The Spanish Journal of Psychology (2020), **23**, e31, 1–22.
© Universidad Complutense de Madrid and Colegio Oficial de Psicólogos de Madrid 2020
doi:10.1017/SJP.2020.33

Does Item Difficulty Affect the Magnitude of the Retrieval Practice Effect? An Evaluation of the Retrieval Effort Hypothesis

Marcos Felipe Rodrigues de Lima, Sebastião Venâncio, Júlia Feminella and Luciano Grüdtner Buratto

Universidade de Brasília (Brazil)

Abstract. Retrieving information by testing improves subsequent retention more than restudy, a phenomenon known as the *retrieval practice effect.* According to the *retrieval effort hypothesis* (REH), difficult items require more retrieval effort than easier items and, consequently, should benefit more from retrieval practice. In two experiments, we tested this prediction. Participants learned sets of easy and difficult Swahili–Portuguese word pairs (study phase) and repeatedly restudied half of these items and repeatedly retrieval practiced the other half (practice phase). Forty-eight hours later, they took a cued-recall test (final test phase). In both experiments, we replicated both the retrieval practice and the item difficulty effects. In Experiment 1 ($N = 51$), we found a greater retrieval practice effect for easy items, $M_{\text{Difference}} = .26$, $SD = .17$, than for difficult items, $M_{\text{Difference}} = .19$, $SD = .19$, $t(50) = 2.01$, $p = .05$, $d = 0.28$. In Experiment 2 ($N = 28$), we found a nonsignificant trend—$F(1, 27) = 2.86$, $p = .10$, $\eta_p^2 = .10$—toward a greater retrieval practice effect for difficult items, $M_{\text{Difference}} = .28$, $SD = .22$, than for easy items, $M_{\text{Difference}} = .18$, $SD = .21$. This was especially true for individuals who benefit from retrieval practice (difficult: $M_{\text{Difference}} = .32$, $SD = .18$; easy: $M_{\text{Difference}} = .20$, $SD = .20$), $t(24) = -2.08$, $p = .05$, $d = -0.42$. The results provide no clear evidence for the REH and are discussed in relation to current accounts of the retrieval practice effect.

Received 3 August 2019; Revised 2 July 2020; Accepted 10 July 2020

Keywords: cued recall, desirable difficulties, retrieval practice, retrieval effort, testing effect

Introduction

A student has just read an introductory chapter in a cognitive psychology textbook. She plans to have another study session the next day and wonders what would be the best way to go through said chapter again in order to boost her long-term memory. A growing body of research has shown that retrieving information by testing improves its subsequent retention more than restudy, a phenomenon known as *retrieval practice effect* (Whiffen & Karpicke, 2017). It is assumed that tests are useful because they allow learners to engage in retrieval processes (Karpicke, 2017), which alter memory representations of the practiced items, making them more recallable in the future (Bjork, 1994).

The traditional procedure used to investigate retrieval practice effects involves three phases. After the initial study of the items (*study phase*), a *practice phase* takes place, in which learners either restudy them or perform an initial test that aims to induce the retrieval practice (Karpicke, 2017). In the *final test phase*, the learners perform a final memory test (*criterion test*) that refers to all previously studied items. Mnemonic benefits of the retrieval practice are indicated by better performance on the criterion test for previously retrieval practiced items than for restudied items. A meta-analysis indicated that, in 81% of the studies analyzed, retrieval practice led to a better performance on the criterion test than restudy (Hedges' $g = 0.50$; Rowland, 2014). The benefits of retrieval practice have been observed even after a short 5-min final test (Guran et al., 2020; but see Roediger & Karpicke, 2006). These

Correspondence concerning this article should be addressed to Marcos Felipe Rodrigues de Lima. Universidade de Brasília. Departamento de Processos Psicológicos Básicos. Campus Darcy Ribeiro, Instituto de Psicologia, Sala ASS-12/5. 70910-900. Brasília, DF (Brazil).

E-mail: lima.piraju@gmail.com

Authors would like to thank Beatriz A. Cavendish, for her helpful comments on a previous version of the article, and Gabriela Y. Iwama, for her suggestions and comments regarding mixed logit regression models

Conflicts of Interest: None

This work was based on a Master's thesis submitted to the University of Brasília, and it was supported by the National Council for Scientific and Technological Development (CNPq).

How to cite this article:
Lima, M. F. R., Venâncio, S., Feminella, J., & Buratto, L. G. (2020). Does item difficulty affect the magnitude of the retrieval practice effect? an evaluation of the retrieval effort hypothesis. *The Spanish Journal of Psychology, 23.* e31. Doi:10.1017/SJP.2020.33

benefits occur across a wide range of materials, contexts, criterion tests, and learners' characteristics (Dunlosky et al., 2013), and tend to be even greater when feedback is provided during retrieval practice, especially under conditions in which initial retrieval rate is lower (Rowland, 2014). Furthermore, retrieval practice seems promising in educational settings (Moreira et al., 2019) and in cognitive rehabilitation of patients with language (Middleton et al., 2016) and memory impairments (Sumowski et al., 2010).

Several hypotheses have been proposed to account for the retrieval practice effect, ranging from descriptive to explanatory accounts (for a review, see Karpicke et al., 2014). These accounts differ on the purported cognitive mechanism underlying the benefits of the retrieval practice. However, they agree on the idea that an initial test involves greater *cognitive effort* than restudy. Some authors suggest that it is this effort that is responsible for the beneficial effects of retrieval practice (e.g., Bjork, 1994; Pyc & Rawson, 2009). *Effort* is usually a vaguely defined construct. In attentional capacity models, cognitive effort is the proportion of processing dedicated to perform a task given a limited capacity central, which can allocate processing capacity in a highly flexible manner (Kahneman, 1973). Although this is an abstract definition, it is understood that the allocation of effort varies between tasks, depending on the manipulation of task difficulty.

The desirable difficulties framework, an influential idea in the learning and memory literature, posits that greater memory gains are expected in conditions that require greater retrieval effort from the learner (Bjork, 1994). Such conditions include spaced practice (Cepeda et al., 2008) and retrieval practice (Roediger & Karpicke, 2006), among others. In the latter case, long-term benefits of retrieval practice tend to be greater due to the fact that retrieval is, allegedly, a more difficult process (Bjork, 1994). This is exactly the core of the *retrieval effort hypothesis* (REH), a descriptive account derived from the desirable difficulties framework. The REH predicts that successful, more difficult retrievals will yield greater memory benefits than successful, easier ones (Pyc & Rawson, 2009). Rowland's (2014) meta-analysis supported this prediction, showing that more difficult initial tests (i.e., free and cued-recall) produce greater retrieval practice effects (g = 0.81 and 0.72, respectively) than less difficult ones (i.e., recognition; g = 0.36).

Task and Item Difficulty

The difficulty of a given retrieval task has been variously operationalized by manipulating (a) the degree of informativeness of a cue at the practice phase (e.g., Carpenter & DeLosh, 2006), (b) the time interval between successive retrieval attempts (e.g., Middleton et al., 2016; Pyc & Rawson, 2009), (c) the attentional demands imposed during the practice phase (e.g., Buchin & Mulligan, 2019; Gaspelin et al., 2013), and (d) the number of times an item was required to be correctly recalled (i.e., *criterion level*; e.g., Pyc & Rawson, 2009; Vaughn et al., 2013). A series of results that indicate better performance on the criterion test for items initially tested under more difficult conditions support the REH (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009).

A distinct way of operationalizing difficulty is selecting items from normative studies that provide information about item difficulty (e.g., Lima & Buratto, 2019; Nelson & Dunlosky, 1994). When the data set comprises two languages, the choice of the foreign language is based on its desirable features, namely, it should (a) be unknown or unusual to the learners, (b) share few cognates with the learners' native tongue, (c) be written in the same alphabet as the learner's native language, and (d) not produce floor and ceiling effects (see Nelson & Dunlosky, 1994). The aforementioned normative studies used a multitrial learning paradigm, in which the learner engages in a predetermined number of successive study–test cycles. In a study block, the learner must study a set of intact pairs (e.g., *wingu–cloud*), whereas in a test block, she must recall the target word (*cloud*), given the cue (e.g., *wingu–?*). The main measure in these studies is the proportion of participants who correctly recalled the target word across test blocks (Lima & Buratto, 2019). This proportion provides an estimate of the relative difficulty of the pair. When the item difficulty is known, participants tend to recall more easy items than difficult ones (e.g., Cull & Zechmeister, 1994), a phenomenon called *item difficulty effect*.

Nelson and Dunlosky's (1994) norms have been widely used in the retrieval practice research. Most studies have used normative measures to establish experimental control, balancing item difficulty across experimental conditions (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2009). Only a few studies have used normative measures as an independent variable (e.g., Carpenter, 2009; Minear et al., 2018; Vaughn et al., 2013).[1] Vaughn et al. argued that at high criterion levels, retrieval practice would benefit difficult items more than easy items, since across practice cycles there would be a decrease in retrieval effort for easy items but not for difficult ones. In two experiments, they found that across several criterion levels, performance on the criterion test was better for easy items than for difficult ones, contrary to REH predictions.

[1] Carpenter (2009) used norms of associative strength, which provide an estimate of the probability of producing the target given the cue. Although this estimate was not based on performance on a multitrial learning paradigm, it is also a measure of pair difficulty, since paired associate learning is affected by the degree of relatedness between cues and targets.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

Carpenter (2009) manipulated the associative strength between cue and target word pairs and, consequently, the difficulty of item retrieval. These pairs were either restudied or retrieval practiced. In two experiments, it was observed that the advantage that pairs with a strong cue had over pairs with a weak cue in the initial test was either reversed (Experiment 1) or eliminated (Experiment 2) on the criterion test. Moreover, in Carpenter's Experiment 1, the proportion of correctly recalled items during the final test was greater for pairs with weak cues than for pairs with strong cues, but only in the retrieval practice condition. This suggests that different levels of retrieval effort induced by the initial test were responsible for this effect (see Carpenter, 2009, Table 3). In an investigation about metamemory, Cull and Zechmeister (1994, Experiment 2) found an analogous interaction, although this should be interpreted with caution, since a self-paced procedure was used, leading to different exposure times under different conditions, which may partly explain the results.

In sum, difficulty can be operationalized either as the amount of demands placed on the learner in the practice phase (e.g., Pyc & Rawson, 2009) or as the relative item difficulty, based upon normative studies (e.g., Vaughn et al., 2013). In both cases, it is assumed that more difficult tasks and items tend to require more effort than their easier counterparts (Kahneman, 1973). Although the REH has gained empirical support in studies that investigated task difficulty (Carpenter & DeLosh, 2006; Pyc & Rawson, 2009), mixed results were obtained in studies that investigated item difficulty (Carpenter, 2009; Vaughn et al., 2013). A possible explanation for these divergent results is the way evidence was produced: on the one hand, Carpenter's experiments factorially crossed type of practice (restudy and retrieval practice) with item difficulty (weak and strong cues); on the other hand, Vaughn et al. compared easy and difficult items across several criterion levels, under the assumption that only difficult items would benefit further at higher criterion levels, since retrieval would still involve effort on later retrieval attempts.

Present Study

Here, we adopted a similar approach to Carpenter's (2009) experiments, crossing factorially type of practice and item difficulty. Unlike Carpenter, we used a longer retention interval (48 hr instead of 5 min) and repeated practice for each item (instead of only one presentation). Like Vaughn et al. (2013), we tested REH predictions, albeit addressing a slightly different question: Does item difficulty affect the magnitude of the retrieval practice effect? If so, which items benefit most from retrieval practice: easy or difficult items? These questions are theoretically relevant because they provide a new test of the REH with effort manipulated via item difficulty rather than task difficulty, which has been more usually investigated. These questions are also of particular applied relevance for teachers and educators, as some materials are more difficult to learn than others and require more effort. If retrieval practice can benefit learning of these materials to a similar or higher extent than learning easier materials, teachers and educators may decide to invest time and resources differently to such difficult materials.

In two experiments we sought (a) to replicate both the retrieval practice and item difficulty effects, and (b) to investigate whether item difficulty affects retrieval practice effect sizes. Participants learned a set of word pairs (study phase) and repeatedly restudied half of this set and repeatedly retrieval practiced another half (practice phase). Forty-eight hours later, they took a cued-recall test (final test phase). Three predictions were made about the experimental results. First, it was hypothesized that participants would recall more retrieval practiced items than restudied ones (Dunlosky et al., 2013; Rowland, 2014). Second, it was hypothesized that participants would recall more easy items than difficult ones (Cull & Zechmeister, 1994). Third, based on the REH (Pyc & Rawson, 2009) and on previous experimental results (e.g., Carpenter, 2009), it was hypothesized that the retrieval practice effect would be greater for difficult items when compared to easy ones.

Experiment 1

Method

Participants and design

Fifty-two undergraduates from the University of Brasília were recruited to take part in the experiment. Sample size was based on Pyc and Rawson (2012), because their Experiment 1b used stimuli, retention intervals, and initial and final tests similar to the present design. One participant was excluded prior to data analysis because she failed to return for the second session. Thus 51 participants comprised the final sample (females = 46; age range = 18–32 years, *M* = 20.29, *SD* = 3.05). All participants were native Brazilian Portuguese speakers and gave written informed consent. The experiment followed a 2 × 2 factorial design, with the factors of type of practice (restudy, retrieval practice) and difficulty (easy, difficult) both manipulated within-participants.

Materials

Forty Swahili–Portuguese word pairs were selected. Based on the memorability normative measures provided by Lima and Buratto (2019), twenty pairs were

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

labeled as *easy* (M = .60, SD = .10), and 20 pairs as *difficult* (M = .24, SD = .05; see Appendix A). Word pairs were divided into two sets, each one with 10 easy and 10 difficult items. Both sets were equated in terms of familiarity, concreteness, arousal, valence for Portuguese words; wordlikeness (similarity to Portuguese) for the Swahili words, and difficulty (ts ≤ 0.78, ps ≥ .44). The assignment of both sets to experimental conditions was counterbalanced across participants. Additionally, twenty math problems, 10 easy (e.g., 7 × 8) and 10 difficult (e.g., 17 × 18), were created.[2] Instructions and materials were presented on a computer screen controlled with PsychoPy (Peirce, 2007).

Procedure

Figure 1a depicts a general schematic representation of Experiments 1 and 2. In Experiment 1, at the beginning of the first session, participants completed an initial training task, in which stimuli were unrelated to Swahili–Portuguese word pairs. This task aimed to train participants on how to use the keyboard and to help them understand the feedback that would be provided throughout this session. Next, in the study phase, participants were presented with 40 Swahili–Portuguese word pairs in random order. Each trial began with a fixation cross on the center of the screen for 4 s, which was followed by the presentation of a word pair, also on the center of the screen (Swahili word on top; Portuguese word below). Participants were instructed to study the pairs. After the study phase and after each practice cycle, participants engaged on a distracter task, which consisted of four math problems. Each distracter task cycle lasted 1 min.

After the first distracter task cycle, participants were informed that all word pairs would be practiced again by one of two different methods (“Method A” or “Method B”). Assignment of method name (A, B) to type of practice (restudy, retrieval practice) was counterbalanced across participants. Examples of trials on both types of practice are depicted in Figure 1b. In both types of practice, each trial started with a fixation cross, with the same location and presentation time of the study phase. On retrieval practice trials, the Swahili word, alongside question marks that replaced the Portuguese word, were presented for 9 s. Participants were then asked to covertly recall the Portuguese translation of the Swahili word. After 6 s of the sole presentation of the Swahili word, four alternatives (three letters and one question mark) were presented on the bottom of the screen. Participants were then asked to press the keyboard arrow that represented the penultimate letter of the Portuguese word they recalled.[3] They had 3 s to indicate their response, which was followed by a 2-s feedback consisting of one of three symbols, indicating whether the response was correct or incorrect (a check mark and an X mark, respectively; see Figure 1b), or whether there was no response at all (a warning sign). Additionally, during the feedback, the correct Portuguese word replaced the question marks on the screen, giving participants a new opportunity to encode the correct translation of the given Swahili word. We chose the penultimate letter for two reasons: (a) The first letter could potentially encourage a strategy in which alternatives could be used as retrieval cues (see Wing et al., 2013); and (b) given the fact that Portuguese words tend to end with a strict set of letters, when asked to indicate the last letter of a word, the participant’s range of potential responses would be rather limited. On restudy trials, word pairs were presented intact for 11 s each. Participants were instructed to use these trials as an additional opportunity to study the pairs. After 6 s onset of the word pair, four alternatives (similar to the retrieval practice trials) were presented on the bottom of the screen, amongst which participants should indicate the penultimate letter of the Portuguese word. They had 3 s to indicate their response, which was once more followed by a 2 s feedback. In both types of practice, participants were encouraged to select “?” if they were not sure of the answer. Word pairs were presented in a random order, the position of the three alternatives (only the letters) was also randomized, and type of practice was mixed across trials. Four cycles of practice were performed in the practice phase.

After the last cycle of the distracter task (after the fourth practice cycle), we assessed participants’ metacognitive knowledge of the effectiveness of both types of practice by having them make two judgments of learning (JOLs). Participants estimated what percentage of Swahili words they believed they would remember two days later. These two judgments were made on a 0–100 scale (0 = *I think I’ll remember nothing*; 100 = *I think I’ll remember all*). Participants saw images representing each method (“A” and “B”) to ensure that they would make the judgment based on the appropriate method. Upon finishing the JOLs, participants were reminded to

[2] We chose easy and difficult math problems because we originally intended to measure participants’ pupil size as a function of task difficulty. There is evidence that the eye’s pupil dilates more while participants perform more difficult tasks than easier ones (see Kahneman, 1973). The math problems would thus serve both as a retention interval filler and as a control task to assess pupil size sensitivity for our eye-tracker.

[3] This procedure was adapted from Wing et al. (2013), as our original aim was to measure pupil size with an eye-tracker. Consequently, we collected discrete responses instead of a full typed response, in both conditions. This minor design feature should not affect the results, as retrieval practice effects also occur when participants make covert recall (van den Broek et al., 2014) or when they emit a discrete response on the keyboard (Racsmány et al., 2018, Experiment 1). A full description of how alternative answers were created was presented in Appendix B.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

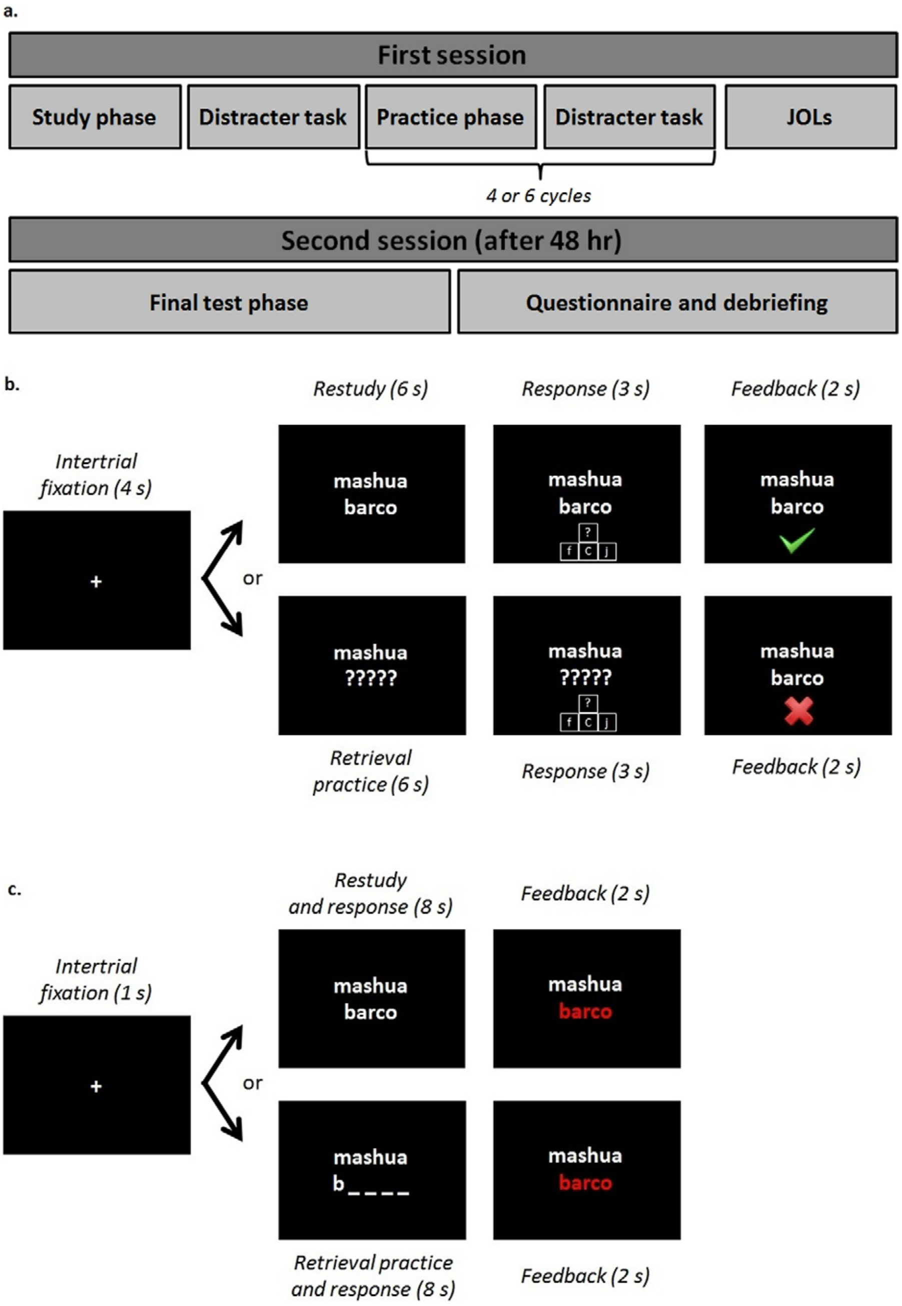

Figure 1. (a) General schematic representation for Experiments 1 and 2. Examples of trials on practice phase for restudy and retrieval practice conditions are depicted for both (b) Experiment 1 and (c) Experiment 2.

return to the lab two days later and dismissed. Forty-eight hours after the first session (range = 42–53 hr), participants returned to the lab. The second session started with a criterion cued-recall test. On each trial, participants were prompted with a Swahili word and were asked to recall its Portuguese translation. Participants typed each word onto the computer, and after they pressed the "Enter" key, the next trial began. The maximum response time allowed on each trial was 15 s. All 40 items previously studied were tested and no feedback was provided. The order of items was randomized. After this task, participants answered a brief questionnaire, were debriefed, thanked, and dismissed.

Statistical Analyses

Overview. An alpha level of .05 was used for all statistical tests, unless otherwise stated. When the assumption of sphericity was violated, as indicated by Mauchly's test, the Greenhouse–Geisser correction was applied to adjust for degrees of freedom. Measures of effect size

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

were reported as Cohen's *d* (*t*-tests), as partial eta-squared (η_p^2; analyses of variance [ANOVAs]), or as log odds (β; mixed logit regression models), when appropriate.

Practice phase. It should be noted that, for retrieval practiced items, correct answers represent learning across cycles, whereas for restudied items, correct answers only indicate that participants paid attention to the task throughout the cycles. Since the participants' responses represented different cognitive processes for each type of practice, we conducted two 2 (difficulty) × 4 (cycle) repeated measures ANOVA, separately for each type of practice, using the proportion of correct answers as our dependent variable. Additionally, following previous studies (e.g., Pyc & Rawson, 2009; Vaughn et al., 2013), we used reaction times (RT) on practice cycles as a measure of task difficulty during practice cycles. Here we are interested in investigating whether our intended manipulation of difficulty was successful, based on Lima and Buratto's (2019) normative measure of difficulty. RT is here defined as the time between the onset of the screen showing response alternatives and the participant's response. We computed median RT for each participant (by condition), considering all trials.[4] Next, we entered these median RT scores in two 2 (difficulty) × 4 (cycle) repeated measures ANOVA, separately for each type of practice. It was hypothesized that, if our manipulation was successful, RTs would be higher for difficult items than for easy items, but only for retrieval practice trials. Interactions were further explored using either paired-samples *t*-test or one-way ANOVAs (Bonferroni corrected), as appropriate. JOLs were analyzed using a paired-samples *t*-test. Exploratory analyses unrelated to the main hypotheses of this study (e.g., those who sought to test alternative explanations for the results) are presented in Appendix C.

Final test phase. Two independent judges were trained to assess the participants' responses on the final test phase. They were blinded to what condition each item pertained to. To assess inter-rater agreement, Cohen's kappa (κ) was computed. Performance on the final test was defined as the proportion of items correctly recalled. To test main effects, the retrieval practice effect was defined as better performance on the criterion test for previously retrieval practiced items than for restudied items. To test the Type of Practice × Difficulty interaction, we first computed retrieval practice effects as difference scores (i.e., $\text{Performance}_{\text{retrieval practice}} - \text{Performance}_{\text{restudy}}$), calculated separately for easy and difficult items. Positive values indicate retrieval practice effects. Next, we compared these scores with a paired-samples *t*-test. We also used the RT as an alternative index of performance on the final test (Racsmány et al., 2018; van den Broek et al., 2014). RT represents the time between stimulus onset (cue word) and participant's first keypress. Both measures were analyzed using 2 (type of practice) × 2 (difficulty) repeated measures ANOVAs, followed by paired-samples *t*-tests, when the interaction term was significant.

Conditional probability analyses. Finally, we further analyzed performance on the final test using conditional analyses (see Finley et al., 2011; van den Broek et al., 2014, for similar procedures). We computed the conditional probability of a correct recall on the final test, given that an item was correctly answered *n* times on practice cycles. We modeled these data using two mixed logit regression models, one for restudied and another for retrieval practiced items (for rationale, see Jaeger, 2008). Fixed effects for difficulty and number of correct answers were estimated in the model (using mean centered variables), with random intercepts for each participant. The model was defined by

$$\text{logit}(p(\mathrm{Y}_{ij})) = b_{0j} + b_1\mathrm{D}_{ij} + b_2\mathrm{C}_{ij} + b_3\mathrm{DC}_{ij} \tag{1}$$

$$b_{0j} = b_0 + u_{0j} \tag{2}$$

where logit($p(\mathrm{Y}_{ij})$) is the log odds of a participant *j* correct recalling an item *i* on final test, b_0 is the fixed intercept, u_{0j} is the variance estimate of participants from the fixed intercept, b_1, b_2, and b_3 are regression coefficients according to the item difficulty, *D* (where 0 = *difficult*, 1 = *easy*), the number of correct answers on practice phase, *C* (ranging from 0 to 4), and the interaction between the item difficulty and the number of correct answers during the practice phase, *DC* (defined as $D \times C$), respectively. These models were compared with base models, which estimate whether the odds of recall and non-recall vary between participants.[5]

[4] Unless otherwise stated, analyses using only correct answers yielded a similar pattern of results. Analyses with all the answers were conducted to prevent the loss of statistical power due to listwise exclusion of missing cases in repeated-measures ANOVA.

[5] We modeled our data separately for restudied and retrieval practiced items because, as already mentioned, we assume that the number of correct answers on practice phase indexed different cognitive processes. Although some authors suggest that modeling could be done jointly, our aim was to overcome limitations related to the different successful recall rates between easy and difficult items in the practice phase (see Results). Specifically, in the retrieval practice model, if the advantage of easy over difficult items was mainly driven by retrieval success, we would expect a significant Difficulty × Number of Correct Answers ($D \times C$) interaction. Because our main interest was on the retrieval practice model, analyses on restudied items were conducted only for control purposes and completeness. The non-inclusion of items as random intercepts is justified due to concerns about multicollinearity, since difficulty and items were associated. Alternatively, in both Experiments 1 and 2, we also modeled our data replacing dichotomous difficulty variable by

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

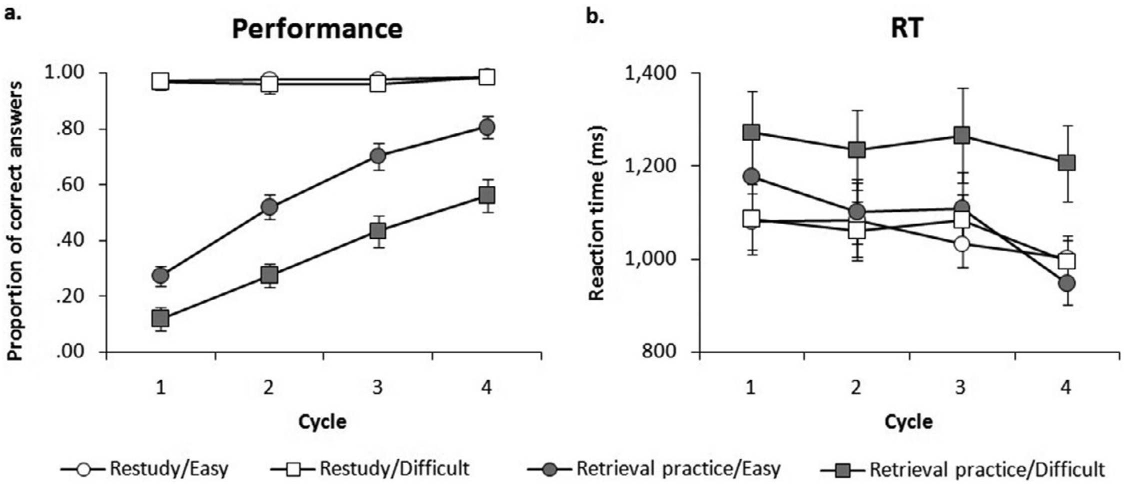

Figure 2. Data from practice cycles of Experiment 1. (a) Proportion of correct answers across the four cycles of the practice phase. (b) Reaction time across the four cycles of the practice phase. Error bars represent the 95% withinparticipant confidence interval (Cousineau, 2005).

Likelihood-ratio tests were used to test if additional variables improved the model when compared to base models.

Results

Practice Phase

Performance on practice cycles. Figure 2a depicts the proportion of correct answers on cycles of the practice phase. For retrieval practiced items, there were main effects of difficulty, $F(1, 50) = 127.14$, $p < .001$, $\eta^2_p = .72$, and cycle, $F(1.91, 98.53) = 204.83$, $p < .001$, $\eta^2_p = .80$. The effect of difficulty indicates that easy items ($M = .58$, $SD = .19$) were better learned than difficult items ($M = .35$, $SD = .20$), whereas the effect of cycle showed that all pairwise comparisons were significant, $ps < .001$ ($M_{\text{Cycle 1}} = .20$, $SD = .13$, $M_{\text{Cycle 2}} = .40$, $SD = .20$, $M_{\text{Cycle 3}} = .58$, $SD = .24$, $M_{\text{Cycle 4}} = .68$, $SD = .24$), indicating an increasing linear trend across the four cycles. Importantly, there was a Difficulty × Cycle interaction, $F(2.34, 117.13) = 4.29$, $p = .01$, $\eta^2_p = .08$. The advantage for easy items over difficult ones was lower in the first cycle, $t(50) = 5.27$, $p < .001$, $d = 0.74$, than in the other cycles, $ts(50) \geq 8.26$, $ps < .001$, $ds \geq 1.16$. We will return to this interaction later in the discussion. For restudied items, there were no significant effects, $Fs \leq 2.32$, $ps \geq .10$.

RT on practice cycles. Figure 2b depicts average RT on cycles of the practice phase. For restudied items, there was only a main effect of cycle, $F(2.52, 126.03) = 4.07$, $p = .01$, $\eta^2_p = .08$. This effect showed that, on average, median RT was shorter for the fourth cycle (998 ms) than for the first, second, and third cycles (1,084 ms, 1,072 ms, and 1,058 ms, respectively), all $ps < .05$. For retrieval practiced items, there were main effects of difficulty, $F(1, 50) = 26.68$, $p < .001$, $\eta^2_p = .35$, and cycle, $F(3, 150) = 5.622$, $p = .001$, $\eta^2_p = .10$. These effects showed that, on average, median RT was shorter (a) for easy items than for difficult ones (1,083 ms vs. 1,244 ms), and (b) for the fourth cycle (1,077 ms) than for the first and third cycles (1,224 ms and 1,187 ms, respectively), all $ps < .02$. Other comparisons were not significant, $Fs \leq 2.14$, $ps \geq .10$. Taken together, these results suggest that retrieval was more effortful for difficult items, as indexed by RTs.

JOLs. The upper side of Table 1 shows the number of participants that judged they would remember more restudied items, more retrieval practiced items, or an equal number of items on both conditions (i.e., a tie). Participants' average JOLs (converted into proportions) for restudied and retrieval practiced items were almost identical (.41 vs. .42), a non-significant difference, $t(50) = .37$, $p = .71$, $d = 0.05$.

Final Test Phase

Scoring. The two judges showed high level of agreement on scorings, $\kappa = .97$, $p < .001$, which could be considered almost perfect (Landis & Koch, 1977). The scores of one of the judges were randomly selected and used on subsequent analyses.

Performance on final test. Following Minear et al.'s (2018) recommendations, we report the proportion of participants showing different patterns of performance (see Table 1, "Experiment 1—All items"). Retrieval practice effects were more frequent for easy items than for difficult ones (.84 vs. .75). Moreover, when we consider all

available normed data as a continuous variable (see Memorability, Appendix A). The results from these analyses led to the same conclusions, so we do not mention them further.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

Table 1. *Number of Participants Showing Different Patterns of JOLs and Performance in Experiments 1 and 2*

	Advantage		
Measure	Retrieval practice	Restudy	Tie
JOLs			
Experiment 1	21 (.41)	25 (.49)	5 (.10)
Experiment 2	21 (.75)	4 (.14)	3 (.11)
Performance			
Experiment 1—Easy	43 (.84)	0 (.00)	8 (.16)
Experiment 1—Difficult[a]	38 (.75)	4 (.08)	9 (.18)
Experiment 1—All items	50 (.98)	1 (.02)	0 (.00)
Experiment 2—Easy[a]	20 (.71)	6 (.21)	2 (.07)
Experiment 2—Difficult	23 (.82)	1 (.04)	4 (.14)
Experiment 2—All items	25 (.89)	1 (.04)	2 (.07)

Note. Sample proportions are reported in parentheses.

[a] Sum of proportions does not total 1.00 due to rounding.

items, almost all participants (.98) showed retrieval practice effects. Figure 3a depicts recall performance on the final test. We found significant main effects of type of practice, $F(1, 50) = 159.70$, $p < .001$, $\eta^2_p = .76$, and difficulty, $F(1, 50) = 171.62$, $p < .001$, $\eta^2_p = .77$. The main effect of type of practice reflects overall higher recall in the retrieval practice condition ($M = .52$, $SD = .26$) than in the restudy condition ($M = .30$, $SD = .22$), whereas the main effect of difficulty shows that recall was higher for easy items ($M = .56$, $SD = .27$) than for difficult ones ($M = .26$, $SD = .23$). The Type of Practice × Difficulty interaction was significant, $F(1, 50) = 4.05$, $p = .05$, $\eta^2_p = .08$, which revealed that the retrieval practice effect was greater for easy items ($M_{\text{Difference}} = .26$, $SD = .17$) than for difficult ones ($M_{\text{Difference}} = .19$, $SD = .19$), $t(50) = 2.01$, $p = .05$, $d = 0.28$ (see Figure 3a).[6]

RT on final test. Figure 3b depicts RT on the final test. We found significant main effects of type of practice, $F(1, 46) = 16.82$, $p < .001$, $\eta^2_p = .27$, and difficulty, $F(1, 46) = 37.98$, $p < .001$, $\eta^2_p = .45$. The effect of type of practice reflects overall shorter RT in the retrieval practice condition than in the restudy condition (3,338 ms vs. 4,171 ms), whereas the effect of difficulty shows that RT was shorter for easy items than difficult ones (3,110 ms vs. 4,399 ms). The Type of Practice × Difficulty interaction was not significant, $F(1, 46) = 1.42$, $p = .24$, $\eta^2_p = .03$.

Conditional probability analyses. Figure 4 depicts the probability of correct recall on the final test, given the difficulty and the number of correct answers on practice cycles (see Finley et al., 2011; van den Broek et al., 2014, for similar procedures). The center of each bubble corresponds to the probability of recall for a given item, whereas bubble diameter represents the proportion of cases falling into each category. Comparisons of final models against base models, using likelihood-ratio tests, indicated that the addition of fixed and random terms improved the prediction for both restudy, $\chi^2(3) = 113.80$, $p < .001$, and retrieval practice models, $\chi^2(3) = 357.22$, $p < .001$.

Table 2 shows model summaries. For the restudy model, difficulty was a significant predictor of successful final recall. The odds favored recall of easy items over difficult ones on the final test ($OR = 5.61$). Other predictors were non-significant. For the retrieval practice model, difficulty was also a significant predictor, with the odds of recalling an easy item on the final test greater than the odds of recalling a difficult one ($OR = 3.71$). Furthermore, the odds of correct recall increased with the number of correct answers ($OR = 2.92$). Difficulty did not interact with the number of correct answers ($OR = 0.93$), which suggests that, across different number of correct answers, easy items still are favored over difficult ones. In sum, difficulty predicts successful recall on the final test for both models, but only for the retrieval practice model the number of

[6] We also ran a 2 × 2 analysis of covariance (ANCOVA) using as a covariate the mean difference of easy over difficult items on the retrieval practice condition on the practice phase. When controlling for mean difference on practice phase, both type of practice, $F(1, 49) = 32.75$, $p < .001$, $\eta^2_p = .40$, and difficulty, $F(1, 49) = 16.76$, $p < .001$, $\eta^2_p = .26$, remained significant, although with smaller effect sizes. The Type of Practice × Difficulty interaction, however, was no longer significant after controlling for mean difference on practice phase ($F < 1$, $p = .77$, $\eta^2_p = .002$). The covariate interacted with difficulty, $F(1, 49) = 30.67$, $p < .001$, $\eta^2_p = .39$. Covariate median split ($Mdn = .23$) showed that the difficulty effect was greater for participants who had a greater advantage for easy items on practice phase, $t(24) = 12.38$, $p < .001$, $d = 2.43$, than for participants who had a smaller advantage for easy items on practice phase, $t(24) = 8.42$, $p < .001$, $d = 1.68$. Taken together, these results suggest that the difficulty effect was partially – but not totally – due to learning rates in the practice phase. Furthermore, the ANCOVA suggests that the Type of Practice × Difficulty interaction can be an artifact of these learning rates.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

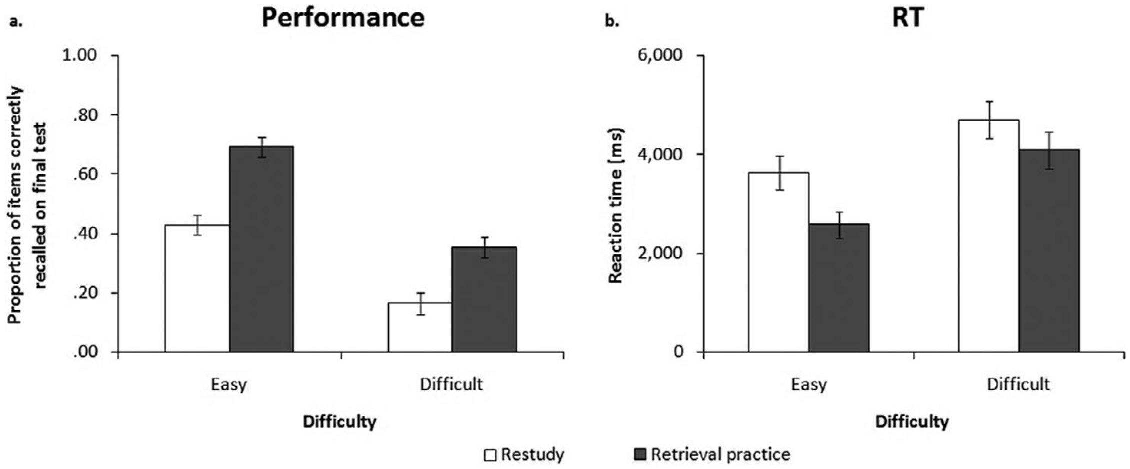

Figure 3. (a) Proportion of correct recall and (b) Reaction time on the final test of Experiment 1. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

correct answers had a predictive value. This result suggests that the benefits of retrieval practice are partially conditioned on successful retrieval during the practice phase, although this result should be interpreted with caution, because this is a post-hoc analysis.

Discussion

Experiment 1 sought (a) to replicate both the retrieval practice and item difficulty effects, and (b) to investigate whether item difficulty affects the retrieval practice effect sizes. The first aim was achieved, replicating previous studies (Cull & Zechmeister, 1994; Dunlosky et al., 2013; Rowland, 2014). Contrary to our hypothesis, however, the magnitude of the retrieval practice effect favored easy over difficult items, not the other way around. One limitation of Experiment 1 is that easy and difficult items had very different proportions of correct answers in the four practice cycles. This indicates that difficult items probably were not as well-learned as easy ones (see Figure 2a). The fixed number of practice cycles used in the present study tends to favor easier items due to the greater proportion of initial recall in the retrieval practice condition (Vaughn et al., 2013). Although the presence of feedback could partially overcome this problem (Finley et al., 2011; see also Tse et al., 2010), the design of Experiment 1 does not allow disentangling the relative contributions of retrieval effort (indexed by item difficulty) and retrieval success (indexed by proportion of correct answers on practice cycles) to final recall performance. Experiment 2 sought to balance the initial performance on the last practice cycle in order to eliminate this confounding factor present in the Experiment 1.

Experiment 2

In this experiment, easy items were practiced four times, whereas difficult items were practiced six times. If the direction of the interaction effect observed on Experiment 1 was due to different learning rates for easy and difficult items, a reversion of the interaction direction would be expected on Experiment 2, supporting the REH (Pyc & Rawson, 2009). If retrieval effort is not related to retrieval practice effect, then either no interaction should be found or an interaction favoring easy items over difficult ones should be found.

Method

Participants, Design, and Materials

We ran an a priori power analysis using G*Power.1.9.4 (Faul et al., 2007), with power set at .95, alpha level set at .05, and Cohen's f = .29 (equivalent to our Type of Practice × Difficulty effect size, η_p^2 = .08), which suggested a sample of 28 participants. Thirty-three undergraduates were recruited from the University of Brasília. In total, five participants had to be excluded, three of them because they failed to return for the second session, one participant due to power outage during the session, and one due to failure to follow the instructions. Thus, the final sample size consisted of 28 participants, as suggested by the a priori power analysis (females = 17; age range = 17–34 years, M = 20.29, SD = 3.87). All participants were native Brazilian Portuguese speakers and gave written informed consent. Design and materials were the same as those used in the Experiment 1.

Procedure

The study phase was the same as in Experiment 1, except that the fixation cross lasted only 1 s. Participants then engaged on a distracter task for 1 min (also presented after each practice cycle). At the beginning of the practice phase, participants were informed that all word

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

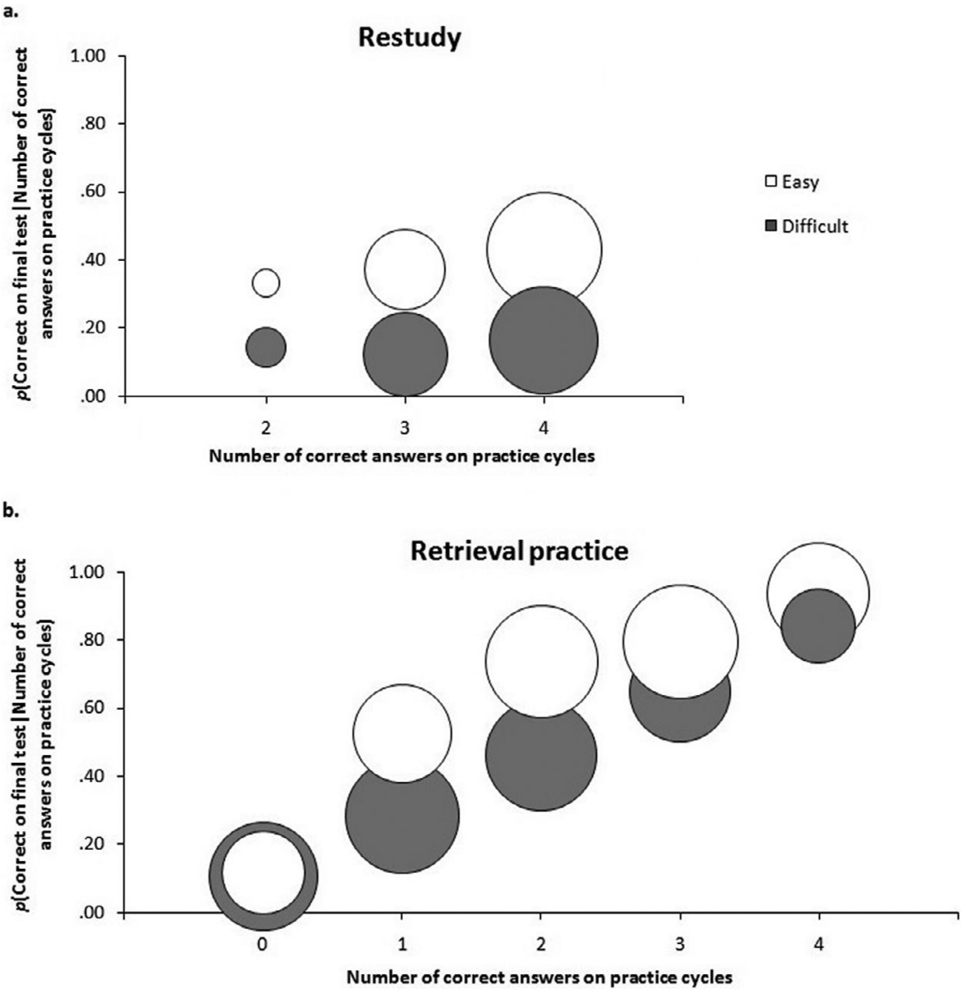

Figure 4. Conditional probabilities depicting the probability of correct recall on final test, given number of correct answers on practice cycles (range = 0–4 times). (a) Restudy condition and (b) retrieval practicecondition.

Table 2. *Fixed Effects for the Mixed Logit Regression Models Predicting Correct Final Recall in Experiment 1*

Fixed effects	β	*SE*	Wald *Z*	*p*
Restudy model[a]				
Intercept	–1.28	0.22	–5.96	< .001
Number of correct answers	0.19	0.27	0.70	.48
Difficulty	1.72	0.18	9.79	< .001
Interaction	0.71	0.52	1.37	.17
Retrieval practice model[b]				
Intercept	0.21	0.19	1.14	.25
Number of correct answers	1.07	0.09	12.41	< .001
Difficulty	1.31	0.19	6.96	< .001
Interaction	–0.04	0.15	–0.26	.80

Note. Beta represents log odds.

[a] Random effect variance (u_{0j}) = 1.88

[b] Random effect variance (u_{0j}) = 1.35

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

Table 3. *Fixed Effects for the Mixed Logit Regression Models Predicting Correct Final Recall in Experiment 2*

			Wald	
Fixed effects	β	*SE*	*Z*	*p*
Restudy model[a]				
Intercept	–0.81	0.34	–2.36	.02
Number of correct answers	0.41	0.31	1.30	.19
Difficulty	2.32	0.66	3.53	< .001
Interaction	0.36	0.53	0.68	.50
Retrieval practice model[b]				
Intercept	0.40	0.15	2.64	.008
Number of correct answers	0.68	0.08	8.00	< .001
Difficulty	1.61	0.23	6.98	< .001
Interaction	0.54	0.17	3.28	< .001

Note. Beta represents log odds.
[a] Random effect variance (u_{0j}) = 0.89.
[b] Random effect variance (u_{0j}) = 0.29.

pairs would be practiced again by one of two different methods. In both types of practice, each trial started with a fixation cross, with the same location and presentation time of the study phase. On retrieval practice trials, the Swahili word, alongside the cue containing the first letter of the Portuguese word (for an example, see Figure 1c), were presented for 8 s. Participants were then asked to recall and type the Portuguese translation of the Swahili word, which was followed by feedback (2 s), consisting of the replacement of the given cue for the correct Portuguese word in red color. On restudy trials, word pairs were presented for 10 s each. Participants were instructed to use these trials as an additional opportunity to study the pairs. They also typed the Portuguese translation of the Swahili word. In the last 2 s, the Portuguese word's color changed from white to red, to balance this feature between conditions. Participants practiced all word pairs for four cycles, but difficult pairs were practiced for two additional cycles. After the last cycle of the distracter task, participants made JOLs on a 0–100 scale and they were reminded to return to the lab two days later. The second session occurred 48 hr after the first session (range = 46–53 hr) and was identical to the second session in Experiment 1.

Statistical Analyses

The same analyses were conducted in Experiment 2. Again, data were modeled using two mixed logit regression models. The only difference in these models is that the number of correct answers on practice phase, *C*, ranged from 0 to 6. In addition, since *p*-values do not provide support for H_0 (Dienes, 2014), we computed the *Bayes Factor* (BF_{10}) in our analyses regarding practice and final test phases. Briefly, BF_{10} informs the strength of evidence in favor of H_1 relative to H_0. Dienes states that values smaller than 0.33 provide support for H_0, whereas values higher than 3 provide support for H_1.

Results

Scoring

The two judges showed an almost perfect level of agreement on scoring, both in the practice (κ = .97) and in the final test phases (κ = .98), *p*s < .001 (Landis & Koch, 1977). The scores of one of the judges were randomly selected and used on subsequent analyses.

Practice Phase

Performance on practice cycles. Figure 5a depicts the proportion of correct answers across cycles of the practice phase. Considering only the first four cycles, for restudied items, there was a significant effect of cycle, *F* (1.13, 30.57) = 8.65, *p* = .005, η_p^2 = .24. Performance was worse on the first cycle (*M* = .83, *SD* = .27) compared to the other cycles ($M_{\text{Cycle 2}}$ = .99, *SD* = .03, $M_{\text{Cycle 3}}$ = .99, *SD* = .04, $M_{\text{Cycle 4}}$ = .98, *SD* = .06), all *p*s < .05. The other effects were non-significant, *F*s ≤ 1.02, *p*s ≥ .37, all $\eta_p^2 \leq$ .04. Like in Experiment 1, for retrieval practiced items, there were significant effects of difficulty, *F*(1, 27) = 39.75, *p* < .001, η_p^2 = .60, and cycle, *F*(3,81) = 309.98, *p* < .001, η_p^2 = .92, when considered only the first four cycles. The effect of difficulty indicates an advantage for easy (*M* = .60, *SD* = .14) over difficult items (*M* = .42, *SD* = .15). The effect of cycle indicates significant increases in the proportion of correct answers across cycles ($M_{\text{Cycle 1}}$ = .18, *SD* = .12, $M_{\text{Cycle 2}}$ = .42, *SD* = .16, $M_{\text{Cycle 3}}$ = .65, *SD* = .14, $M_{\text{Cycle 4}}$ = .81, *SD* = .15), all *p*s < .001. The Difficulty × Cycle interaction was also significant, *F*(3, 81) = 4.01, *p* = .01, η_p^2 = .13, a pattern also found in Experiment 1. The advantage for easy items over difficult ones was lower for the first, *t*(27) = 3.29, *p* = .003, *d* = 0.62, and fourth cycles, *t*(27) = 3.20, *p* = .003, *d* = 0.61, than for the second, *t*(27) = 5.45, *p* < .001, *d* = 1.03, and third cycles, *t*(27) = 6.43, *p* < .001, *d* = 1.22. When we compared items in the last cycle of retrieval practice (easy, cycle 4 vs. difficult, cycle 6), we found no significant difference between easy and difficult items, *t*(27) = –1.99, *p* = .06, *d* = –0.38, BF_{10} = 1.11. In addition, for retrieval practiced items, we also summarized performance across all four or all six cycles, for easy and difficult items, respectively. Our aim was to test whether, across cycles, easy and difficult items led to similar levels of successful retrievals. Because the number of practice cycles differed across conditions, we analyzed our data using both the proportion of correct recalls and the average number of correct recalls per item across cycles. Regarding the proportion of correct recalls, we found a non-significant

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

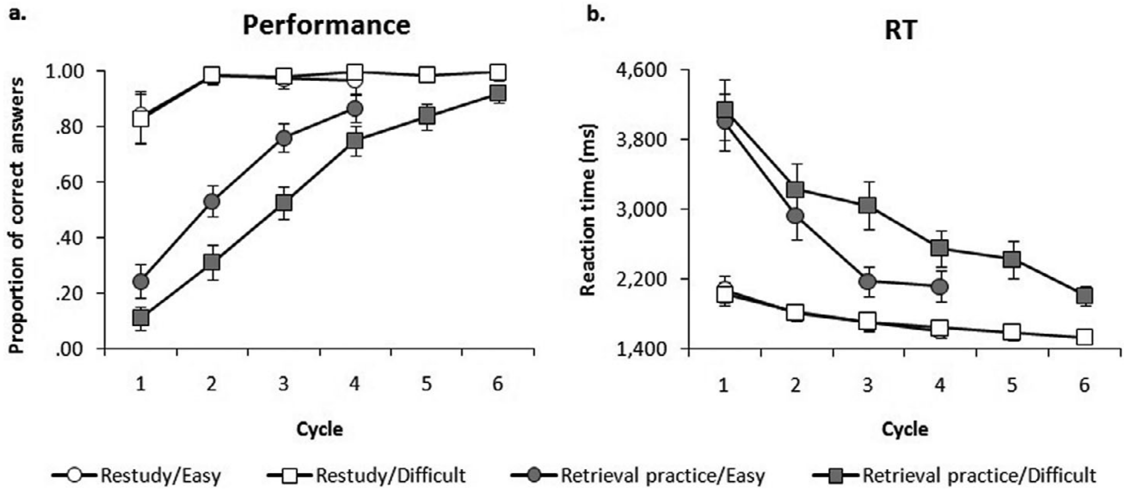

Figure 5. Data from practice cycles of Experiment 2. (a) Proportion of correct answers across the six cycles of the practice phase. (b) Reaction time across the six cycles of the practice phase for correct responses. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

advantage of easy items ($M_{\text{all cycles}}$ = .60, SD = .14) over difficult ones ($M_{\text{all cycles}}$ = .58, SD = .13), t(27) = 1.09, p = .29, d = 0.21. However, regarding the average number of correct recalls, difficult items were recalled significantly more times (M = 3.45 times, SD = 0.79) than easy items (M = 2.41 times, SD = 0.57), t(27) = –7.87, p < .001, d = –1.49. We will return to these seemingly contradictory results in the General Discussion.

RT on practice cycles. Figure 5b depicts average RT on cycles of the practice phase. We again computed median RT for each participant, considering all trials.[7] We conducted two 2 (difficulty) × 4 (cycle) repeated measures ANOVA, separately for each type of practice. Since our aim was to check the effectiveness of retrieval effort manipulation and the RT was available for both difficulty levels up to the fourth cycle, we entered only the first four cycles in these analyses. For restudied items, there was only a main effect of cycle, F(1.66, 44.66) = 34.53, p< .001, η_p^2 = .56. This effect showed that, on average, median RT decreased across cycle 1 (2,053 ms), cycle 2 (1,811 ms), cycle 3 (1,706 ms), and cycle 4 (1,622 ms), all ps < .001. For retrieval practiced items, there were main effects of difficulty, F(1, 27) = 11.01, p < .001, η_p^2 = .29, and cycle, F(2.14, 57.79) = 69.09, p< .001, η_p^2 = .72. These effects showed that, on average, median RT (a) was shorter for easy items than for difficult ones (2,804 ms vs. 3,239 ms), and (b) decreased across cycle 1 (4,069 ms), cycle 2 (3,074 ms), cycle 3 (2,609 ms), and cycle 4 (2,335 ms), all ps < .001. There was a Difficulty × Cycle interaction, F(2.40, 64.79) = 3.78, p = .02, η_p^2 = .12, which indicates that decreasing linear trend was steeper for easy items (η_p^2 = .80) than for difficult ones (η_p^2 = .62). These results suggest again that retrieval was more effortful for difficult items, as indexed by RTs.

JOLs. The upper side of Table 1 shows the number of participants that judged they would remember more restudied items, more retrieval practiced items, or an equal number of items on both conditions (i.e., a tie). Contrary to Experiment 1, participants' average JOLs (converted into proportions) was higher for retrieval practiced items than for restudied ones (.55 vs. .38), t (27) = 3.62, p = .001, d = 0.69.

Final Test Phase

Performance on final test. The bottom side of Table 1 shows different patterns of performance. Unlike Experiment 1, retrieval practice effects in Experiment 2 were more frequent for difficult items than for easy ones (.82 vs. 71). Moreover, when all items are considered, most participants (.89) showed retrieval practice effects. Figure 6a depicts recall performance on the final test. We found main effects of type of practice, F(1, 27) = 72.19, p < .001, η_p^2 = .73, BF_{10} = 5.83 × 10^5, and difficulty, F(1, 27) = 61.75, p < .001, η_p^2 = .70, BF_{10} = 1.14 × 10^5. The main effect of type of practice reflects overall higher recall in the retrieval practice condition (M = .55, SD = .18) than in the restudy condition (M = .32, SD = .20), whereas the main effect of difficulty shows that recall was higher for easy items (M = .54, SD = .20) than for difficult ones (M = .33, SD = .18). More important, the Type of Practice × Difficulty interaction was not significant, F(1, 27) = 2.86, p = .10, η_p^2 = .10, BF_{10} = 0.95.

[7] In four cases, participants had missing RT data because they did not respond in any trials of a given condition. In these cases, the missing values were replaced by the mean of all participants in that condition.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

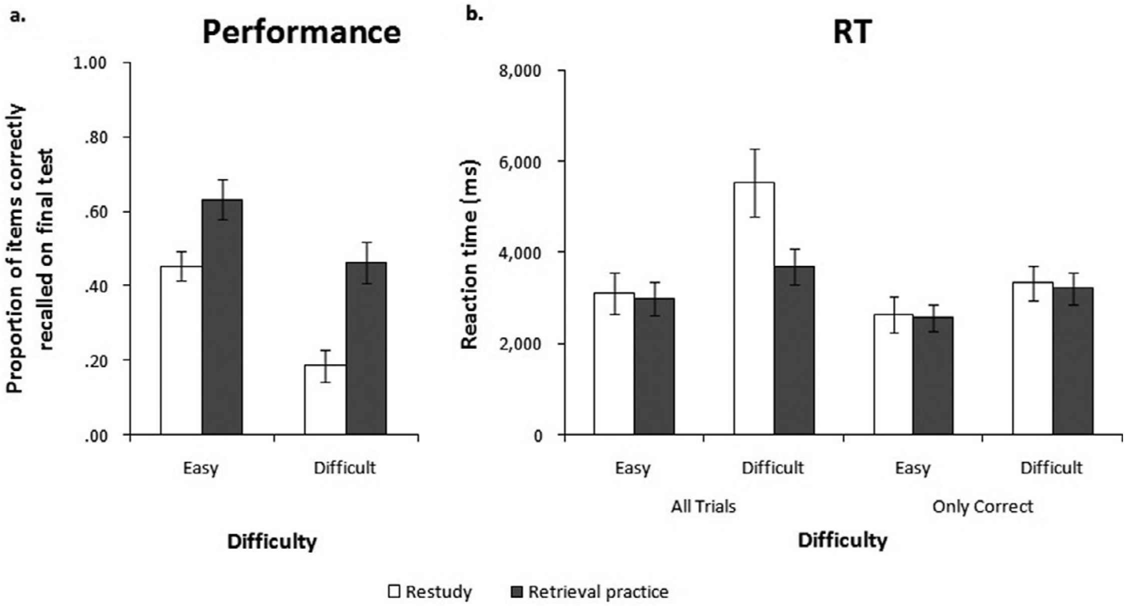

Figure 6. (a) Proportion of correct recall and (b) Reaction time on the final test of Experiment 2. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

Nonetheless, a trend towards interaction was observed in the opposite direction to that in Experiment 1, which suggests that the retrieval practice effect was higher for difficult items ($M_{\text{Difference}}$ = .28, *SD* = .22), than for easy ones ($M_{\text{Difference}}$ = .18, *SD* = .21). We note that this result should be interpreted with caution, as the standard frequentist statistic was not significant and BF_{10} suggests we did not find strong evidence in favor of any of the hypotheses.[8]

A recent study published by Minear et al. (2018) found a non-significant trend toward a greater retrieval practice effect for easy items. However, when they constrained the subsequent analyses to positive testers (i.e., individuals who benefit from retrieval practice), a different result emerged: Participants with high fluid intelligence (gF) showed a greater retrieval practice effect for difficult items than for easy items, whereas low gF showed the opposite pattern. Following Minear et al.'s procedure, we also reanalyzed performance on the final test only for positive testers (considering all items, this analysis was done with 25 participants; see Table 1). In this constrained analysis, the Type of Practice × Difficulty interaction was significant, *F*(1, 24) = 4.34, *p* = .05, η_p^2 = .15, BF_{10} = 2.10, indicating that the retrieval practice effect was greater for difficult items ($M_{\text{Difference}}$ = .32, *SD* = .18) than for easy ones ($M_{\text{Difference}}$ = .20, *SD* = .20), *t*(24) = –2.08, *p* = .05, *d* = –0.42.[9] This finding should be taken cautiously, however, as BF_{10} suggests that there is only anecdotal evidence to choose a model including the interaction term against a model including only the two main effects (see Dienes, 2014).

RT on final test. Figure 6b depicts RT for all trials on the final test. We found main effects of type of practice, *F* (1, 25) = 11.38, *p* = .002, η_p^2 = .31, and difficulty, *F*(1, 25) = 26.97, *p* < .001, η_p^2 = 0,52. The main effect of type of practice reflects overall shorter RT for retrieval practiced items than for restudied ones (3,349 ms vs. 4,322 ms), whereas the main effect of difficulty reflects overall shorter RT for easy items than for difficult ones (3,053 ms vs. 4,618 ms). The Type of Practice × Difficulty interaction was also significant, *F*(1, 25) = 19.12, *p* = .008, η_p^2 = .25. Paired comparisons indicate that, for difficult items, RT was significantly shorter for retrieval practiced items than for restudied ones (3,702 ms vs. 5,533 ms), *t*(25) = –3.56, *p* = .002, *d* = –0.70. For easy items, there were no significant differences in RT between retrieval

[8] We also ran a 2 × 2 ANCOVA, with mean difference of easy over difficult items on retrieval practice condition on practice phase. When controlling for mean difference on practice phase, both type of practice, *F*(1, 26) = 68.72, *p* < .001, η_p^2 = .73, and difficulty, *F*(1, 26) = 56.69, *p* < .001, η_p^2 = .69, remained significant and showed greater effect sizes. The Type of Practice × Difficulty interaction term reminded non-significant after controlling for mean difference on practice phase, *F*(1, 26) = 3.79, *p* = .06, η_p^2 = .13. The covariate was not significant and did not interact with any one of the variables, *F*s ≤ 1.69, *ps* ≥ .21, η_p^2s ≤ .06.

[9] For completeness, we report the main effects of type of practice, *F* (1, 24) = 116.30, *p*< .001, η_p^2 = .83, BF_{10} = 7.74 × 10^6, and difficulty, *F*(1, 24) = 62.12, *p*< .001, η_p^2 = .73, BF_{10} = 5.99 × 10^4. Performance was better for retrieval practiced items (*M* = .56, *SD* = .15) than for restudied ones (*M* = .30, *SD* = .16), and was also better for easy items (*M* = .54, *SD* = .18) than for difficult ones (*M* = .32, *SD* = .14). Thus, main effects were similar for both all participants and positive testers only.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

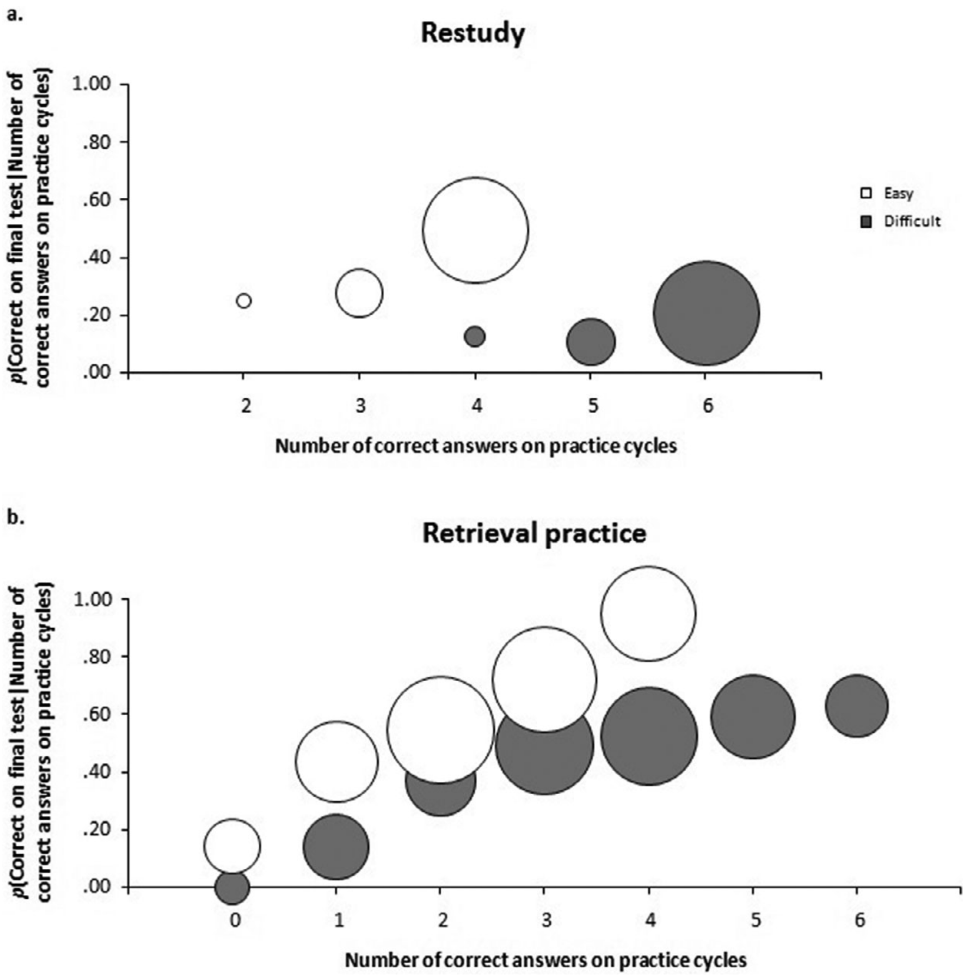

Figure 7. Conditional probabilities depicting the probability of correct recall on final test, given number of correct answers on practice cycles (range = 0–6 times). (a) Restudy condition and (b) retrieval practicecondition.

practiced and restudied items (2,995 ms vs. 3,110 ms), t (25) = –0,41, p= .69,d = –0.08. When analysis was restricted to correct trials,[10] only the difficulty effect remained significant, F(1, 25) = 11.69, p = .002, η_p^2 = 0.32, reflecting overall shorter RT for easy items than for difficult ones (2,608 ms vs. 3,279 ms). The other effects were non-significant, Fs(1, 25) ≤ 0.15, ps ≥ .70, all $\eta_p^2 \leq .006$ (see Figure 6b).

Conditional probability analyses. Figure 7 depicts the probability of correct recall on the final test, given the difficulty and the number of correct answers on practice cycles. When modeling our data with mixed logit regression models, likelihood-ratio tests indicated that the addition of fixed and random terms improved the prediction for both restudy, $\chi^2(3) = 57.69$, $p < .001$, and retrieval practice models, $\chi^2(3) = 94.81$, $p < .001$.

Table 3 shows summaries for these models. Like in Experiment 1, for the restudy model, difficulty was a significant predictor of successful final recall. The odds of recalling an easy item on the final test was more than ten times greater than the odds of recalling a difficult item. For the retrieval practice model, difficulty was also a significant model. The odds of recalling an easy item was five times greater than the odds of recalling a difficult item. Again, the odds of correct recall increased with the number of correct answers (OR = 1.97). The Difficulty × Number of Correct Answers interaction was also significant (OR = 1.72). Increasing linear trends shown in Figure 7b suggest that the advantage of easy over difficult items has increased across different number of correct answers. In sum, like in Experiment 1, this result did not support REH claims.

[10] Six participants did not recall any difficult words in the restudy condition. To reduce the loss of statistical power due to missing cases, they were replaced by the mean of all participants in that condition. We also entered our data into a Linear Mixed Model (LMM). Such models do not require listwise deletion for missing cases and thus allow the inclusion of data from all participants (Hoffman & Rovine, 2007). The results from the LMM analyses and the ANOVAs led to the same conclusions.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

Discussion

Experiment 2 sought to balance initial performance on the practice cycles to eliminate the confounding factor present in the Experiment 1. Averaged across all cycles, the proportion of retrieval success was similar between easy and difficult items. However, averaged across items, the absolute number of correct recalls was greater for difficult items than for easy ones. We replicated both the retrieval practice and the item difficulty effects found in Experiment 1. Unlike Experiment 1, we found a Type of Practice × Difficulty interaction trend (p = .10) in the hypothesized direction, with a slightly greater retrieval practice effect for difficult items than for easy items.

General Discussion

In two experiments, we replicated both the retrieval practice effect and the item difficulty effect. The retrieval practice effect was large and reliable, having been observed in most participants in both Experiments 1 and 2 (.98 and .89, respectively). We also investigated whether item difficulty affects the magnitude of the retrieval practice effect. According to the REH, difficult items require more retrieval effort than easier items and, consequently, should benefit more from retrieval practice. In Experiment 1, we found a greater retrieval practice effect for easy items. In Experiment 2, a non-significant trend emerged in the predicted direction. However, in both cases, recall levels on the practice phase mirrored retrieval practice effects: Easy items outperformed difficult ones in Experiment 1, whereas the opposite pattern was observed in Experiment 2. In the following sections, we elaborate on these findings. First, we discuss in more detail the validity of the item difficulty manipulation. Second, we relate our findings to previous literature, and discuss implications for current memory theory. Third, we discuss other findings of our exploratory analyses. Finally, we present some limitations of our experiments and future directions.

Validity of the Item Difficulty Manipulation

The REH posits that successful, more difficult retrieval attempts will yield greater memory benefits than successful, but easier, retrieval attempts. So, two conditions must be satisfied to test the REH. First, difficulty should vary between retrieval attempts (Pyc & Rawson, 2009). The selection of Swahili–Portuguese word pairs previously normed for item difficulty (Lima & Buratto, 2019) sought to satisfy the first condition. Two findings suggest that this was an effective strategy. First, Lima and Buratto's difficulty measure strongly correlated with performance (on an item-by-item basis) on the final test, both in Experiment 1, r = .88, 95% BCa CI [.82, .93], and 2, r = .72, 95% BCa CI [.56, .83], ps < .001. Second, like in previous studies (Pyc & Rawson, 2009; Vaughn et al., 2013), RT was used as a measure of task difficulty during practice cycles. In both experiments, we found greater RT for difficult items than for easy ones during retrieval practice (but not during restudy), indicating that retrieval effort induction was indeed effective.

The second condition necessary for proper testing of the REH is that retrieval must be successful (Pyc & Rawson, 2009). In Experiment 1, participants had an initial low performance for difficult items on practice phase cycles. Despite the possibility that feedback could mitigate the negative effects of both retrieval failures and intrusion errors (Finley et al., 2011; Tse et al., 2010), a possible explanation for a greater retrieval practice effect for easy items is that difficult items were not as well learned as easy ones. Conditional probability analyses have suggested that easy items, across different number of correct answers in the practice phase, benefit more from retrieval practice than difficult items. However, contrary to previous studies (Pyc & Rawson, 2009; Vaughn et al., 2013), in which number of correct recalls (i.e., criterion level) was directly manipulated, our analyses were post-hoc and therefore should be interpreted with caution. One reason for this concern is that these analyses are susceptible to item selection effects, that is, more retrievable items are also recalled more often in the practice phase (see Buchin & Mulligan, 2019).

Experiment 2 sought to balance average initial performance on the practice cycles in order to eliminate low initial performance as a confounding factor. The goal was to make sure that easy and difficult items were recalled at approximately the same rate in the final cycle of the practice phase. Easy items were practiced four times, whereas difficult items were practiced six times. Averaged across all cycles, the proportion of retrieval success was similar between easy and difficult items. However, averaged across items, the absolute number of correct recalls was greater for difficult items than for easy ones. In the final test, we found a trend pointing to a greater retrieval practice effect for difficult items. Although non-significant, this trend is consistent with the REH.

One possible explanation for the non-significant result is lack of statistical power. Our power analysis was based on η_p^2, a biased estimator, instead of the partial omega-squared, (ω_p^2; see Albers & Lakens, 2018). Had we used ω_p^2 as our estimator, the estimated sample size for Experiment 2 would increase from 28 to 38, and we might have observed a significant interaction in the predicted direction. Alternatively, the trend to reverse the direction of the Type of Practice × Difficulty interaction in Experiment 2 may have been an artifact of the absolute number of

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

correct recalls in the practice phase. It is already known that increasing the number of successful retrievals during practice phase enhances later recall in the final test phase (Pyc & Rawson, 2009; Vaughn et al., 2013). So, although we ensured an approximately similar proportion of recalls across cycles, it is possible that matching the absolute number of recalls is more important in this regard. Unfortunately, the absolute number of recalls was not matched in both experiments. Our procedures represent what Pyc and Rawson (2009) called *duration-based procedures*, in which items are practiced a predetermined number of times. This contrasts to *criterion-based procedures*, in which items are practiced until a criterion is met. Vaughn et al. equated recall levels in the practice phase for easy and difficult items, but the average numbers of trials (i.e., the re-exposure time) to reach criterion differed across item difficulties. Pyc and Rawson (2009) also equated recall levels in the practice phase, but they did not provide the average number of trials to reach criterion. Thus, for both duration- and criterion-based procedures, it seems difficult to equate, at the same time, retrieval success and the exposure time across item difficulty.

In the absence of experimental control of the retrieval success in the practice phase, we again supplemented our results with conditional probability analyses. The retrieval practice model showed that the odds of recalling an easy item was greater than the odds of recalling a difficult item, and that this advantage tended to increase across different number of correct answers (see Figure 7b), contrary to that would be expected according to REH (Pyc & Rawson, 2009).

Relation to Previous Studies and Theoretical Implications

Although the difficulty of retrieval tasks has been investigated in several ways (Buchin & Mulligan, 2019; Carpenter & DeLosh, 2006; Gaspelin et al., 2013; Middleton et al., 2016; Pyc & Rawson, 2009), only a few studies investigated the relationship between retrieval practice effects and item difficulty (Carpenter, 2009; Minear et al., 2018; Vaughn et al., 2013). Whereas Carpenter's (2009) Experiment 1 found a greater retrieval practice effect for weakly related pairs than for strongly related pairs (see Carpenter's Table 3), in two experiments, Vaughn et al. (2013) found that, regardless of the number of times an item was required to be correctly recalled, performance on the criterion test was always better for easy items, contrary to REH prediction.

One key difference in the aforementioned experiments is that Carpenter (2009) manipulated *type of practice* across items, whereas Vaughn et al. (2013) manipulated *criterion level* across items (i.e., there was no restudy vs. retrieval practice contrast in their experiments). In this regard, our manipulations were more similar to Carpenter's design, despite differences in materials, retention intervals, and number of practice cycles. Therefore, it remains unclear whether a greater retrieval practice effect for difficult items can be replicated for retention intervals longer than 5 min. On the other hand, our conditional probability analyses showed that more effortful, successful retrieval did not reverse the advantage for easy items over difficult items. This set of analyses is analogous to Vaughn et al.'s manipulation of criterion level, but since the assignment of items to different number of correct answers in the practice phase was not random, our evidence is purely correlational.

A recent study on individual differences found a three-way interaction (Group × Type of Practice × Difficulty), suggesting that participants with high fluid intelligence (gF) showed a greater retrieval practice effect for difficult items, whereas low gF showed a greater retrieval practice effect for easy ones (Minear et al., 2018). Nonetheless, it should be noted that this pattern had emerged only when their analysis was restricted to positive testers, whereas an analysis with all data showed a non-significant trend toward a greater retrieval practice effect for easy items. When we reanalyzed Experiment 2's data constrained to positive testers, we found a greater retrieval practice effect for difficult items, consistent to REH claims. However, a greater retrieval practice effect for difficult items was not found in Experiment 1, where virtually all participants (.98, see Table 1) were positive testers. Again, one possible explanation for these different patterns is that, since retrieval practice performance on practice cycles was worse for difficult items in Experiment 1, final test performance approached floor effects, preventing us from drawing stronger conclusions.

In light of the foregoing considerations, we conclude that results provide no clear evidence for the REH. At a first look, these mixed results seem at odds with previous studies showing that task difficulty affects the magnitude of the retrieval practice effect (Carpenter & DeLosh, 2006; Pyc & Rawson, 2009). However, Minear et al.'s (2018) findings suggest that we should take into account both item difficulty and learners' skills when introducing desirable difficulties.[11] It is not well clear, for example, why Minear et al.'s study had a smaller proportion of positive testers than ours. In fact, to date, little is known about the relationship between individual differences and the retrieval practice effect. A recent

[11] In their study, negative testers (i.e., individuals who benefit from restudy) outperformed positive testers in the overall final recall performance. Positive and negative testers did not differ in working memory, gF, and crystallized intelligence, but they did differ in self-reported encoding strategies: Positive testers more frequently reported using shallow processing strategies, whereas self-testing was used more frequently by negative testers (Minear et al., 2018).

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

review about individual differences in long-term memory states that relations observed "are inconsistent and are potentially moderated by factors such as the ability range of the sample, difficulty of the items, presence or absence of feedback, delay length, and potentially other factors" (Unsworth, 2019, p. 118). Future research is needed to better examine relations between learners' skills and item difficulty.

One criticism to the REH is that this account does not provide a cognitive mechanism to explain why effortful retrieval benefits memory (Karpicke, 2017). Karpicke et al. (2014) argue that not all difficult (or effortful) retrievals are beneficial to memory. As an example, they mentioned that dividing attention during retrieval practice did not improve memory (e.g., Gaspelin et al., 2013). In fact, some difficulties are *undesirable* difficulties, "if the learner, by virtue of prior knowledge and current cues, is not equipped to respond to them successfully" (Bjork & Kroll, 2015, p. 242; see also Bjork, 1994). To put it another way, it is imperative to balance retrieval success on the one hand and retrieval effort on the other (Finley et al., 2011).

One possible way to reconcile our results with explanatory accounts is to consider the episodic context account, which proposes that effort is important as it leads to context reinstatement (Karpicke et al., 2014). In this sense, we can argue that a longer interval between successive retrieval attempts leads to a greater degree of context reinstatement than a smaller interval, because contextual cues change more during longer intervals than during shorter intervals. It is not clear, however, how item difficulty could engage different degrees of context reinstatement during retrieval practice. Thus, according to the episodic context account, task difficulty effects (e.g., time interval between successive retrieval attempts) can be explained by degree of context reinstatement, whereas there is no clear reason to expect a greater retrieval practice effect for difficult items.

Exploratory Analyses

RT

RT was used as an alternative index of performance on the final test (Racsmány et al., 2018; van den Broek et al., 2014). In both experiments, when we analyzed the RT measure for all trials, we found retrieval practice effect and item difficulty effect. The episodic context account suggests that retrieval practice (but not restudy) helps the learner to restrict the search set of potential targets in subsequent retrieval attempts, including a criterion test. A reduced search set should be translated into shorter RTs for previously retrieval practiced items than for previously restudied ones (see Karpicke et al., 2014), as found in our experiments and elsewhere (Racsmány et al., 2018; van den Broek et al., 2014). However, in Experiment 2, when the analysis was restricted to correct trials, differences in RT between restudy and retrieval practice disappeared. These results suggest that, regarding the RT measure, only Experiment 1 supported the episodic context account.

In a similar vein, both the decreasing linear trends in RT across retrieval practice cycles (see Figures 2b and 5b) and the shorter RT in the criterion test for retrieval practiced items (see Figures 3b and 6b) are consistent with an automatization account of the retrieval practice effect (Racsmány et al., 2018). As automatization takes place, effortful processing decreases. This should entail in shorter RT across cycles as well as shorter RT for retrieval practiced items in the final test, as observed in our experiments. However, it is not yet clear how these two concepts (effort and automatization) are interrelated in the accounts of retrieval practice.

Particularly, in our Experiment 1, RTs were measured in a different way, given that the participant could give an answer only after 6 s (see Figure 1b). Despite this feature of our design, possibly, the decreasing linear trends across cycles were observed because participants could recall items faster. An alternative, related explanation is that participants *learned to respond* faster, when the alternatives had presented. Although cycles effects for both restudied and retrieval practiced items suggests that what increased was *speed of response*, the effect of difficulty observed only for retrieval practiced items suggest that *speed of recall* can also be affected across cycles. Moreover, in Experiment 2, in which response was permitted as long as the stimuli were present on screen (see Figure 1c), the overall RT pattern on practice cycles was similar to that found in Experiment 1.

JOLs

It is noteworthy that participants' JOLs did not differ in Experiment 1 and favored retrieval practice in Experiment 2. Although several between-participant designs showed that JOLs favored restudied over retrieval practiced condition (e.g., Roediger & Karpicke, 2006), Tullis et al. (2013) have suggested that one factor (out of four) that is important for accurate metacognitive judgments is the opportunity to compare different conditions of processing. Hence, JOLs should be more accurate in both within-participant and mixed-list designs, which is the design employed here. Possibly, higher JOLs for retrieval practice condition in Experiment 2 can be an artifact from both higher retrieval success for difficult items achieved in this experiment and more opportunities to compare different types of practice. This hypothesis should be further explored in future studies.

Our design did not allow us to assess learner's metacognitive judgments as a function of item difficulty. To do so, it would be necessary to separate easy and

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

difficult items (i.e., a "blocked-by-difficulty design"), so that participants knew which set of items was being judged. Another way to conduct this analysis would be by using an item-by-item-JOL procedure instead of an aggregate-JOL procedure (Tullis et al., 2013). Although the REH is silent about the metacognitive results after practicing easy and difficult items, it is an interesting question whether these different judgments are accurate, since they can guide subsequent time allocation by learners in self-regulated scenarios.

Individual differences

The only individual-differences variable in our study was the participants' number of fluent languages. One of our exploratory analyses showed that the benefits of retrieval practice are greater for multilinguals than for monolinguals (see Appendix C). Bjork and Kroll's (2015) review proposes that the learner's known languages are active and competing, which may be a desirable difficulty for learning a new vocabulary. However, this finding was not replicated in Experiment 2. Our question about the number of fluent languages spoken did not provide a clear operational definition of what participants should consider as "fluent". Future studies on individual differences in retrieval practice may benefit from a greater control over several language variables (e.g., number of fluent languages and age-of-acquisition), to further explore this relationship.

Our study has some limitations. First, one source of concern could be regarding female–male ratios in both experiments. However, to the best of our knowledge, no study showed a moderator role of sex on retrieval practice effects, so we believe it may not have had a large impact on the results. Second, the presence of feedback introduces indirect effects of retrieval practice. Retrieval practice can benefit subsequent memory both through retrieval processes themselves and through feedback, exposing learners to correct answers (Karpicke, 2017). Despite this drawback, we chose to provide feedback like a previous study (Minear et al., 2018), (a) to mitigate the negative effects of retrieval failures (Finley et al., 2011), and (b) to ensure that participants would have new opportunities to encode the word pairs, thus decreasing the influence of the intrusion errors during practice on subsequent memory performance (Tse et al., 2010). Due to the features of our experimental design, it is not possible to disentangle the relative contributions of the direct benefits of retrieval practice and indirect benefits introduced by feedback. Future studies may attempt to replicate our results addressing the REH by not providing feedback to participants.

Third, we used a cued-recall task in both practice and final test phases, as the REH has been primarily tested through this type of task. It is important, however, to test this hypothesis using other tasks, such as free recall and recognition. Free recall, in particular, may prove an interesting test for the REH because it yields a larger retrieval practice effect than cued-recall and recognition, and it is associated with less frequent use of feedback (Rowland, 2014). Fourth, we may have inserted a confounding factor in our Experiment 2's design, since the number of cycles on practice phase co-varied with the difficulty, namely, four cycles for easy items and six cycles for difficult items. Karpicke and Roediger (2008) showed that, after a first correct recall of a foreign word, repeated retrieval practice is the key factor for improving memory on a delayed criterion test, probably because repeated restudy is a "shallow" encoding strategy (for a discussion about control conditions in retrieval practice literature, see Moreira et al., 2019). If only retrieval practice cycles conceive additional benefits for later memory, we may have inadvertently biased conditions in favor of difficult items in Experiment 2. We were aware that this could be a problem, but we kept this design feature in order to avoid another concern, namely, fatigue effects resulting from a rather long experiment. Future studies could extend our experiments by either (a) equally increasing both the number of successful recalls and the number of cycles for easy and difficult items, so that the items differ only in relation to the type and practice and item difficulty, or (b) comparing retrieval practice with "deeper" encoding strategies (Moreira et al., 2019).

The results presented in this study are also informative for practitioners. Reliable retrieval practice effects for both easy and difficult materials suggest that this technique can be extended to a wide range of materials, not only easier ones – as long as high rates of initial recall of the material are achieved. These findings are important in both educational and clinical contexts. The evidence that retrieval practice boosts retention for easy and difficult items alike has important implications for its use both as a learning tool and as a rehabilitation technique.

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

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

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, *128* (2), 241–252. https://doi.org/ 10.5406/amerjpsyc. 128.2.0241

Buchin, Z. L., & Mulligan, N. W. (2019). Divided attention and the encoding effects of retrieval. *Quarterly Journal of*

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

Experimental Psychology, *72*(10), 2474–2494. https://doi.org/10.1177/1747021819847141

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

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition*, *22*, 249–257. https://doi.org/10.3758/bf03208896

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, Article e00781. https://doi.org/10.3389/fpsyg.2014.00781

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., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*, 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Gaspelin, N., Ruthruff, E., & Pashler, H. (2013). Divided attention: An undesirable difficulty in memory retention. *Memory & Cognition*, *41*(7), 978–988. https://doi.org/10.3758/s13421-013-0326-5

Guran, C.-N. A., Lehmann-Grube, J., & Bunzeck, N. (2020). Retrieval practice improves recollection-based memory over a seven-day period in younger and older adults. *Frontiers in Psychology*, *10*, Article e02997. https://doi.org/10.3389/fpsyg.2019.02997

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kahneman, D. (1973).*Attention and effort*. Prentice Hall.

Karpicke, J. D. (2017).Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (pp. 487–514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. *61*, pp. 237–284). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lima, M. F. R., & Buratto, L. G. (2019). *Norms for familiarity, concreteness, valence, arousal, wordlikeness, and memorability for Swahili–Portuguese word pairs* [Manuscript submitted for publication]. Institute of Psychology, University of Brasília.

Middleton, E. L., Schwartz, M. F., Rawson, K. A., Traut, H., & Verkuilen, J. (2016). Towards a theory of learning for naming rehabilitation: Retrieval practice and spacing effects. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1111–1122. https://doi.org/10.1044/2016_JSLHR-L-15-0303

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education*, *4*, Article e00005. https://doi.org/10.3389/feduc.2019.00005

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

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 737–746. https://doi.org/10.1037/a0026166

Racsmány, M., Szőllősi, Á., & Bencze, D. (2018). Retrieval practice makes procedure from remembering: An automatization account of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44* (1), 157–166. https://doi.org/10.1037/xlm0000423

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

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

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

Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology*, *24* (2), 267–272. https://doi.org/10.1037/a0017533

Tse, C.-S., Balota, D. A., & Roediger, H. L., III. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging*, *25*(4), 833–845. https://doi.org/10.1037/a0019933

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*, 429–442. https://doi.org/10.3758/s13421-012-0274-5

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. https://doi.org/10.1037/bul0000176

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. https://doi.org/10.1080/09658211.2013.831455

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review*, *20*, 1239–1245. https://doi.org/10.3758/s13423-013-0434-z

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia*, *51*(12), 2360–2370. https://doi.org/10.1016/j.neuropsychologia.2013.04.004

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press

Appendices

Appendix A: Experimental Stimuli Used in Experiments 1 and 2

Table A1 *Parameters of Swahili-Portuguese Word Pairs Used in Experiments 1 and 2*

Difficulty	Swahili	Portuguese	English[a]	Memo-rability[b]
Easy				
	roho	alma	soul	.67
	pipa	barril	barrel	.75
	punda	burro	donkey	.70
	mbwa	cachorro	dog	.65
	pombe	cerveja	beer	.65
	elimu	ciência	science	.47
	godoro	colchão	mattress	.47
	goti	joelho	knee	.51
	buu	larva	maggot	.53
	wasaa	lazer	leisure	.59
	nafaka	milho	corn	.51
	wingu	nuvem	cloud	.57
	lulu	pérola	pearl	.73
	nabii	profeta	prophet	.49
	malkia	rainha	queen	.73
	chura	sapo	frog	.47
	chama	sociedade	society	.65
	dafina	tesouro	treasure	.63
	nyanya	tomate	tomato	.75
	usingizi	sono	sleep	.47
Difficult				
	lozi	amêndoa	almond	.22
	nanga	âncora	anchor	.27
	ambo	cola	glue	.25
	iktisadi	economia	economy	.26
	bahasha	envelope	envelope	.21
	samadi	estrume	manure	.25
	ankra	fatura	invoice	.22
	jeraha	ferida	wound	.27
	hadithi	história	story	.28
	bustani	jardim	garden	.26
	yamini	juramento	oath	.25
	ziwa	lago	lake	.28
	hamira	levedura	yeast	.18
	inda	malícia	spite	.29
	utenzi	poema	poem	.23
	lango	portão	gate	.28
	ladha	sabor	flavor	.14
	ruba	sanguessuga	leech	.28
	hariri	seda	silk	.28
	handaki	trincheira	trench	.14

[a] Original English word normed for Nelson and Dunlosky (1994).

[b] Memorability was computed as the average proportion of participants that correctly recalled items over three test cycles (see Lima & Buratto, 2019).

Appendix B: Alternative Answers for Experiment 1

The correct answers were the penultimate letter of the Portuguese word presented to (restudy) or recalled by (retrieval practice) the participant (e.g., for *nuvem*, which means *cloud*, the correct answer was the letter *e*). The number of times a given letter was a correct answer varied across letters ($1 = j, n, v$; $2 = a, e, g, l, o, p$; $3 = d, h, m, t$; $6 = i$; $7 = r$). These letters were also used as incorrect alternatives in other trials (e.g., letter *p* could be an incorrect alternative for *nuvem*).

The remaining letters of alphabet that were never a correct answer also were used as incorrect answers as follows: 6 times = b, c, f, s, u; 5 times = x, z. Letters k, q, w, and y were not used as alternatives, because they are uncommon in Brazilian Portuguese, especially as the penultimate letter in words. Finally, the question mark was always one of the four alternatives.

Appendix C: Exploratory Analyses

For completeness, exploratory analyses conducted are reported next, for both Experiments 1 and 2.

Experiment 1

Relationship between performance on main and distracter tasks

It is possible that some participants used subvocal rehearsal or another kind of retrieval practice strategy during distracter task periods. The use of such strategies could possibly lead to lower performance in the distracter task and to produce an enhancement in performance on main task. To check this possibility, we correlated performance on distracter task with performance on the immediately subsequent practice cycle. We hypothesize that if some participants engaged in retrieval practice strategy during distracter task, a negative correlation could be expected between performance in this task and in main task. No significant correlations were observed across cycles, $rs \leq .06$, $ps \geq .69$. This suggests that, in fact, participants engaged in the distracter task, instead of some kind of retrieval practice strategy.

Analysis using covariates

Retention interval differed between participants (range = 42–53 hr). Moreover, the number of fluent languages differed according to each participant's report (range = 1–5). Therefore, we reanalyzed our main dependent variable (i.e., proportion of items correctly recalled on final test) considering this variation. We entered our data in a 2 (type of practice) × 2 (difficulty) repeated-measures analysis of covariance (ANCOVA), including both retention interval and number of fluent languages as covariates. Although the two covariates were nonsignificant, ANCOVA revealed a significant Type of Practice × Number of Fluent Languages interaction, $F(3, 32) = 5.74$, $p = .003$, $\eta_p^2 = .35$. Pearson's correlation indicated that there was a positive relationship between retrieval practice effect size and number of fluent languages, $r = .41$, $p = .003$. After recoding participants as monolinguals (i.e., speakers of the native language only) and multilinguals (i. e., speakers of two or more languages), a *t*-test showed that the retrieval practice effects are greater for multilinguals than for monolinguals, $t(49) = 2.92$, $p = .005$, $d = 0.34$.

Experiment 2

Relationship between performance on main and distracter tasks

We again correlated performance on distracter task with performance on the immediately subsequent practice cycle. No significant correlations were observed across cycles, *r*s ≤ .32, *p*s ≥ .10, suggesting that participants engaged in distracter task, instead of some kind of retrieval practice strategy.

Analysis using covariates

Retention interval differed between participants (range = 46–53 hr). Similarly, the number of fluent languages differed according to each participant's report (range = 1–3). Like in Experiment 1, we reanalyzed our main dependent variable entering our data in a 2 (type of practice) × 2 (difficulty) repeated-measures ANCOVA, with both retention interval and number of fluent languages as covariates. Unlike Experiment 1, the Type of Practice × Number of Fluent Languages interaction was nonsignificant, $F(2, 18) = 0.25$, $p = .79$, $\eta_p^2 = .03$. All other effects were nonsignificant, *F*s ≥ 1.60, *p*s ≤ .21.

https://doi.org/10.1017/SJP.2020.33 Published online by Cambridge University Press