Learning from failure?

No effect of errorful generation on a cued recall test

Note: 2021 edition (Do not include this note in your report)

ABSTRACT

Potts and Shanks (2014) reported that making mistakes improved the encoding of novel information, compared to simply studying. Following their work, our participants attempted to learn the definitions of very rare English words. During training, participants either guessed the definition (and nearly always made an error) before the correct defintion was shown, or simply studied the words. In contrast to Potts and Shanks (who used a recognition test), we tested memory using cued recall. We employed two different cued recall tasks, differing in difficulty. In the hard task, memory performance was equivalent under guessing and studying conditions, with Bayesian evidence for the null. However, performance was also very low in both conditions (around 15% correct), implying that our result may have been due to a floor effect. In the easier task, performance was higher (around 30% correct), but again no benefit of guessing was found. We hypothesise that errorful generation in Potts and Shanks (2014) may improve memory for definitions, rather than memory for associations.

Optimising the learning of educational materials is of critical importance to educators and students alike. Testing is one technique that has been particularly endorsed in recent years, and it is now well-established that the process of retrieving information from memory in an initial test can improve retrieval in a later test relative to simply restudying information. This effect is known as the testing effect (see Roediger & Karpicke, 2006, for a review).

Students do not always do well on tests though, which might lead educators to worry whether failed tests could do more harm than good. In this context, the results of Potts and Shanks (2014) are particularly interesting. Their first three experiments involved asking participants to learn the definitions of very rare English words (or the English translations of a foreign language). During encoding, the participants were sometimes shown the cue (the rare word, e.g., *valinch*) and were asked to guess the target (the definition), before the correct target (*tube*) was revealed. Other times, the participants simply studied the intact word pairs (e.g., *valinch = tube*) for the entire trial. In a subsequent multiple-choice test, the cues were presented one at a time, and the participants were asked to select the correct target, which was placed among three novel words (“foils”) that were created for each word pair. Since the cues were archaic English words, participants were very unlikely to have known the answers pre-experimentally. In line with this assumption, the participants’ guesses during encoding were almost always incorrect. Nevertheless, in a series of studies, generating errors enhanced performance on the final test compared with studying. Thus, Potts and Shanks (2014) demonstrated a benefit of generating errors over studying with novel word pairs. The implication seems to be that testing is beneficial to learning, even if students nearly always get the answers wrong.

However, one potential issue with Potts and Shanks’ (2014) work is that participants did not actually have to learn the word → meaning associations in order to perform well on the test. This is because the correct answer was always placed among three words that had not previously been seen in the experiment (novel foils). It is therefore possible to do well on the final test just by picking the only definition that seems familiar from the four available. This raises the issue of whether the benefit of errorful generation observed by Potts and Shanks (2014) would still be observed in a more challenging test that requires knowledge of the word→ definition association. In the present experiments, we investigated this question using a cued recall test.

**Experiment**

Experiment 1 was similar to the work reported by Potts and Shanks (2014), except that memory was tested via cued recall, rather than by a multiple-choice test. We ran two versions of our cued-recall task, with different groups of participants. The two conditions differed in their likely difficulty. In the *hard* condition, all items were presented before memory was tested. In the *easy* condition, memory was tested after each within-subject encoding condition (Study, Generate). We ran two conditions in this way in order to reduce the chances that any effects of encoding condition would be obscured by a floor or ceiling effect.

**Method**

*Participants*

We tested 52 participants, drawn from those attending PSYC520 (a second-year psychological research methods module) at Plymouth University. The data from two participants were unusable due to technical errors, leaving 25 participants per condition. This sample was determined in advance of data collection, and has sufficient power to detect medium-to-large effect sizes, both within-subjects (*dz=* 0.58) and betweeen-subjects (*d = 0.81*)

*Apparatus and materials*

The experiment was designed in OpenSesame (Mathôt et al., 2012), and presented online using a JATOS server (Lange et al., 2015). Stimuli were presented in a black 16-point font on a white background, and responses were collected using the keyboard. We used thirty word pairs of rare English words and their definitions, taken from Potts (2013). The full set are included in the Appendix

*Procedure*

Participants were told that they would be studying rare English words and their definitions in two formats. The order of the two encoding conditions (Study, Generate) was counterbalanced between participants. Before each encoding coding, participants read onscreen instructions stating that they would see rare English words and that they should try to remember the correct definitions, because there would be a test. The key instruction for each stage (which told the participants to either study the words or guess the meaning) was presented in red. All other text was presented in black.

Each encoding condition (Study, Generate) consisted of one presentation of each of 15 different word pairs. In the Study condition, each word pair was presented centrally for 17 seconds (e.g. *gadoid = fish)*. The participants simply studied the words. In the Generate condition, the cue was presented alone for 10 seconds (e.g *gadoid*), along with the question “What do you think this word means?”. Participants were strongly encouraged to type a one-word definition, and their answer appeared below the question. The target appeared after 10 seconds, along with the cue, for a further seven seconds.

In the *difficult* condition, the test phase immediately followed the encoding phase. All thirty cues from the encoding phase were presented in random order. The question “What does this word mean?” was presented beneath each cue, and participants were strongly encouraged to type the correct target. The procedure was the same for the *easy* condition, except that a cued- recall test followed each encoding condition (Generate, Study). Each test comprised the 15 cues presented in the immediately preceding encoding condition..

Each trial of the experiment was separated by a 3-second interval.

**Results**

Analysis was conducted using the *BayesFactor* package (Morey & Rouder, 2018) in R (R Core Team, 2019).

*Hard condition:* Mean performance was 13% correct in the Generate condition, and 14% correct in the Study condition. These two conditions did not differ, with Bayesian evidence for the null, *BF =* 0.30. Figure 1 shows the distribution of performances across participants.

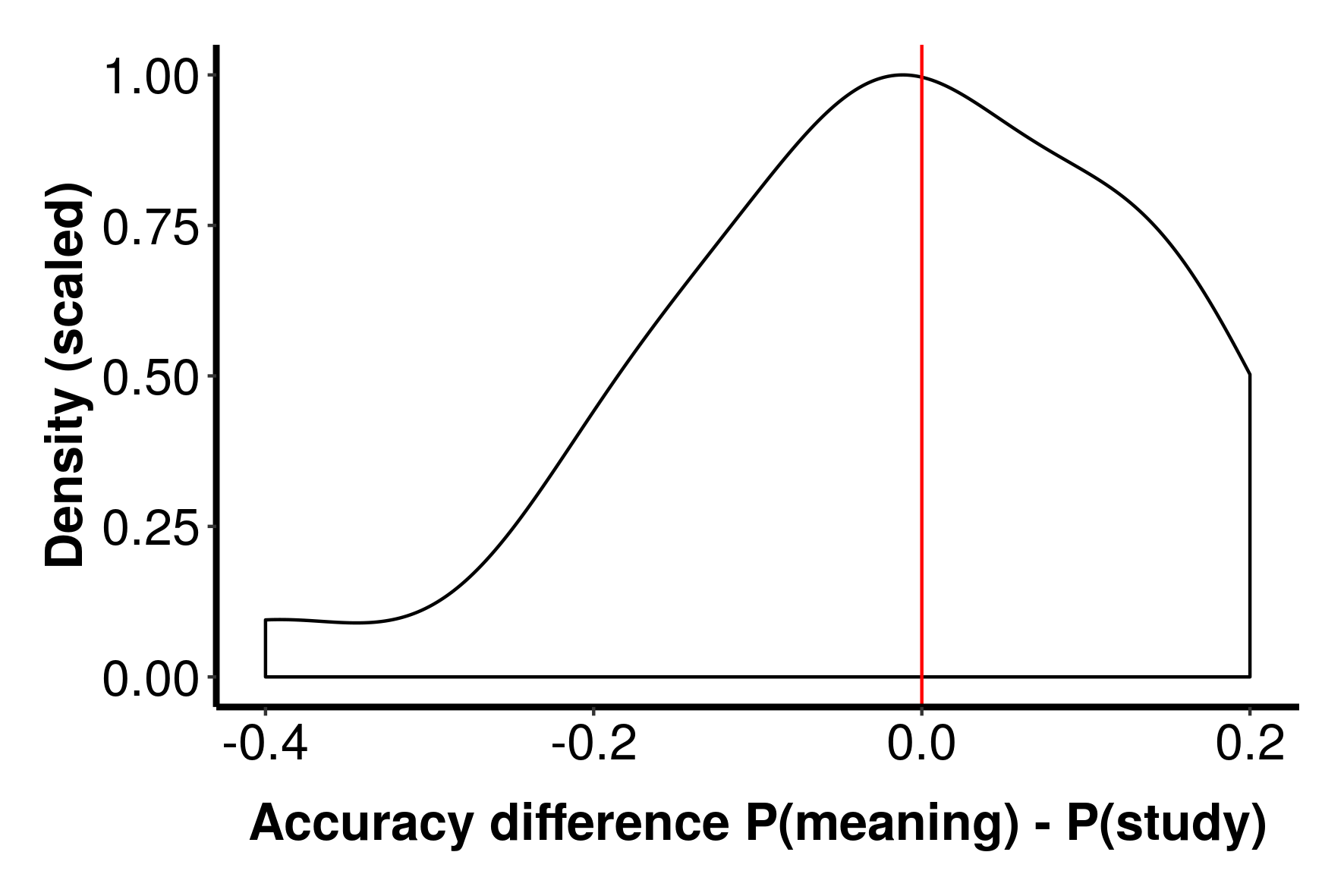


Figure 1. Distribution of difference scores in the *hard* condition.

*Easy condition:* Mean performance was 29% correct in the Generate condition and 35% correct in the Study condition. Bayesian evidence for the presence or absence of a difference was inconclusive, *BF* = 0.86. Figure 2 shows the distribution of performances across participants.

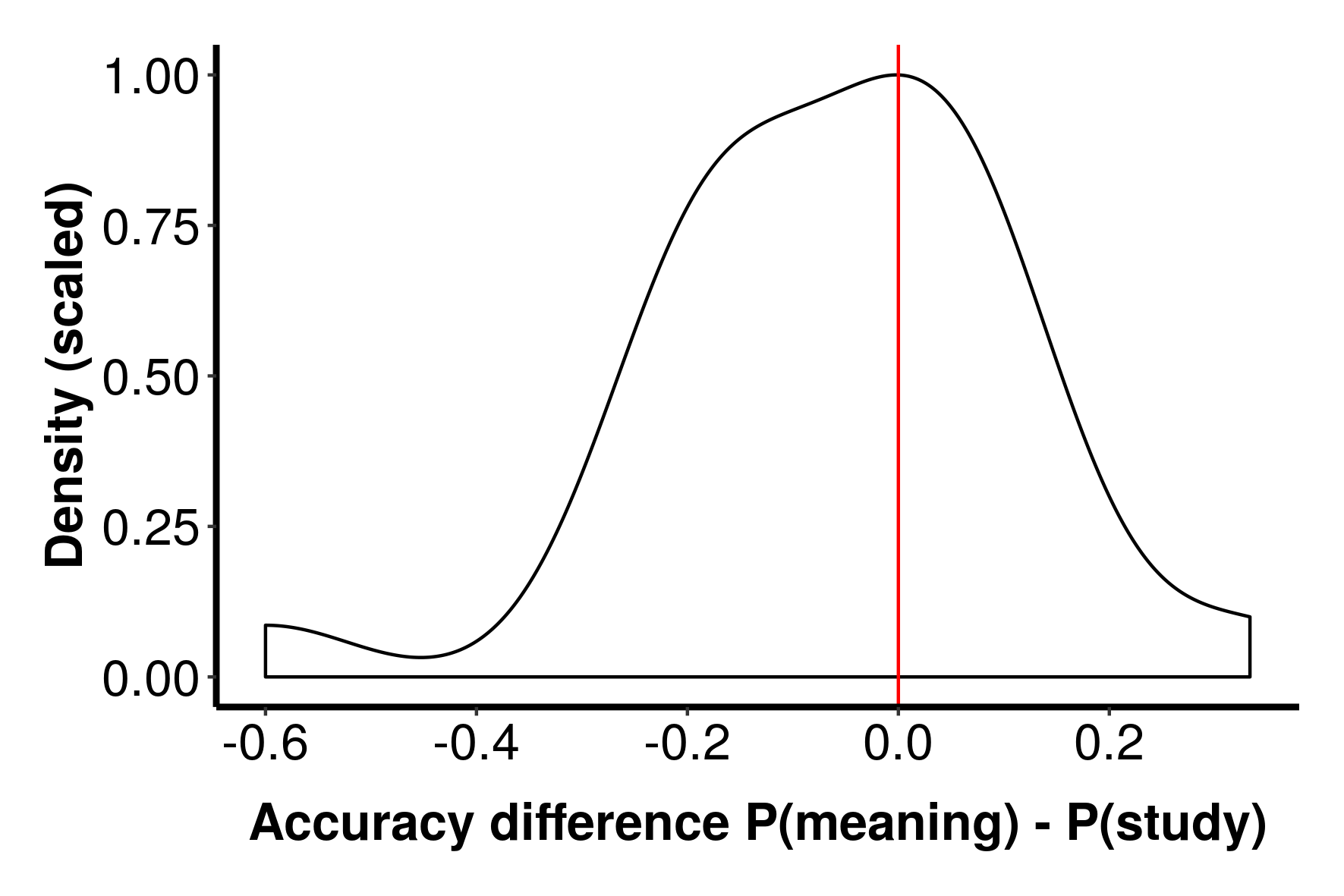


Figure 2. Distribution of difference scores in the *easy* condition.

A factorial Bayesian ANOVA, including Difficulty (Easy, Hard) as a between-subjects factor, and Encoding (Generate vs. Study) as a within-subjects condition, revealed a main effect of Difficulty, BF = 3605. There was weak Bayesian evidence for the absence of an effect of Encoding, BF = 0.39, and for the absence of an interaction between Difficulty and Encoding, BF = 0.38.

**Discussion**

We examined the effect of generating errors versus studying when learning the definitions of rare English words. We used two version of a cued-recall task, one easier than the other, as the test of performance. There was no effect of errorful generation on either version of the task. In the harder task, there was Bayesian evidence for the null but also low performance, meaning the results were possibly due to a floor effect. In the easier task, performance was higher and there was still no effect of errorful generation, but in this case the Bayesian evidence for a difference was inconclusive.

Our results stand in contrast to those of Potts and Shanks (2014), who found benefits of errorful generation across multiple experiments. Perhaps the key difference between our result and theirs is that they used a multiple-choice test as their measure of performance, while we used cue recall. This is potentially important because their multiple-choice test contained the correct definition along with three novel definitions not seen previously in the experiment. It is therefore possible to do well in their experiments using just a sense of familiarity, rather than knowledge of the word → definition association. In contrast, above-chance performance on cued recall requires a memory for the specific associations between the words and their definitions. So, it is possible that errorful generation boosts familiarity with the definitions, rather than boosting the word → definition associations. In most applied situations (e.g. foreign vocabulary learning), it is these associations, rather than overall familiarity, that are critical.

The biggest limitation of the current work is the strength of evidence for an absence of an effect of errorful generation. The results in the harder test could have been due to a floor effect, and the Bayesian evidence in the easier test was inconclusive. In both experiments, performance was numerically lower in the Generate condition than in the Study condition, but the effect size was small-to-medium (dz = 0.34 in the easier test) . Given this estimate of effect size, we would have needed to test at least 55 people in the easier test for adequate statistical power. Running a large-scale replication of the easier test is thus the obvious next step.

Another potential limitation of our work is that participants in the Generate condition see the word pair (e.g. roke – mist) for much less time (7 seconds) than participants in the Study condition (17 seconds). This was done to equate the overall trial length across conditions (17 seconds), and followed the procedure of Potts and Shanks (2014). However, their design must be seen in the context of their result that performance in the Generate condition is better than performance in the Study condition. In this context, their result is all the more surprising given that the condition with the shorter duration of presentation for the word pair is the condition with the better performance. Given our trend for the opposite result (i.e. Study > Generate), the difference in word-pair presentation duration is a potential artifact. In future work, we could run a version of the experiment that equated word-pair presentation time across conditions. This of course would result in overall trial duration becoming unmatched, so there would be some value in running both versions of this experiment.

In conclusion, errorful generation did not improve performance in our cued recall tests. This might imply that errorful generation boosts familiarity with the definitions rather boosting knowledge of the word → definition associations. However, further research is needed.

**Ethical statement**

The stimuli and procedures of these experiments were innocous, and posed no risk to participants or experimenters. Participants gave informed consent, and had the right to withdraw from the study at any point. Data has been stored anonymously.

**References**

Lange, K., Kuhn, S. & Filevich, E. (2015). Just another tool for online studies (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS One, 10,* e0130834.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. Behavior Research Methods, 44(2), 314-324.

Morey, R.D. & Rouder, J.N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-4.2. <https://CRAN.R-project.org/package=BayesFactor>

Potts, R. (2013). *Memory interference and the benefits and costs of testing*. University College London.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. R version 3.6.1. <https://www.R-project.org/>.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

**Appendix**

The words, and defintions, we used were as follows:

