.67$, $BF_{10} = 5.60e + 19$; $a = 0.05$, respectively. Mean RTs in the overt and covert retrieval conditions also differed significantly, $t(258) = 12.35$, $p < 0.001$, $d_z = 0.77$, $BF_{10} = 4.17e + 24$; $a = 0.025$.

### Discussion

The aim of Experiment 1 was to explore whether items overtly retrieved during learning would be recalled most effectively in a memory test compared with those that were (in order of performance) retrieved covertly, restudied, or studied once. However, because study-once performance was near floor, we did not include that condition in the analysis. In addition, we anticipated that the difference between overt and covert retrieval would lessen if the learning experience required fewer overt (vs. covert) responses. In short, the tested predictions were either not supported (i.e., recall accuracy for overt retrieval was not better than covert retrieval), or there were effects in the opposite direction (i.e., recall accuracy for covertly retrieved items was *worse* than for restudied items).

[6] Note that, as with many single-session testing effect experiments, the design does not balance the amount of time that participants were exposed to the *intact pair*, which would have been greater in the restudy condition than the retrieval conditions. However, the design does balance the amount of time per trial that participants spent processing or "thinking about" each item (up to the point of the memory rating prompt).

The most striking result from Experiment 1 was the unexpected superiority of the restudy condition over all other conditions. Three practice sessions of successive restudying following an initial study session resulted in final recall accuracy equal to 0.74. In contrast, three sessions of spaced retrieval practice (either covert or overt) only resulted in final recall equal to 0.64. The exploratory RT analysis indicated that participants' superior memory performance in the restudy (vs. retrieval) condition could not be explained by participants spending longer completing the memory rating (resulting in greater exposure to the items). Rather, it occurred despite participants in the restudy condition spending the *least* amount of time completing the memory rating. Higham et al. (2022) also found that spaced restudying greatly improved recall. However, their comparison of final recall performance for items that had been successively restudied versus encountered once during a lecture was confounded by the items. That is, because it was a classroom intervention, the items representing the two conditions could not be counterbalanced, so conceivably, the items representing the restudy condition were easier than those that were not practiced. In the current study, however, we assigned items to conditions using three different randomization procedures, making it very unlikely that item differences could explain any differences between the conditions.

In most circumstances, restudying is considered a low-utility study strategy (e.g., Dunlosky et al., 2013). However, most testing effect research with a restudy condition has included a single practice session (see Rowland, 2014 for a review). The results of the current study raise the possibility that repeated restudying over spaced intervals converts a low-utility learning strategy into one that has high utility. After all, the vast majority of research on the robust *spacing effect* in memory research involves repeated *presentation* of to-be-remembered information over spaced intervals, not *retrieval* of that information (see Benjamin & Tullis, 2010 for a review).

An alternative explanation of the results in Experiment 1 is that is there was a specific feature of our methodology that may have enhanced the utility of repeated spaced restudying and/or reduced the utility of retrieval. In the current study, participants in the covert retrieval condition in Experiment 1 were not required to respond with the to-be-remembered information, whereas they were in the overt condition. To make the overt and covert conditions more comparable in their study, Putnam and Roediger (2013) asked participants on covert retrieval trials "Did you remember?" to which a "yes" or "no" response was required. That way, both groups had to make an overt response, but only the overt retrieval condition overtly responded with the retrieved information. We adjusted this procedure slightly in Experiment 1 by asking participants on both overt and covert retrieval trials "Was your answer correct?" to which they provided a 0–100 rating. To equate the restudy and retrieval trials in terms of overt responding, we also asked participants on restudy trials "Would you have remembered the correct translation?," again with a 0–100 scale. This design feature meant that all three groups were directed to consider the state of their memory (or the hypothetical state of their memory) with respect to each item at the end of each trial, which may have produced reactivity like that observed with JOLs (e.g., Soderstrom et al., 2015). However, if these ratings produced reactivity, they may have had differential effects on memory in the restudy and retrieval conditions. As noted above, we are aware of no reactivity studies that have used English

translation pairs as materials, and finding both positive and negative reactivity in the same experiment is rare in the literature (although see Mitchum et al., 2016). We therefore considered it worthwhile to directly manipulate the requirement to make memory ratings in Experiment 2 and observe the effect on restudy and retrieval memory performance.

## Experiment 2

The focus of Experiment 2 was to understand the superior recall performance obtained in the restudy condition of Experiment 1. Specifically, we investigated the impact on recall of making a memory rating at the end of each trial. Consistent with Experiment 1, participants completed five sessions aimed at learning and testing memory for Swahili–English word translations (i.e., initial study, three LPSs, and a final cued recall test, with two-day intervals between each session). As there were no differences in recall between the covert and overt retrieval conditions in Experiment 1, Experiment 2 only included the overt retrieval, restudy, and study-once conditions. No VASs were used in Experiment 2 as we were focused on the cause of the superior final test recall performance in the restudy condition obtained in Experiment 1.

To manipulate the requirement to make a memory rating during the LPSs, participants made the same memory ratings that they made in Experiment 1 on half the overt and restudy trials, whereas for the other half, there was no rating requirement. Specifically, on rated restudy trials, participants were asked "Would you have remembered the English translation?" whereas on rated overt retrieval trials, they were asked "Was your answer correct?". Both responses were made on a 0–100 slider scale as in Experiment 1. We manipulated the requirement to make memory ratings within-subjects to increase power and to reduce the number of participants that would be needed. One potential concern with this design is carry-over effects. That is, if participants make memory ratings for some items and not for others, then there is a concern that they will make covert ratings on no-rating trials, thereby weakening the manipulation. We were not overly concerned about this possibility given that Rivers et al. (2021) found that there was little evidence for carry-over effects in their experiments in which the requirement to make JOLs was manipulated within subjects. Nonetheless, to minimize the possibility of participants covertly making memory ratings in the no-rating condition, the rating and no-ratings conditions were blocked rather than randomly intermixed. We reasoned that by blocking the rating requirement, but randomly intermixing the practice type (restudy vs. retrieval), any contaminating effects of making the ratings would likely only occur on the first few trials when the rating requirement changed (from rating to no rating) if they occurred at all.

This design allowed us to directly test whether restudying was effective compared to retrieval in Experiment 1 because of the requirement to make a memory rating. Our preregistered hypothesis was that the memory rating would moderate the effect of practice type (Table 1). That is, when a memory rating was required, restudied items would be recalled better than overtly retrieved ones, replicating the results

of Experiment 1. On the other hand, when a memory rating was not required, then overtly retrieved items would be recalled better than the restudied ones, replicating the typical testing effect (e.g., Rowland, 2014).

## Method

### Participants

Participants were recruited and enrolled the same way as in Experiment 1 with the same inclusion and exclusion criteria. Ninety-three participants took part in Session 1. However, 13, 10, 4, and 0 participants did not return for Sessions 2–5, respectively. Of the remaining 66 participants, three had their data excluded (see "Data Exclusions"). Sixty-three participants (42 female) remained in the sample with a mean age of 32.30 ($SD$ = 9.89). Participant characteristics and demographics are shown in Table S2.

### Design and Materials

Experiment 2 had a 2 (condition: retrieve and restudy) × 2 (memory rating: present and absent) within-subjects design. The stimuli for the study were the same 40 Swahili–English word pairs (e.g., *bustani–garden)* used in Experiment 1. Eight pairs were initially randomly assigned to each of five within-subject conditions: study-once, restudy-rating, retrieve-rating, restudy-no-rating, and retrieve-no-rating. After the pairs were randomly assigned to these conditions, the 32 word pairs that were seen during the LPSs were rotated through the restudy-rating, retrieve-rating, restudy-no-rating, retrieve-no-rating conditions across participants using a Latin Square design with four counterbalance formats. The eight pairs randomly assigned to the study-once condition that were not seen during the LPSs (and hence not part of the main 2 × 2 design) were not included in the counterbalancing; that is, word pair assignment for this condition was held constant across all participants. The study-once items were not included in the main counterbalanced design because Experiment 1 showed that performance on them was at floor. We included the study-once condition for completeness, but our main focus was on the restudy and retrieval conditions.

### Procedure

***Session 1: Study Phase.*** The study phase procedure was identical to that of Experiment 1, except that individual differences were not measured. It took place two days prior to the first LPS, as in Experiment 1.

***Sessions 2–4: LPSs.*** The timing of the LPSs was identical to Experiment 1 in that they were approximately two days apart. Participants completed the first practice cycle of eight restudy trials with a memory rating, eight restudy trials with no memory rating, eight retrieval trials with a memory rating, and eight retrieval trials with no

memory rating. Restudy and retrieval practice trials were intermixed, but rating versus no rating at the end of each trial was blocked; half of participants completed the memory rating condition first and half completed the no memory rating condition first.

The trial-level procedures for the retrieval-rating and restudy-rating conditions were identical to the overt retrieval and restudy conditions from Experiment 1, respectively (see Fig. 2). On restudy and retrieval trials that did not require a rating, the timing of the events was the same except that the trial ended at the point that the rating question would have appeared in the rating conditions. The no-rating trials were therefore shorter than the rating trials in that the final 8 s that was available to participants in the rating conditions to enter their rating was not available in the no-rating conditions. Note, however, that most participants in the rating conditions did not use the full 8 s to make a response, so the timing differential was considerably less than that (see exploratory analyses on RTs later).

After the first cycle of all 32 experimental trials, participants completed a second pass through the items. The order of the rating/no-rating blocks was the same between the two passes, but trial order was randomly shuffled within blocks for each pass. After completing the practice session, participants completed an exit survey about data quality and any technical issues.

***Session 5: Final Test.*** The final test was identical to that in Experiment 1, except participants did not complete the VAS at the end.

## Analysis

### Data Exclusions

No participants were excluded for timeouts on encoding, practice, or final test trials. One participant was excluded for answering affirmatively to giving poor quality data, and one participant was excluded for technical issues. One additional participant was excluded for a reason that was not preregistered: they explained in the exit questionnaire that they covered the English word during restudy trials, which effectively turned them into retrieval trials.

In Experiment 1, we excluded 17 participants who had an intra-class correlation (between the researchers' vs. the participants' assessment of recall accuracy during the LPSs) that was less than 0.75. However, this criterion resulted in the loss of data from 17 participants. Given that correspondence between the two types of recall accuracy assessment was not critical or central to any of our aims in Experiment 2, we determined that this criterion was overly harsh, so it was not applied in Experiment 2.

### Hypotheses and Analysis Plans

Our preregistered hypothesis for Experiment 2 is shown in Table 1. We tested it with 2 (condition: retrieval and restudy) × 2 (memory rating: present and absent)

within-subjects ANOVA on accuracy on the final cued–recall test, with the expectation of obtaining a significant interaction (followed up with two-tailed, Bonferroni-corrected *t*-tests). Specifically, we predicted that on trials ending with a memory rating, cued–recall performance on the final test would be better in the restudy condition than the retrieval condition, replicating the findings from Experiment 1. However, for trials where no memory rating was required, final test performance would be better in the retrieval condition than the restudy condition, replicating the standard testing effect and reversing the pattern observed when memory ratings were required.

We preregistered that we would base our sample size on the results of Experiment 1 in which we observed an effect size of $d_z = 0.42$ for the performance benefit of restudying versus overt retrieval practice when ratings were required. A power analysis for a one-tailed *t*-test using $d = 0.42$, alpha level = 0.025 (to account for the family error of performing two *t*-tests), and with a power of 0.90, yielded a sample size estimate of $n = 62$.

To compensate for expected attrition, we over-recruited participants at Session 1 and stopped when we reached our sample size (at least 62 participants) at Session 5, keeping all over-recruited participants in the dataset. As the estimate for the effect size for a retrieval practice effect over restudy is Hedge's $g = 0.51$ according to a recent meta-analysis (Adesope et al., 2017), we should also be adequately powered to find this effect in the no-rating condition.

### Deviations from Preregistered Protocol

For the power analysis in the preregistration, we used the effect size $d_z = 0.42$ from Experiment 1, but we made a rounding error and should have used $d_z = 0.41$. Also, because we corrected for multiple comparisons with Bonferroni corrections in the following analyses, there was no need to reduce the alpha level a priori to 0.025. Finally, we conducted two-tailed *t*-tests instead of one-tailed tests as there is some debate whether it is appropriate to use one-tailed *t*-tests solely because the effect being tested was predicted (e.g., Ruxton & Neuhäuser, 2010). We repeated the power analysis with these corrections (i.e., using $d_z = 0.41$, no reduction of alpha level, and two-tailed tests), and it revealed that the desired sample size should have been 65. Consequently, our sample size was slightly under this estimate after the exclusions, but the difference was small (i.e., estimated sample: $n = 65$; actual sample: $n = 63$).

### Results

As in Experiment 1, we computed mean ISIs and retention intervals for each condition to ensure that there was not excessive deviation from 48 hr for either interval type. As before, the results showed that both mean interval types were close to 48 hr (lowest mean interval = 47 hr, 37 min; *SD* = 7 hr, 22 min; highest mean interval = 50 hr, 31 min; *SD* = 6 hr, 11 min).

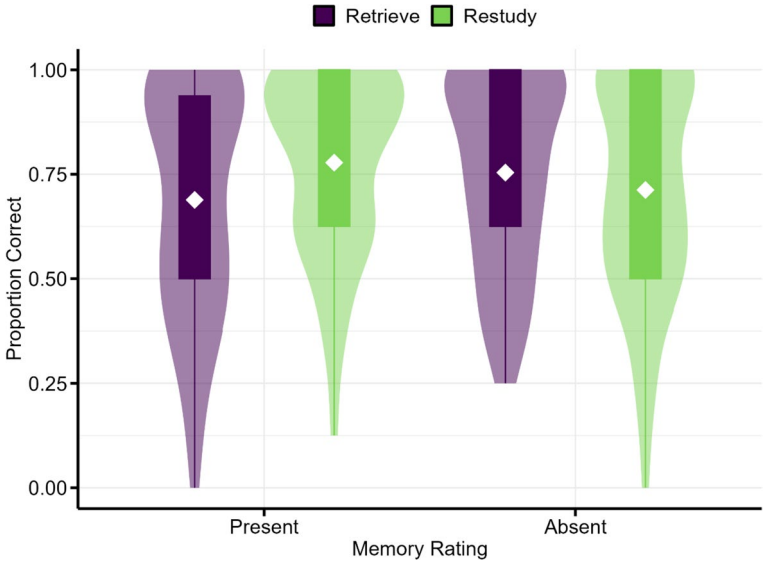



**Fig. 4** Cued–recall accuracy on the final test as a function of condition (retrieve and restudy) and memory rating (present and absent)

Recall accuracy during the LPSs is shown in Table 2. As in Experiment 1, recall performance on the final test for pairs that were studied once was near floor ($M = 0.05$; $SD = 0.12$) and was only slightly better for retrieved items in the first pass through the items in LPS 1, where the two-day retention interval was matched to that in the restudy and retrieval conditions (see Table 2). Consequently, the study-once data were again excluded from the main analysis. The main results for Experiment 2 are shown in Fig. 4. A 2 (condition: retrieve, restudy) × 2 (memory rating: present, absent) within-subjects ANOVA on final recall accuracy showed no significant main effect of condition, $F(1, 62) = 1.15$, $p = 0.29$, $\eta^2_G < 0.01$, $BF_{10} = 0.26$, nor memory rating, $F(1, 62) < 0.01$, $p > 0.99$, $\eta^2_G < 0.01$, $BF_{10} = 0.14$. However, there was a significant interaction, $F(1, 62) = 12.58$, $p < 0.001$, $\eta^2_G = 0.02$, $BF_{10} = 20.53$. The interaction was followed up with two Bonferroni corrected paired sample *t*-tests. The first indicated that final test recall for restudied items was better than for retrieved items when a memory rating was required during the LPSs, $t(62) = 2.89$, $p = 0.003$, $d_z = 0.36$, $BF_{10} = 5.95$, $a = 0.025$, replicating our findings from Experiment 1. A second paired sample *t*-test comparing retrieved versus restudied pairs when no ratings were required showed descriptive (but not statistically significant) superiority for the retrieved pairs, $t(62) = 1.56$, $p = 0.062$, $d_z = 0.20$, $BF_{10} = 0.44$, $a = 0.025$.

### Additional Exploratory Analyses

To test for reactivity effects, we compared final test performance between the rating and no-rating conditions separately for retrieval and restudy conditions with two

exploratory paired-sample *t*-tests. The analyses revealed that there was statistically significant positive reactivity in the restudy condition, $t(62) = 2.64$, $p = 0.010$, $d_z = 0.33$, $BF_{10} = 3.31$, but statistically significant negative reactivity in the retrieval condition, $t(62) = 2.20$, $p = 0.032$, $d_z = 0.28$, $BF_{10} = 1.29$.

As in Experiment 1, we compared RTs to complete the memory rating between the retrieve and restudy conditions. Consistent with Experiment 1, the RT for the restudy condition ($M = 3174$ ms, $SD = 667$ ms) was less than the RT in the retrieve condition ($M = 3619$ ms, $SD = 655$ ms), $t(62) = 9.02$, $p < 0.001$, $d_z = 1.14$, $BF_{10} = 11{,}669{,}243{,}669$. Therefore, the superior recall in the restudy condition compared to the retrieve condition when ratings were required cannot be attributed to the restudy items being processed for longer after the memory rating prompt.

## Discussion

The results of Experiment 2 mostly confirmed our hypotheses. If participants were required to provide a memory rating during the LPSs, restudied items were recalled better than retrieved ones on the final test, replicating the results of Experiment 1. Conversely, if no memory rating was required, retrieved items showed superior performance to restudied ones, consistent with prior research on retrieval practice. However, the latter difference was only descriptive and not statistically significant. Figure 4 shows that, of the four experimental conditions, the best recall performance was in the restudy condition that required memory ratings. At the same time, retrieval trials that required a memory rating produced the *worst* final recall performance out of the four conditions.

These results suggest that the memory ratings produced reactivity like that observed in studies on JOL reactivity (e.g., Soderstrom et al., 2015). However, there are some key differences between the reactivity observed in our research compared to previous work with JOLs. First, novel materials were used (English translation pairs). When first learning these pairs, the members of the pair were not related in any manner. Indeed, the Swahili words would initially be equivalent to meaningless nonwords at first, only taking on meaning with repeated practice. Although we are not the first to observe reactivity effects with unrelated pairs (e.g., Mitchum et al., 2016; Rivers et al., 2021), it is unusual and the effects are typically small, unlike the large reactivity we observed in the restudy condition. Second, our studies yielded both positive and negative reactivity in the same experiment. This dissociative pattern constrains some of the theoretical accounts of reactivity, a point we return to in the General Discussion.

It is noteworthy that even when participants were not required to make memory ratings, we did not obtain a robust retrieval practice effect. This result is surprising given the long history of testing effects that have been obtained in over 100 years of

memory research (see Rowland, 2014; Yang et al., 2021 for reviews).[7] One possible reason for this finding is that restudying shifts from being a low-utility strategy in single-session experiments to being a moderate- or high-utility strategy when it is repeated over spaced intervals. Because we only tested recall of restudied items in Experiments 1 and 2 after they were practiced three times, it was not possible to compare recall of restudied and retrieved items with fewer practice sessions. Experiment 3 was conducted to address this issue.

## Experiment 3

In Experiment 3, we compared the retrieval and restudy conditions over one, two, and three practice sessions. This design allowed us to determine the benefit that each successive LPS had on recall performance when participants were asked to either retrieve or restudy items. As in the previous experiments, participants learned Swahili–English word pairs over five sessions (initial study, three LPSs, and final test) with each session separated by approximately two days. So that we could assess the relative effects of restudying and retrieving without the potential influence of memory ratings, there was no requirement to make memory ratings in any condition. By eliminating this rating requirement altogether, Experiment 3 provided a purer test of the learning effectiveness of spaced restudying versus spaced retrieval practice.

To observe learning over sessions in Experiment 3, the study-once items were dropped, and all 40 pairs were either restudied or retrieved during the three LPSs. However, after studying all pairs once during the initial study session, the items selected to be restudied versus retrieved changed over the three LPSs. Specifically, in the first LPS, most items were restudied and a few were retrieved. In the second LPS, some items that were restudied in the first LPS were now retrieved. In the third LPS, more items that were restudied in both the first and second LPS were now retrieved. By examining recall performance on both the final test and the first pass through the items in the retrieval condition of the LPSs, it was possible to determine spaced repetition effects after either restudying or retrieving the items once, twice, or thrice. If spaced restudying is a low-utility strategy initially but becomes a moderate-to-high utility strategy with spaced repetitions, then recall for restudied items should start out poorer than for retrieved items on the first practice session but catch up with (or exceed) recall performance for retrieved items on later sessions. Thus, we hypothesized that we would observe a testing effect in the first LPS of Experiment 3, but this effect would be eliminated by the final test, replicating the results of Experiment 2 when no memory rating was required (Table 1).

[7] Although it was surprising that we did not observe a significant testing effect when no ratings were required, this result should not be given too much weight because the Bayes Factor for the marginal testing effect in the no-rating conditions of Experiment 2 was inconclusive ($BF_{10} = 0.44$). The important result from Experiment 2 was the cross-over interaction between memory rating and condition, which yielded a Bayes factor that fell in the range of "strong evidence for alternative hypothesis" ($BF_{10} = 20.53$).

## Method

### Participants

Twenty undergraduates from the University of Southampton enrolled in Session 1 in exchange for course credits, and 114 participants on Prolific enrolled in Session 1 in exchange for payment at a rate of £5 per hour (total = 134 participants). Twenty-two participants did not return for Session 2, 15 for Session 3, three for Session 4, and five for Session 5. Of the remaining 89 participants, 12 had their data excluded (see "Data Exclusions"). Seventy-seven participants (57 female) remained in the sample with a mean age of 33.24 (*SD* = 11.44). Participant characteristics and demographics are shown in Table S3.

### Design and Materials

Experiment 3 had a 2 (condition: restudy and retrieval) × 3 (practice frequency: 1, 2, and 3) within-subjects design. As in the previous two experiments, participants engaged in five learning/testing sessions (initial study, three LPSs, and a final recall test) involving 40 Swahili–English pairs with each session separated by approximately two days. After studying all 40 pairs during the initial learning session, participants restudied or retrieved the 40 items during each LPS. The restudy and retrieval trials during these sessions were identical to the corresponding conditions in Experiment 2 where no memory rating was required. The first LPS consisted of 30 restudy trials and 10 retrieval trials. In the second LPSs, 10 previously restudied items were shifted to the retrieval condition such that there were now 20 restudied and 20 retrieval items. In the third LPS, 10 additional items that were restudied in the previous two LPSs were retrieved such that there were now 10 restudy and 30 retrieval items. Finally, a retrieval attempt was made on all items on the final cued-recall test. A Latin square design was used to counterbalance the items across participants. With this design, it was possible to compare recall performance after items were restudied or retrieved once, twice, or thrice.[8]

### Procedure

The study phase in Session 1 was identical to that of Experiment 2. After approximately two days, participants returned for three LPSs (Sessions 2–4) with a two-day ISI. In each LPS, participants completed the first pass through all 40 word pairs, followed by a second cycle. The procedure on individual restudy and retrieval trials was identical to that in the no-rating conditions in Experiment 2, but the ratio of

[8] Note that this design meant that in the third LPS, there were also 10 items that were studied twice and retrieved once, and on the final test, there were 10 items that were studied twice and retrieved twice, and 10 items that were studied three times and retrieved once. However, as the purpose of Experiment 3 was to compare recall performance for items with a pure history of restudying versus testing, we do not report the data from these other conditions.

restudy and retrieval trials in each LPS changed across LPSs as explained earlier. Once a word pair had been assigned to the retrieval practice condition, whether that be in the first LPS or a later one, it stayed assigned to that condition for the remainder of the experiment. Participants then completed a final cued recall test in Session 5 which was identical to Experiment 2.

## Analysis

### Data Exclusions

As per our preregistration, 12 participants met the exclusion criteria. One participant stated that their data was of low quality, one missed more than 10 slider responses in the encoding session, and ten participants missed responses during the practice sessions (more than 5 during LPS 1, more than 10 during LPS 2, or more than 15 during LPS 3). This left 77 participants available for the analysis.

### Hypotheses and Analysis Plans

We preregistered the prediction that there would be an advantage for retrieved items over restudied items in the first LPS, but that this effect would be eliminated by the final recall test after three practice sessions. This hypothesis was tested with a preregistered 3 (practice frequency: 1, 2, and 3) × 2 (condition: restudy, retrieval) within-subjects ANOVA. We anticipated that we would obtain an interaction, and if so, it would be followed up with paired sample *t*-tests.

We also preregistered the prediction that both restudy and retrieval performance would increase across the LPSs. To test this hypothesis, we preregistered that we would conduct two Bonferroni corrected, paired sample *t*-tests. The first compared recall in the tested-once condition with the tested-thrice condition and the second compared recall in the restudied-once condition with the restudied-thrice condition. These comparisons allowed us to determine the effect of retrieving versus restudying once versus thrice.

We preregistered a power calculation to determine the sample size for this experiment using the same effect (i.e., restudy over retrieval advantage on the final test when memory ratings were required) and effect size as in Experiment 2. For a two-tailed *t*-test with an effect size of $d_z = 0.42$, power = 0.90, and an adjusted alpha level = 0.01 (due to a greater number of post hoc comparisons), the resulting estimate was $n = 88$ participants.

### Deviations from Preregistered Protocol

There were errors in the preregistered power analysis. As in Experiment 2, we used the effect size $d_z = 0.42$ from Experiment 1, but we made a rounding error and should have used $d_z = 0.41$. Also, because we corrected the alpha level for multiple comparisons in the following analyses, there was no need to reduce the alpha level a priori. Moreover, because there were no ratings in Experiment 3, we should not have

based our power analysis on the effect size associated with the restudy advantage over retrieval on the final test when memory ratings were required. Finally, contrary to the preregistration, all conducted *t*-tests were two-tailed. To explore the effect size we were able to detect with our final sample size, we conducted a post hoc sensitivity analysis for a two-tailed *t*-test with $n = 77$, power = 0.90, and alpha = 0.05. It revealed that we had enough power to detect an effect of $d_z = 0.37$.

There was a technical error with one word pair (*mbwa–dog*) for 18 participants on the first pass of LPS 3 which meant the trial was skipped. Consequently, there was one less presentation of this word pair for these participants. This error was unlikely to affect the results, so the item was not excluded from the dataset.

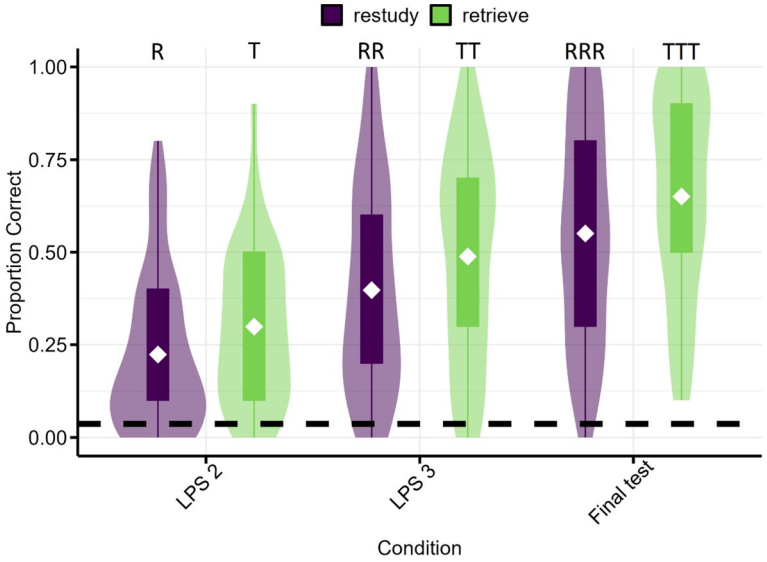
## Results

As in the previous two experiments, we computed mean ISIs and retention intervals for each condition to ensure that there was not excessive deviation from 48 hr for either interval type. As previously, the results showed that both mean interval types were close to 48 hr (lowest mean interval = 46 hr, 27 min; *SD* = 7 hr, 4 min; highest mean interval = 49 hr, 9 min; *SD* = 6 hr, 47 min).

Recall accuracy on the first pass in LPSs 2 and 3 and on the final recall test are shown in Fig. 5. Although it was not possible to compare recall accuracy for items that had previously been retrieved versus restudied items for the first LPS (because all items had only been studied once during the initial study session), we included recall accuracy for these items as a dotted line in Fig. 5 for completeness. The remaining recall accuracy data were subjected to a 3 (practice frequency: 1, 2, 3) X 2 (condition: restudy, retrieval) within-subjects ANOVA. It revealed a significant main effect of condition, $F(1, 76) = 19.03$, $p < 0.001$, $\eta^2_G = 0.03$, $BF_{10} = 853.10$, a significant main effect of practice frequency, $F(1.79,135.8) = 174.35$, $p < 0.001$, $\eta^2_G = 0.23$, $BF_{10} = 3.95e48$, but no interaction, $F(2,152) = 0.47$, $p = 0.63$, $\eta^2_G < 0.01$, $BF_{10} = 0.06$. Items that were retrieved ($M = 0.48$, $SD = 0.29$) were recalled better than items that were restudied ($M = 0.39$, $SD = 0.29$). Because there was no significant interaction, we did not conduct any follow up *t*-tests.

Additionally, two preregistered paired sample *t*-tests with Bonferroni correction showed that items that were practiced with retrieval and pairs that were restudied were both recalled on the final test (retrieval: $M = 0.65$, $SD = 0.27$; restudy: $M = 0.55$, $SD = 0.29$) better than in LPS 2 (retrieval: $M = 0.30$, $SD = 0.21$; restudy: $M = 0.22$, $SD = 0.21$), $t(76) = 15.16$, $p < 0.001$, $d_z = 1.73$, $BF_{10} = 3.63e+21$; $a = 0.025$, and $t(76) = 13.14$, $p < 0.001$, $d_z = 1.50$, $BF_{10} = 1.55e+18$; $a = 0.025$, respectively.

For completeness, we conducted an additional exploratory *t*-test to test whether we replicated the finding in Experiment 2 that there were no significant differences in recall accuracy between the restudy and retrieval practice conditions on the final test (i.e., TTT vs. RRR; Fig. 5). However, unlike Experiment 2, recall for items that had a history of retrieval (TTT) were better recalled at final test ($M = 0.65$, $SD = 0.14$), than items with a history of restudying (RRR; $M = 0.55$, $SD = 0.18$), $t(76) = 3.82$, $p < 0.001$, $d_z = 0.44$, $BF_{10} = 81.77$.




*Note*: T = tested; R = restudied; LPS = learning practice session. The initials indicate the items' history across LPSs. For example, "R" denotes recall accuracy for items restudied once in LPS 1, whereas "TT" denotes recall accuracy for items that were tested twice, once in LPS 1 and once in LPS 2. All retention intervals are approximately two days.

**Fig. 5** Mean recall accuracy on the first pass in the second and third LPS, and on the final test, for items with a history of either restudying or retrieving. The dotted line represents mean recall accuracy on the first LPS for items studied once

## Discussion

Our hypothesis received partial support in Experiment 3. As predicted, both spaced restudy and retrieval performance on the final cued–recall test improved with greater practice frequency (Fig. 5). However, we did not observe that performance would be better for retrieved items than restudied items in the second LPS, but that this difference would reduce over subsequent sessions such that by the final test, performance for the two item types would be equal. Instead, the advantage of retrieval over restudying was observed in the second and third LPSs as well as on the final test and the size of the advantage did not vary over the number of practice sessions (i.e., the interaction for the ANOVA was not significant). Thus, there was no evidence that restudying on its own caught up to retrieval in terms of recall performance when the two learning strategies were repeated and spaced over several sessions. In short, once the memory rating was eliminated in Experiment 3, robust retrieval practice effects were obtained regardless of the number of spaced repetitions.

That said, spaced restudying also produced reasonably good performance in Experiment 3. Although Dunlosky et al. (2013) concluded that restudying was of low utility, this conclusion was primarily based on research that had a single restudy

session that occurred in close temporal proximity to initial study. In Experiment 3, we tested recall accuracy for items that were restudied once, twice, or thrice over spaced LPSs and performance was quite respectable, reaching 0.55 after being restudied thrice. Three sessions of spaced retrieval practice only improved accuracy over three sessions of spaced restudying by approximately 0.09, to 0.64. Moreover, we did not include a memory rating in Experiment 3 at all because we wanted a "pure" comparison of spaced restudying and spaced retrieval practice. In both Experiments 1 and 2, restudying coupled with a memory rating produced the best performance of all. Conceivably, had a memory rating requirement been added to the spaced restudy condition in Experiment 3, recall accuracy may have increased further still.

## General Discussion

In three experiments, we investigated both spaced retrieval practice and spaced restudying as methods to learn Swahili–English translation pairs. In all experiments, following an initial study phase, participants took part in three LPSs during which they practiced restudying or retrieving the word pairs multiple times, and then took a final recall test. All sessions were spaced two days apart.

In short, there are three key results from our research. First, response format during retrieval (i.e., covert vs. overt) had little effect on final-test performance (Experiment 1). Second, if participants were required to rate whether they would have been able to recall the intact word pairs while restudying them, spaced restudying produced good final test performance (Experiments 1 and 2). Indeed, performance in the spaced restudying condition exceeded performance in the spaced retrieval practice condition in which participants also evaluated the accuracy of their retrieval attempts. Third, even without a concomitant memory rating, the spaced restudying condition produced good learning, but performance on the final test was poorer when compared with spaced retrieval practice. That is, without the memory rating, the standard retrieval practice effect (i.e., retrieval > restudying) was observed both on the final test, and during the LPSs (Experiment 3). In what follows, we elaborate on these key results.

### Response Format

One overarching aim of our research was to determine the effect of response format (i.e., covert vs. overt retrieval) on spaced retrieval practice. Although prior research has compared response formats with single session learning, none has examined how response format might impact spaced retrieval practice. In Experiment 1, we manipulated the completeness of covert retrieval by varying participants' expectations about the need to make an overt response. Following others (e.g., Putnam & Roediger, 2013; Sundqvist et al., 2017), we reasoned that if participants could predict that an overt response was unlikely, such as if covert trials were blocked (blocked-50% group), or if the sequencing was random but likelihood of an overt

response was low (random-25% group), then covert retrieval would be incomplete. Conversely, if the sequencing was random but the likelihood of an overt response was high (random 75% group), then participants' covert retrieval attempts would be more thorough.

However, we did not observe any difference between overt versus covert retrieval in any group. These null results are consistent with prior research on response format where material is learned in a single session rather than over several spaced LPSs (e.g., Putnam & Roediger, 2013; Smith et al., 2013; see Sundqvist et al., 2017 for a meta-analysis). One might have assumed that overt retrieval would yield memory benefits analogous to *producing* (i.e., saying, writing, or typing) responses in research on the *production effect* (e.g., MacLeod & Bodner, 2017). Those benefits are commonly attributed to enhanced distinctiveness of produced items compared to covertly considered ones. Thus, it appears that overt responding in retrieval practice experiments does not create more distinctive and retrievable memory traces than covert responding. The null effect of response format is consistent with Tulving's (1983) GAPS framework, which proposed that covert processing ("thinking about" material) is equivalent to overt processing (responding to explicit questions about the material). The null result also has practical implications in that students who practice covertly retrieving information, perhaps because silence is required (e.g., studying in the library), can enjoy the memory benefits of retrieval practice as much as students who overtly produce the products of their retrieval attempts.

One reason that overt retrieval may not have produced more distinctive memory traces than covert retrieval in Experiment 1 is because we required participants to make some type of overt response on all trials. Like Putnam and Roediger (2013), our participants rated whether their retrieval attempt was successful on covert retrieval trials. This aspect of the procedure ensured that the covert retrieval condition was comparable to the overt retrieval condition in that participants in both conditions made an overt response, but only the overt retrieval condition required an overt response that included the to-be-remembered information. Potentially, without this requirement to rate the success of the covert retrieval attempt, results more akin to the production effect might have obtained.

### Spaced Restudying

Our second major aim of the current research was to investigate the efficacy of spaced restudying as a learning strategy when there are repetitions both within and between the LPSs. As noted earlier, restudying has typically been identified as a low-utility learning technique (Dunlosky et al., 2013). However, its efficacy in learning conditions typical of spaced retrieval practice or successive relearning research, which are characterized by repeated practice over multiple spaced sessions, has received very little attention. Higham et al. (2022) were the first to investigate how successive restudying fares relative to successive relearning when exposure to the material has been equated between the conditions. They found that, although successive relearning produced the best long-term recall, successive restudying also produced performance that was much better than a single study session. Given these

results, the possibility exists that repeated, spaced restudying may be a moderate-to-high-utility learning strategy. Indeed, decades of research on the positive effects of spacing repetitions of to-be-remembered material (*spacing effect*) would suggest that this is a reasonable explanation of our findings (see Benjamin & Tullis, 2010 for a review).

The results of Experiment 1 indeed suggested that spaced restudying was an effective learning strategy. In fact, restudying unexpectedly led to *better* performance than either covert or overt spaced retrieval. We hypothesized that the memory ratings that participants were required to make in Experiment 1 may have had a dissociative effect on restudying versus retrieval performance. In terms of restudying, the hypothetical memory rating ("Would you have remembered the correct translation?") may have encouraged participants to adopt a more effective study strategy than they normally would if they were only reading the material. One possibility is that participants covered the English word on some restudy trials so that the memory rating was easier to make, effectively turning some restudy trials into a retrieval trials. However, we asked participants to report their learning strategy at the end each experiment and only one participant across all three experiments reported using this strategy and their data were excluded because of it. Another way that restudy trials might have been effectively converted to retrieval trials is through *study-phase retrieval* (e.g., Thios & D'Agostino, 1976). That is, the memory rating on restudy trials may have prompted participants to think back to earlier encounters with those same items (i.e., on an earlier LPS or on an earlier pass in the same LPS), encouraging retrieval of the restudied items during the LPSs. However, for both these accounts, it is difficult to explain why retrieving with a memory rating on restudy trials led to better performance in both Experiments 1 and 2 than retrieving with a memory rating in the retrieval condition.

Another potential explanation of superior performance in the restudy-rating condition is that the memory ratings improved factors associated with learning such as attention, confidence and enjoyment, study time, and/or effort while (re)studying the items. The literature on *reactivity* has shown that making JOLs during study sometimes enhances memory for the studied material (see review by Double et al., 2018). For example, Putnam and Roediger (2013) found that asking participants to make JOLs on a 0–100 scale while restudying weakly related English cue-target word pairs boosted later recall. To explain effects like these, Zhao et al. (2022) proposed the *Enhanced Learning Engagement* (ELE) theory of positive JOL reactivity. In support of this theory, Shi et al., (2022) found that making JOLs during study reduced mind wandering, and that mind wandering mediated the JOL reactivity effect. Although JOLs and the restudy memory rating used in the current study are not identical (e.g., JOLs query prospective memory performance whereas our memory rating was retrospective), perhaps both serve to enhance learning engagement, thereby boosting restudy memory performance.

Another potential candidate account of reactivity is the cue-strengthening hypothesis (e.g., Myers et al., 2020; Soderstrom et al., 2015). The cue-strengthening hypothesis posits that making JOLs strengthens the cues that support the JOL judgment (e.g., semantic relatedness between the cue and target). If the subsequent

memory test is sensitive to those cues (e.g., the test taps cue-target relatedness information), then memory performance is enhanced relative to no JOLs.

However, just as with the target-covering and study-phase retrieval explanations described earlier, neither the ELE or cue-strengthening account explains why memory ratings had opposing effects on retrieval and restudy performance. Presumably, making memory ratings during retrieval should also enhance engagement and/or strengthen cues, improving memory performance instead of impairing it. Even if participants were already fully engaged in the retrieval task, neither the ELE account nor the cue-strengthening account has a mechanism to *reduce* performance by adding a memory rating requirement to the task (i.e., negative reactivity). As Rivers et al, (2021) noted with respect to the latter account, "Any amount of negative reactivity is difficult to explain solely with the cue-strengthening hypothesis" (p. 1351).

One candidate explanation that may be able to account for both the positive and negative reactivity occurring simultaneously in the same experiment is the changed-goal hypothesis (Chang & Brainerd, 2023; Mitchum et al., 2016). According to this account, making JOLs highlights the fact that some items are easy to learn whereas others are hard. Being alerted to this fact causes participants to change their goal from trying to learn all the items to focusing on learning the easier items at the expense of the harder ones. In the typical JOL reactivity paradigm, the study list consists of intermixed related and unrelated English words. Therefore, the changed goal in this case would be to concentrate on learning the easy, related items (causing positive reactivity) while disregarding the hard, unrelated items (causing negative reactivity). Of course, our materials were English–Swahili pairs which are unique with respect to reactivity. However, one could speculate that the memory rating caused participants to focus on learning the easy, restudy pairs (which did not require effortful retrieval) at the expense of learning the hard, retrieval pairs (for which retrieval was often unsuccessful), which would be consistent with the core principles of the changed-goal hypothesis.

A second explanation that has the potential to explain how ratings both enhanced restudy performance and impaired retrieval performance is *mediator generation*. Pyc and Rawson (2010) defined a *mediator* as "…a word, phrase, or concept that links a cue to a target" (p. 335). They suggested that participants do not typically generate mediators when restudying, but spontaneously generate mediators on spaced retrieval trials (see also Morehead et al., 2018). However, the requirement to provide a memory rating may change this typical situation. On restudy trials, if participants are asked to rate the hypothetical likelihood that they would be able to recall the English translation, they might more closely examine the relationship between the Swahili cue word and English target word. This assessment may have inadvertently caused participants to generate mediators (e.g., "I would definitely have remembered *cloud* in response to *wingu* because birds use their wings and fly near the clouds"), thereby benefitting performance for restudied items on the final recall test. However, the memory rating for retrieval trials in Experiments 1 and 2 required participants to compare their response to the feedback, assess the degree of correspondence, and then to make an overt 1–100 memory rating. The memory rating in this case may have acted like a secondary task and interfered with spontaneous

mediator generation, thereby compromising final recall performance for rated items if retrieval was required during the LPSs.

Although it is plausible that reactivity effects are caused by differential learning strategies (e.g., mediator generation promoting learning on restudy trials vs. impaired mediator generation reducing learning on retrieval trials), Rivers et al. (2021) asked participants to report the memory strategy that they used when learning word pairs while either making JOLs or not. They found no evidence that the learning strategies were different between the JOL and no-JOL conditions. However, they also noted that the timing of memory strategy reports was not ideal in that participants were asked to report their learning strategies after a final recall test, which could have complicated interpretation of the data. For example, participants may have forgotten the strategy that they used, and such forgetting might be particularly likely if recall was unsuccessful, leading to a confounding of recall success and strategy accessibility. Thus, future research without these problems might further investigate how making JOLs potentially affects the learning strategies that people use. For example, participants might be asked to report their strategies while learning the items. In this vein, Muncer et al. (2021) found in a recent meta-analysis that "online" measures of metacognition (e.g., think-aloud protocols) better predicted adolescents' mathematics learning than "offline" measures (e.g., questionnaires about past behavior).

### Dissociative Effects of Memory Ratings

In Experiment 2, we manipulated the requirement to make memory ratings to more directly examine the possibility that it had a dissociative effect on restudying versus retrieval performance. As predicted, final recall performance in the spaced restudying condition was superior to that in the spaced retrieval practice condition when a memory rating was required, but descriptively worse when the memory rating requirement was absent. However, even with the memory rating requirement removed, a robust retrieval practice effect was not obtained (i.e., the effect was of only marginal significance). This result stands in contrast to a significant body of work on the testing effect (see Rowland, 2014; Yang et al., 2021 for reviews). One possible reason for the weak retrieval practice effect in Experiment 2, even when no memory ratings was required, is that restudying becomes an effective learning strategy when it is repeated over spaced intervals. A single study episode may not be very effective, but after six restudy episodes (three LPSs with two episodes per LPS), restudy performance may effectively "catch up" to six retrieval episodes.

This hypothesis was tested in Experiment 3 where we compared recall performance after a two-day retention interval for items restudied versus retrieved once, twice, or thrice. If restudy performance only becomes effective after multiple, spaced encounters, then there would have been a retrieval advantage after one practice encounter, replicating the typical testing effect obtained in single-session studies, but little difference or even a reversal (i.e., a restudy advantage) after three episodes. However, we did not observe that pattern of results. Instead, retrieval was

superior to restudying regardless of the amount of practice and there was no interaction between condition (restudy and retrieval) and practice frequency (1, 2, and 3).

Why, then, was there no robust testing effect in Experiment 2 when a memory rating was not required? One possible explanation for the weak testing effect in Experiment 2 can be derived from the changed-goal hypothesis. That is, because all participants rated both restudy and retrieval items in the first LPS (i.e., condition was manipulated within subjects), they may have changed their goal early on to focus on learning the restudy items at the expense of the retrieval items (cf. Rivers et al., 2021). Another possibility is that the need to complete memory ratings on some trials in the same list contaminated the no-rating condition. Hence, some participants may have continued to make ratings covertly on no-rating trials, thereby effectively turning them into rating trials. Such covert activity would be particularly likely on later LPSs if participants became metacognitively aware that the ratings were helping their recall performance for restudied items. If participants continued to make memory ratings on restudy trials even though it was not a requirement, restudy performance may have been inadvertently boosted because of mediator generation, reducing the effect of retrieval practice. However, as we noted earlier, the rating and no-rating conditions in Experiment 2 were blocked thereby reducing the potential for ratings to contaminate the no-rating condition (i.e., perhaps limited to the first few trials in the second block). We also question whether participants would be willing to engage in covert memory ratings when it was not a task requirement, particularly given that the experiment was labor-intensive enough as it was. Nonetheless, contamination of the no-rating condition remains a possibility and future research might manipulate the rating requirement between subjects to alleviate this concern.

### Implications of the Results and Future Directions

Although more work needs to be done, the opposing effects of memory ratings on restudy and retrieval performance limits the number of candidate mechanisms that could explain the results, facilitating future research on the topic. Thus, we consider the current series of experiments a good foundation for developing our understanding of the processes underpinning learning via restudying versus retrieval in both successive relearning and spaced retrieval practice paradigms, and reactivity effects on memory more generally. On the applied side, students are sometimes given practice quizzes, and because they are formative, they may be asked to mark their own (or another student's) answers after being provided with feedback. Marking quiz answers, particularly if part marks are allowed, is essentially the same task as rating the accuracy of a retrieval attempt on a 100-point scale, which was participants' task in Experiments 1 and 2. Similarly, some learning apps (e.g., Remember More: https://www.remembermore.app) require users to judge whether their answers are correct after a retrieval attempt. If this activity undermines the beneficial effects of retrieval practice and/or feedback learning, as our experiments suggest, then educators and app developers may need to rethink such practice.

There are several questions that future research might address. First, neither the changed-goal nor the mediator generation hypotheses that we have introduced here

to explain the dissociative effect of memory ratings on restudy and retrieval performance have direct evidential support. We proposed them here as potential explanations for the serendipitous finding that restudy performance was superior to retrieval performance in Experiment 1. And although Experiment 2 clearly indicated that memory ratings moderated the efficacy of restudying and retrieving, it did not produce direct evidence in support of either hypothesis. Direct support might come from examining the learning strategies that participants report while restudying and retrieving with and without memory ratings (e.g., Morehead et al., 2018; Rivers et al., 2021). Alternatively, with respect to the mediator hypothesis, participants might be asked to recall any mediators that they may have used (e.g., Pyc & Rawson, 2010). If the mediator hypothesis has any validity, we anticipate that the use of mediators would be more evident in the restudy-rating and retrieval-no-rating conditions than in the restudy-no-rating and retrieval-rating conditions, respectively.

A second avenue for future research has a more applied focus. As noted earlier, the best performance across the three experiments we have reported was in the restudy condition when a memory rating was required. If this effect is robust and applies to a variety of materials, it could potentially be exploited as a learning strategy to use in real classrooms. However, before promoting such a strategy, it is important to determine the boundary conditions of the benefit. For example, further research would need to investigate whether restudying with memory ratings is beneficial for learning with material other than English word translations. Memory for key term definitions or expository texts, which have clear educational relevance, may not benefit in the same way.

Another consideration is the student learning experience. Higham et al. (2022) found in their classroom study that students' self-reports of mastery, attentional control, and anxious affect tended to change very little over spaced LPSs if they were successively restudying. In contrast, students who were successively relearning (i.e., retrieving) showed positive change. Specifically, in the first LPS, students reported increased feelings of anxiety and lower perceptions of mastery and attentional control when asked to retrieve (vs. restudy) information. However, as recall performance improved over subsequent LPSs, reports of anxiety reduced while perceptions of mastery and attentional control increased. By the final LPS, retrieving was associated with lower reports of anxious affect, better mastery, and better attentional control than restudying. Additionally, student feedback provided at the end of the semester indicated that, overall, they preferred retrieving to restudying course material.

Importantly, however, these results were obtained with "pure" restudying, without any rating. Potentially, the addition of the rating may yield both better learning and a better learning experience. Only future research can determine whether these outcomes will obtain. At the very least, the results presented here suggest that restudying with rating is a strategy that can be exploited in applied setting such as the learning of foreign vocabularies and can easily be implemented in apps and website that have been designed for this purpose.

**Funding** This work was supported in part by a grant from the Economic and Social Research Council (ESRC), ES/T013664/1 awarded to Philip A. Higham, Julie A. Hadwin, Rosalind Potts, and

Kou Murayama and Leverhulme Trust Research Leadership Award (RL-2016-030) awarded to Kou Murayama.

## Declarations

**Conflict of Interest** The authors declare no competing interests.



## References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review, 33*(2), 693–712. https://doi.org/10.1007/s10648-020-09556-8

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*, 296–308. https://doi.org/10.1037/0096-3445.108.3.296

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*(5), 316–321. https://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*(4), 566–577. https://doi.org/10.1016/j.jml.2005.01.012

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

Carleton, R. N., Norton, M. A. P. J., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders, 21*(1), 105–117. https://doi.org/10.1016/j.janxdis.2006.03.014

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Chang, M., & Brainerd, C. J. (2023). Changed-goal or cue-strengthening? Examining the reactivity of judgments of learning with the dual-retrieval model. *Metacognition and Learning, 18*(1), 183–217. https://doi.org/10.1007/s11409-022-09321-y

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology, 111*(2), 225–236. https://doi.org/10.1037/0021-843X.111.2.225

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of. Psychology, 40*, 104.

Higham, P. A., Zengel, B., Bartlett, L. K., & Hadwin, J. A. (2022). The benefits of successive relearning on multiple learning outcomes. *Journal of Educational Psychology, 114*(5), 928–944. https://doi.org/10.1037/edu0000693

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hope, R. M. (2022). *Rmisc: Ryan miscellaneous. R package version 1.5.*1. https://cran.r-project.org/package=Rmisc

Janes, J. L., Dunlosky, J., Rawson, K. A., & Jasnow, A. (2020). Successive relearning improves performance on a high-stakes exam in a difficult biopsychology course. *Applied Cognitive Psychology, 34*(5), 1118–1132. https://doi.org/10.1002/acp.3699

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review, 25*(6), 2356–2364. https://doi.org/10.3758/s13423-018-1463-4

Jönsson, F. U., Kubik, V., Larsson Sundqvist, M., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research, 78*(5), 623–633. https://doi.org/10.1007/s00426-013-0522-8

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education, 47*(7), 674–682. https://doi.org/10.1111/medu.12141

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review, 33*(3), 959–987. https://doi.org/10.1007/s10648-020-09572-8

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*(4), 390–395. https://doi.org/10.1177/0963721417691356

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200–219. https://doi.org/10.1037/a0039923

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-4.4.* https://cran.r-project.org/package=BayesFactor

Muncer, G., Higham, P. A., Gosling, C. J., Cortese, S., Wood-Downie, H., & Hadwin, J. A. (2021). A meta-analysis investigating the association between metacognition and math performance in adolescence. *Educational Psychology Review*. https://doi.org/10.1007/s10648-021-09620-x

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition, 48*(5), 745–758. https://doi.org/10.3758/s13421-020-01025-5

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2*(3), 325–335. https://doi.org/10.1080/09658219408258951

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36–48. https://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*(3), 283–302. https://doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General, 142*(4), 1113–1129. https://doi.org/10.1037/a0030498

Rawson, K. A., & Dunlosky, J. (2022). Successive relearning: An underexplored but potent technique for obtaining and maintaining knowledge. *Current Directions in Psychological Science, 31*(4), 362–368. https://doi.org/10.1177/09637214221100484

Rawson, K. A., Dunlosky, J., & Janes, J. L. (2020). All good things must come to an end: A potential boundary condition on the potency of successive relearning. *Educational Psychology Review, 32*(3), 851–871. https://doi.org/10.1007/s10648-020-09528-y

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory, 29*(10), 1342–1353. https://doi.org/10.1080/09658211.2021.1985143

Rowbotham, M., & Schmitz, G. M. (2013). Development and validation of a student self-efficacy scale. *Journal of Nursing Care, 2*, 126. https://doi.org/10.4172/2167-1168.1000126

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution, 1*(2), 114–117. https://doi.org/10.1111/j.2041-210X.2010.00014.x

Shi, A., Xu, C., Zhao, W., Shanks, D. R., Hu, X., Luo, L., & Yang, C. (2022). Judgments of learning reactively facilitate visual memory by enhancing learning engagement. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02174-1

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

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

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553–558. https://doi.org/10.1037/a0038388

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Sundqvist, M. L., Mäntylä, T., & Jönsson, F. U. (2017). Assessing boundary conditions of the testing effect: On the relative efficacy of covert vs. overt retrieval. *Frontiers in Psychology, 8*, 1018. https://doi.org/10.3389/fpsyg.2017.01018

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging, 34*(6), 836–847. https://doi.org/10.1037/pag0000376

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition, 7*(1), 106–115. https://doi.org/10.1016/j.jarmac.2016.10.004

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15*(5), 529–536. https://doi.org/10.1016/0022-5371(76)90047-5

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C. (2022). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development, 93*(2), 405–417. https://doi.org/10.1111/cdev.13689

Supplemental materials are available at: https://osf.io/39p7w/?view_only=d2cfd37736d04bb6a57acf467a170e04