Retrieval practice benefits memory precision


To link to this article: https://doi.org/10.1080/09658211.2019.1623260

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 30 May 2019.

Submit your article to this journal

Article views: 304

View Crossmark data
Retrieval practice benefits memory precision

Brendan A. Schuetze a,*, Luke G. Eglington b and Sean H.K. Kang c

aCognitive Science Program, Dartmouth College, Hanover, NH, USA; bDepartment of Psychological and Brain Sciences, Dartmouth College, Hanover, USA; cDepartment of Education, Dartmouth College, Hanover, NH, USA

ABSTRACT

Although previous research on retrieval practice (RP) has predominantly featured stimuli with discrete right-or-wrong answers, continuous measures offer potentially greater sensitivity in assessing the effects of RP on memory precision. The present study used a colour gradient (125 points ranging from magenta to yellow) as a continuous response variable. The colours of different images were learned through either RP or restudy and either one or three cycles of practice after initial study. On a delayed final test, participants’ memory was assessed for each item’s colour. Participants also created per-item intervals representing the region where they believed the correct colour most likely to have been. We found that repeated rounds of RP enhanced the correspondence between responses and the correct colour. In addition, RP led to participants creating more accurate (correct answers were more likely to be within the participant-specified intervals) and more precise (narrower) intervals relative to restudy, suggesting that RP enhances the precision of memories.

Some researchers have advocated for a shift to an accuracy-oriented approach to the study of human memory, which historically has been dominated by a quantity-oriented approach. The former focuses on how closely the retrieved information corresponds to the actual experienced event, whereas the latter emphasises the amount or number of items that is retrieved (Kornack & Goldsmith, 1994, 1996). Accuracy-oriented researchers argue that describing memory performance primarily in terms of proportion of items recalled may constrain the kinds of research questions that can be answered. In particular, the quantity-oriented approaches often disregard questions concerning the level of detail, or the precision, with which a memory is reported (Goldsmith, Kornack, & Pansky, 2005).

The process by which experience is transformed through metacognitive and cognitive processes into mnemonic reports of varying precision has implications for diverse situations, ranging from educational settings to courtrooms (Evans & Fisher, 2011). For example, consider a court case in which a crime is known to have occurred between 1pm and 1:15pm (relatively high precision) versus merely “the early afternoon” (relatively low precision). As such, the precision of memory reports in eyewitness testimony could be the difference between a conviction or an acquittal.

Despite the importance of memory precision, many questions regarding the influence of encoding strategies on memory precision are yet to be answered. An unresolved question is whether and how the precision of memory representations can be improved. A large body of research shows that when a learner practises retrieving information from memory, the information becomes more accessible and resistant to forgetting (e.g., Carpenter, Pashler, Wixted, & Vul, 2008), leading to recommendations for educators and students to better utilise testing as a learning tool (e.g., Benjamin & Pashler, 2015; Roediger, Agarwal, Kang, & Marsh, 2009). Indeed, retrieval practice (RP) has been found to promote long-term learning and retention more than restudying or rereading the material (Karpicke & Roediger, 2008; Rowland, 2014). Beneficial effects of RP have been found with diverse learning materials, such as lists of words (Götz & Jacoby, 1974), historical facts (Pan, Gopal, & Rickard, 2016), scientific concepts (Karpicke & Blunt, 2011), foreign language learning (Kang, Gollan, & Pashler, 2013), and deductive inference (Eglington & Kang, 2018). Together, this research led us to expect that retrieval practice presents a possible strategy for increasing the precision at which memories are reported.

In this study, we used a metacognitive operational definition of memory report precision first developed by researchers in the 1990s (Yaniv & Foster, 1995). In this...
paradigm, participants report their answers in the form of intervals rather than merely through point estimates (e.g., by reporting that an event occurred “between 1920 and 1940” rather than “in 1937”). The use of interval answers as a dependent measure allows for the interplay between accuracy and informativeness (Goldsmith et al., 2005). As intervals become wider (e.g., “between 1920 and 1940” is wider than “between 1930 and 1940”), they become less informative. Nevertheless, as intervals become wider, the likelihood that they contain the correct answer inevitably also increases. For example, Goldsmith, Koriat, and Weinberg-Eliezer (2002) found that participants use differing grain sizes to strategically regulate memory accuracy over time. That is, as the study-test interval increases, and participants’ memory weakens, they compensate through the use of increasingly wider intervals. This allows for the probability of the interval containing the correct answer to stay constant, even though memory precision may deteriorate over time. As further examined in the discussion section of this paper, this interval width operationalisation differs quite dramatically from the measures of precision estimated by researchers using mixture modelling techniques, both in terms of aims and conceptualisation (e.g., Donkin, Nosofsky, Gold, & Shiffrin, 2015; Sutterer & Awh, 2016).

The majority of retrieval practice studies have used verbal stimuli (e.g., word pairs, text passages), with the dependent measures often being number of items (or idea units) recalled or number of questions answered correctly on the final test. Although these dependent measures may correspond well to educational contexts, the binary nature of these materials (i.e., recalled vs. not recalled; correct vs. wrong) limits the ways in which the data can be analysed. Continuous measures, on the other hand, allow for the analysis of the strength and direction of memory errors, rather than just the gross classification of answers as either correct or incorrect.

In the present study, memory precision was analysed using colour as a continuous variable. Participants learned a total of 80 line-drawn items filled with different colours drawn from a continuous colour gradient. We chose colour due to its continuous nature and prevalence as a dependent measure in previous research on memory fidelity (e.g., Donkin et al., 2015). Nevertheless, it should be noted that we were not investigating the effects of RP on colour memory per se, but rather metacognitive precision during retrieval. As such, we did not attempt to disambiguate colour and spatial location in our task.

We manipulated learning in two ways: via the type of practice (retrieval practice versus restudy) and the number of practice trials (one versus three). A day after the learning phase, participants received a final test for the colour of the items they had learned previously. On each test trial, participants indicated their best estimate of the item colour on a colour gradient bar followed by specifying an interval on the gradient within which they were 90% confident that the correct colour lay. We hypothesised that the colour of items in the retrieval practice condition would be remembered more accurately (with lower error, i.e., closer to the correct colour) and with higher precision (narrower intervals) than restudy counterparts.

Method

Participants

Seventy participants were recruited via Amazon Mechanical Turk (median age was 35 years; 42 female). Our sample size was determined via a power analysis based upon Rowland’s (2014) RP meta-analysis (d = 0.5, power = .85, α = .05). All participants reported normal colour vision. Three outliers, as determined by Hoaglin, Iglewicz, and Tukey’s (1986) interquartile range rule, were dropped from our analyses due to poor performance on the first day restudy task involving simple colour matching. Participants were paid $8.50 upon completion of the two experiment sessions. This study was approved by the Dartmouth Committee for the Protection of Human Subjects.

Stimuli

Eighty clipart-style images were collected from the public domain. Images were chosen to have readily apparent verbal labels (e.g., cat, ice cream cone, football), and each image was presented in a single colour from the LAB colour space (the specific colour for each image was randomly determined with the constraint that no colour was repeated). As opposed to more common colour spaces, such as RGB, the LAB colour space is designed to change in a perceptually constant manner (Regier, Kay, & Khetarpal, 2007). The colour gradient consisted of 125 colours ranging from yellow to magenta. This subset of the LAB colour space was chosen due to the relatively low levels of perceptual bias (Bae, Olkkonen, Allred, & Flombaum, 2015).

Procedure

The experiment was run using jsPsych (de Leeuw, 2015), and participants accessed the experiment using a desktop or laptop computer running the Chrome web browser. The experiment had a $2 \times 2$ within-participants design. Items were learned via two levels of practice, retrieval practice and restudy, and with either one or three practice trials each. After an initial study trial, an item assigned to restudy would be restudied either one or three times (referred to as Restudy-1 or Restudy-3, respectively), and an item assigned to retrieval practice would be tested either one or three times with feedback (referred to as Retrieval-1 or Retrieval-3, respectively). Each participant was presented with twenty images in each of the four conditions, and items were randomly assigned to condition. Each condition was run in a separate block, with a round
of study directly preceding the restudy or retrieval attempts. Random chance determined the order of the practice conditions, and within each practice condition, whether the one- or three-practice trial condition was first.

Participants studied sets of twenty images at a time. The order of the images within these blocks was randomly determined (e.g., in the three-practice trial conditions, the 20 images would cycle through three times in a random order in the assigned practice condition). On average, repeated presentations of a particular item were separated by 20 intervening items (lasting approximately 160 s). Participants were given the opportunity to take short breaks between each block of twenty images. These interstitial breaks tended to be fairly short in length (Median = 11.01 s, Mean = 25.68 s, SD = 35.00 s).

During each study/restudy trial, participants were shown an image filled with a single colour selected from the subset of the LAB colour space mentioned above. This subset of the colour space was shown as a gradient (bar) below the stimulus. Participants used a slider to match the correct position on the colour gradient to the colour of the image being shown to them (see Figure 1). Participants were given 5 s to indicate the correct colour, after which there were 3 s of feedback. Feedback consisted of an arrow depicting the correct location of the colour on the gradient. Retrieval practice trials were formatted exactly the same as the study/restudy trials, except the images were presented in black and participants were asked to retrieve the original colour and indicate it by moving the slider on the colour bar. Again, participants had 5 s to answer before 3 s of feedback were given in the form of the image filled with its correct colour. During this feedback phase, an arrow also appeared, which pointed at the correct location of the colour on the gradient. The learning phase lasted approximately 45 min.

A day after initial learning, participants returned for a follow-up session, in which they were tested on their colour memory for the previous day's images (presented in random order). Each test trial had two stages: First, participants were prompted with a black-and-white version of an image and used the same gradient mechanism (colour bar) from the learning phase to indicate what they believed to be the correct colour for the item; second, participants set the width of an interval centred around the participants' point-estimate of the correct colour. Participants were instructed to create an interval such that they were “90% confident that the correct color lay within the interval”. All test trials were self-paced without feedback (Figure 2).

Between each trial, participants had a ten percent chance of seeing a distractor trial, which they had not seen before. These distractor trials did not differ in any other way from the trials that contained previously seen images. Participants were not informed that there would be distractor trials present within this experiment. As a result, each participant saw approximately 88 total items (80 previously seen items in addition to 8 ± 5 distractor items). The distractor items served to determine how participants would assess their confidence in items that they had not seen before. If participants are able to discriminate distractor items from previously studied items, then their reported interval width for distractor items should be much wider than that for previously studied items (since there should be no colour memory for the new/distractor items).

Results

Unless otherwise specified, linear mixed models were used to analyse data from this experiment. These models were constructed using the R (R Core Team, 2018) package lme4 (Bates, Mächler, Bolker, & Walker, 2015). Mixed effects models were chosen for our analyses due to their ability to account for non-independence among observations by estimating the random effects associated with each item and participant. This approach allows for the calculation of a more reliable estimate for the fixed effects of interest and also increased ability to generalise the findings to other items and participants (cf. Baayen, Davidson, & Bates, 2008; Harrison et al., 2018); the number of practice trials (1 vs. 3) and practice type (RP vs. RS) were treated as fixed effects. The individual participants and the correct location of the target colour were incorporated as random intercepts, as well as random slopes of study type nested within participant. Essentially, this means that the linear mixed model estimated a difficulty level (i.e., intercept) for each correct answer location on the colour bar and also estimated the relative benefit of retrieval practice over restudy for each participant (i.e., a slope estimating the difference between the restudy and retrieval conditions).

Figure 1. Schematic of a study/restudy and retrieval practice trial during the learning phase (Day 1).
first day study session reflects the closing gap between restudy and retrieval practice performance over repeated rounds of practice. The narrowing of this gap was driven primarily by the improvements in the retrieval practice performance; restudy performance stayed relatively constant throughout the first day study session, indicating participants reached some sort of perceptual ceiling on performance.

**Final test phase**

As with the learning phase analysis, absolute error was the dependent measure during the final test phase. A linear mixed model, as described above, found that greater number of presentations \( (b = -5.62, SE = 0.87, d = 0.23, p < .001) \) decreased mean absolute error (MAE), even when accounting for between-participants variability and the location of the correct answer on the colour gradient (see Figure 4). The model also found a significant interaction between practice condition and the number of practice trials, with the benefit of RP over restudy being larger with more practice trials \( (b = -3.11, SE = 1.24, p = .012) \). The main effect of retrieval practice was not significant \( (b = -0.92, SE = 1.02, p = .366) \).

Simulations of random guessing found that chance performance resulted in a MAE of approximately 39 units, while always guessing in the middle of the bar would result in a chance performance of 31 units. This latter, more stringent cutoff is shown in Figure 4. The MAE in arbitrary units for each condition was as such: Restudy-1 \( (M = 31.42, SD = 25.36) \), Retrieval-1 \( (M = 30.43, SD = 24.04) \), Restudy-3 \( (M = 25.78, SD = 24.4) \) and Retrieval-3 \( (M = 21.97, SD = 21.41) \). The maximum possible error was 125 units.

**Report intervals**

**Proportion correct**

In addition, we analysed the impact of retrieval practice on the proportion of correct intervals; intervals were considered correct if the correct answer was located within the given interval. The logistic mixed-effects model incorporating the correct colour and participants as random effects found that items in the three-practice trial condition were significantly more likely to be assigned correct intervals \( (\text{Odds Ratio} = 1.35, SE = .09, p < .001) \). The main effect of RP was not significant \( (OR = .87, SE = .10, p = .147) \); however, there was a significant interaction between practice condition and number of practice trials \( (OR = 1.46, SE = .12, p = .002) \), indicating a significant RP advantage when there were three practice trials. The proportion correct by condition was as follows: Restudy-1 \( (.478) \), Retrieval-1 \( (.439) \), Restudy-3 \( (.507) \), Retrieval-3 \( (.546) \).

**Width**

Under Goldsmith et al.’s (2005) framework, higher memory precision should result in the creation of smaller report intervals. As with our analysis of mean absolute error, we
analysed the effect of the number of presentations and study type on interval width (memory precision) using a linear mixed model. We found that RP ($b = -3.13$, $SE = 1.06$, $d = 0.10$, $p = .004$) and greater number of practice trials ($b = -8.15$, $SE = 0.87$, $d = 0.27$, $p < .001$) significantly decreased the width of participant-generated report intervals (see Figure 5). There was no significant interaction found between the number of practice trials and retrieval practice ($b = -0.44$, $SE = 1.23$, $p = .723$).

When analysing only correct intervals, RP ($b = -4.04$, $SE = 1.72$, $d = 0.12$, $p = .019$) and greater number of practice trials ($b = -13.78$, $SE = 1.43$, $d = 0.41$, $p < .001$) still significantly decreased the width of correct participant-generated report intervals (see Figure 6). In other words, the benefit of RP or additional practice for setting intervals that more often contained the correct answer was not due to participants merely increasing the width of the intervals.

**Distractor items**

As mentioned in the method section, distractor items were randomly interspersed among the previously-presented items, and participants were not told which items they had or had not seen before. The widths of the intervals assigned to previously-presented and distractor items were aggregated within subjects. The mean interval width for previously seen items was 40.12, while the mean width for distractor items was 51.95. A dependent t-test was computed between the interval widths of new (not previously studied) and old (previously studied) items. This test found that previously seen items were assigned significantly smaller intervals [$t(66) = -4.67$, $p < .001$]. When restricting our analysis to the one-practice trial conditions, we still found a significant difference between the intervals created for new and old items [$t(66) = -3.29$, $p = .002$]. These results suggest that participants could distinguish between the previously seen and novel (distractor) items, irrespective of condition.

**Discussion**

In this study, we drew upon previous research on memory precision from the metacognitive perspective (Goldsmith et al., 2005; Yaniv & Foster, 1995) to investigate the effects of retrieval practice. Under this framework, the precision of memory reports is understood as the result of strategic regulation of memory strength through metacognitive processes (Goldsmith et al., 2002). In our study, we extended the use of these metacognitive measures of precision to a colour memory task. Items were learned through varied number of practice trials (one versus three) and practice types (restudy versus retrieval practice). Expanding upon previous research in the study of memory precision (e.g., Donkin et al., 2015; Goldsmith et al., 2002), we measured memory performance through the use of two main dependent variables, mean absolute error (distance between the given and correct colour) and interval width (the precision of participants’ responses).

Although many previous studies have demonstrated the benefit of RP for learning, the present study offers additional novel contributions through the use of a continuous dependent measure and very closely matched RP and restudy conditions. The use of a continuous dependent measure enabled us to calculate fine-grained evaluations of trial-by-trial memory performance, such as mean absolute error and interval width. Additionally, the nature of the restudy and RP trials was closely matched. Both conditions required participants to identify the correct colour using a colour gradient response mechanism. Both conditions were also given feedback on the accuracy of their answers. The requirement that participants actively respond during the restudy condition mollifies potential concerns that the restudy performance may be disproportionately impacted by instances of inattention or divided attention (Buchin & Mulligan, 2017).

We found that, although RP performance was significantly worse than restudy during the first day study session, RP benefitted long term learning in terms of reducing error and increasing the proportion of correct intervals during the day-delayed final test, while simultaneously allowing for the creation of narrower (more precise) intervals. These effects were found despite an active restudy
The benefits of RP were most apparent in the Retrieval-3 condition, in line with previous research showing multiple rounds of retrieval practice leads to an even larger gap in performance between RP and restudy conditions (Rowland, 2014). The reduction of mean absolute error in the Retrieval-3 condition relative to Restudy-3 shows that RP produced more accurate memory that was more similar or closer to the correct colour. The creation of smaller intervals after RP (relative to restudy) indicates that participants had greater confidence in the retrieved colour memory and were able to more precisely report their answers. In other words, retrieval practice led to the elicitation of more informative answers (Ackerman & Goldsmith, 2008). Although the production of more precise and informative intervals may appear to be a rather expected benefit of retrieval practice, we did not regard it as a foregone conclusion since there is evidence that prior testing can in some situations lead to underconfidence (e.g., Finn & Metcalfe, 2007). Alternatively, if the RP intervals were more precise (narrower), and correspondingly less likely to include the correct answer, one could argue that the benefit of RP (in terms of mean absolute error) comes at a cost (overconfidence in the precision of one’s memory).

Instead, we found that items in the restudy condition were more likely to be assigned incorrect intervals (i.e., intervals not containing the correct answer). Incorrect intervals could result from guessing and/or the production of poorly calibrated metacognitive judgements. One might argue that the beneficial effect of RP on the proportion of correct intervals is to be expected, given that the point estimates given by participants for items in the Retrieval-3 condition were already closer on average to the correct colour. Seeing as restudy items were more likely to be assigned incorrect and also wider intervals, one might anticipate that the RP advantage might disappear if we removed incorrect intervals from the analysis.

Counter to this conjecture, we found that the benefit of RP increased when we discarded incorrect intervals (see Figure 6). Although restudy items were assigned wider report intervals, these intervals were not wide enough to compensate for the poorer memories in the restudy condition. As such, our study shows that the RP condition led to metacognitive benefits relative to restudy even in the learning of semantically-sparse stimuli in well matched conditions. Our findings conceptually replicate Barenberg and Dutke’s (2019) finding that RP benefits metacognitive judgements concerning the learning of textual materials. Furthermore, the benefit of RP for memory precision is suggestive of Ackerman and Goldsmith’s (2008) strategic regulation hypothesis. The strategic regulation hypothesis states that memory reports would become less precise in order to compensate for reduced memory strength or specificity (decreased accessibility to specific memory details). Correspondingly, we found that restudy items were remembered more poorly as measured by mean absolute error and were also assigned wider report intervals.

The benefit of RP for arbitrary color-image pairings has implications for gaining theoretical understanding of the RP effect. For instance, the elaborative retrieval hypothesis suggests that the benefit of RP stems from an increased generation of semantic pathways between cue and target (Carpenter, 2009). If you consider a cue and target pair of “house” and “tire” a student using retrieval practice would elaborate on this cue-target pair by activating semantic mediators such as “house” → “garage” → “car” → “tire” which can potentially serve later as additional retrieval routes or cues. It is unclear how the elaborative retrieval hypothesis would explain the learning of color-item pairs, as the generation of one-or-more semantic pathways between “the Parthenon” and “magenta” would be difficult in the limited time between trials in this study. Also, given how restricted the colour range was, there would likely have been massive cue overload (Watkins & Watkins, 1975) that would have limited the usefulness of elaborating on semantic colour labels since numerous items appeared in the same general colour (reddish or yellowish).

Rather, the data from these experiments corresponds more directly with other theoretical accounts of retrieval practice, such as Mozer, Howe, and Pashler’s (2004) neural network-based theory or Karpicke, Lehman, and Aue’s (2014) episodic context account, as neither of these theories require an appeal to semantic information. Rather, both of these theories are agnostic to the type of information – whether visual, semantic, or procedural – subject to enhancement by retrieval practice.

According to Mozer et al. (2004), the retrieval practice benefit can be captured by neural network models that learn by error correction (i.e., a comparison of an actual output to the desired output). They showed that when the system has to make a prediction (attempt retrieval) and then receives corrective feedback, the error signal produced is more reliable and yields better tuning of the network (learning). But when the target response is presented together with the cue (as in the RS condition), it interferes with the estimation of error (how discrepant a
recent output is from the desired output). Although our RS condition did require a response during practice (and did receive feedback), one should bear in mind that the learning goal was to remember the colour of the items (not learn to match the colour of an item that is visually present to a particular colour).

Likewise, Karpicke et al. (2014) episodic context account may also provide theoretical understanding for why a retrieval practice benefit was found in this study. Under Karpicke et al.’s model the retrieval process reinstates the previous context(s) in which an item has been studied and drives contextual updating (the item becomes associated with more diverse contextual features). The association with more elaborate contextual features means that on a later test, the search set can be restricted (fewer candidates) and there is a greater likelihood that a cue will be effective for eliciting the target.

Lastly, we seek to clarify the relationship between this present work and that of other researchers in the domain of memory precision and fidelity. Most pertinently, Sutterer and Awh (2016) investigated the possible impact of retrieval practice on memory precision for colour and concluded null effects. Sutterer and Awh’s results were derived from mixture modelling techniques that were aimed at separating “guess” responses from true retrieval responses. Recent work by Schurgin, Wixted, and Brady (2018) has shown that a simple one-parameter model can account for what had appeared to be dissociations between memory accessibility and memory precision by this mixture modelling approach if the underlying perceptual properties of the stimuli (i.e., colour) space is known. Schurgin et al. conclude that the studies showing such dissociations between precision and strength were the result of Type 1 errors and/or unaccounted for biases in the colour space being used. Additionally, Sutterer and Awh’s (2016) study analysed the impact of only one round of retrieval practice versus one round of restudy without the incorporation of feedback during the learning phase.

In line with Schurgin et al.’s (2018) findings that memory precision cannot be manipulated independently of memory strength, our paper shows a clear correlation between memory strength as measured by mean absolute error and memory precision as operationalised by interval construction. In line with Ackerman and Goldsmith’s (2008) strategic regulation hypothesis, retrieval practice appears to improve the strength of memories, which, in turn leads to subsequent improvements in metacognitive precision (see also Goldsmith et al., 2002).

In sum, the present study assessed the effects of RP on the learning of drawing-color associations that were semantically sparse due to the colour range used. Importantly, the use of colour allowed for a continuous measure of memory. We observed a memorial advantage of RP over restudy in terms of three dependent variables: RP decreased the mean absolute error between given responses and correct answers, decreased the width of report intervals and thus made the reports more informative, and lastly, it increased the proportion of correct intervals, despite the fact that these intervals were narrower. Taken together, these results show that RP is a useful learning strategy that improves not only the correspondence between the given and actual answers, but also the precision and accuracy of metacognitive judgements regarding one’s memorial responses.

Notes

1. Further information on the specifications of this color gradient can be found in the online data repository located at http://doi.org/10.6084/m9.figshare.7527188.
2. When applicable, effect sizes (d) were estimated from these linear mixed models via the method described in Westfall, Kenny, and Judd (2014). Significance test (p) values were computed via Satterthwaite approximation.

Acknowledgement

The research was conducted as part of the first author’s Senior Thesis in the Dartmouth College Cognitive Science Program and was supported by a Kaminsky Undergraduate Research Award. The results were presented in a poster at the International Meeting of the Psychonomic Society, Amsterdam, the Netherlands, in May 2018. We acknowledge Amanda Chen, Autumn Diggs, Danielle Fang, and Nandita Kisireddy for their help with data collection. We also thank Deanne Harper for comments on an early draft of this manuscript.

The data from this study are openly accessible on figshare at http://doi.org/10.6084/m9.figshare.7527188.

Disclosure statement

No potential conflict of interest was reported by the authors.

References


