AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2025 American Psychological Association
ISSN: 0096-3445

2025, Vol. 154, No. 3, 658–671
https://doi.org/10.1037/xge0001707

Erring on the Side of Caution: Two Failures to Replicate the Derring Effect

Yeray Mera[1], Ariana Modirrousta-Galian[2, 3], Gemma Thomas[2], Philip A. Higham[2], and Tina Seabrooke[2]

[1] Department of Basic Psychological Processes and their Development, Faculty of Psychology,
University of the Basque Country UPV/EHU
[2] Centre for Perception and Cognition, School of Psychology, University of Southampton
[3] Department of Experimental Psychology, Division of Psychology and Language Sciences, University College London

It has been claimed that deliberately making errors while studying, even when the correct answers are provided, can enhance memory for the correct answers, a phenomenon termed the derring effect. Such deliberate erring has been shown to outperform other learning techniques, including copying and underlining, elaborative studying with concept mapping, and synonym generation. To date, however, the derring effect has only been demonstrated by a single group of researchers and in a single population of participants. This article presents two independent, preregistered replication attempts of the derring effect. In Experiment 1, participants studied 36 term–definition concepts in a within-subjects, laboratory study. On error-correction trials, participants were presented with a term–definition concept and were asked to generate an incorrect definition before correcting it. Error-correction trials were compared with copy trials, where participants simply copied the term–definition concepts and underlined the key concepts. Experiment 2 was an online study in which participants studied trivia facts using a similar protocol. Memory for the studied facts was then tested either immediately (Experiments 1 and 2) or after 2 days (Experiment 1). Unlike the original demonstrations of the derring effect, cued-recall performance did not significantly differ between the error-correction and copy conditions, and the Bayes factors provided moderate support for the null hypothesis in both experiments. We discuss potential explanations for our findings and consider them in relation to key theories and the broader literature on the role of errors in learning.

Public Significance Statement

Four recent articles published in prestigious psychology journals (*Journal of Experimental Psychology: General*, *Journal of Educational Psychology*, and *Educational Psychology Review*) reported that deliberately generating errors when studying, even when the correct answers were provided, substantially boosted subsequent test performance—an effect termed the *derring effect* (Wong, 2023; Wong & Lim, 2022a, 2022b; Yap & Wong, 2024). For the first time, the current article reports two preregistered attempts, which were independent from the original research team, to observe the derring effect. Contrary to expectations, both experiments failed to replicate the derring effect, which strongly questions whether deliberate erring should be incorporated into educational practice.

Keywords: derring effect, error correction, concept learning, deliberate errors, error generation

Supplemental materials: https://doi.org/10.1037/xge0001707.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This article was published Online First January 13, 2025.

Ashleigh Maxcey served as action editor.

Yeray Mera https://orcid.org/0000-0002-8621-9112

Ariana Modirrousta-Galian https://orcid.org/0000-0003-2925-2976

Gemma Thomas https://orcid.org/0009-0006-1164-1729

Philip A. Higham https://orcid.org/0000-0001-6087-7224

Tina Seabrooke https://orcid.org/0000-0002-4119-8389

Yeray Mera and Ariana Modirrousta-Galian contributed equally to this work. All preregistrations, materials, data, analytic code, and additional online materials are available on the Open Science Framework at https://osf.io/j4unw/?view_only=85fb30253d354f4493236461d005cbb5. This work was uploaded as a preprint to PsyArXiv (https://osf.io/preprints/psyarxiv/4u526) and was presented as a poster at the April 2024 Experimental Psychology Society symposium.

The authors have no conflicts of interest to disclose. This study was funded by an Experimental Psychology Society grant awarded to Tina Seabrooke, the Basque Country Government Department of Education (Grant PRE-2020-1-0031) awarded to Yeray Mera, and Economic and Social Research Council South Coast Doctoral Training Partnership studentships awarded to Ariana Modirrousta-Galian and Gemma Thomas. The authors thank Tim Hollins for his thoughts on an earlier draft of the study.

Yeray Mera and Ariana Modirrousta-Galian equally contributed to data curation, formal analysis, investigation, methodology, project administration, software, resources, visualization, writing–original draft, and writing–review and editing. Gemma Thomas contributed to data curation, formal analysis, investigation, software, visualization, and writing–original draft. Philip A. Higham contributed to conceptualization, methodology, supervision, and writing–review and editing. Tina Seabrooke contributed to conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, supervision, validation, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Yeray Mera, Department of Basic Psychological Processes and their Development, Faculty of Psychology, University of the Basque Country UPV/EHU, Tolosa Hiribidea, 70, 20018 Donostia-San Sebastián, Spain. Email: yeray.mera@ehu.eus

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The question of how learners and educators should approach errors in educational settings has been one of long-standing debate. Traditionally, psychologists advocated an *errorless learning* approach, where learners were strongly discouraged from making errors (e.g., Bandura, 1986; Skinner, 1953, 1958). This idea was substantiated by *interference theory* (Melton & Irwin, 1940; Postman & Underwood, 1973), which posits that errors will compete with correct answers during retrieval, thereby impairing recall. More recent research, on the other hand, has shown that errors can sometimes improve subsequent recall (e.g., Kornell et al., 2009; Mera et al., 2022; Metcalfe, 2017), purportedly even when those errors are deliberately generated (Wong, 2023; Wong & Lim, 2022a, 2022b; Yap & Wong, 2024). The present work assesses the conditions in which deliberate errors help, or harm, learning.

Errorful Learning

The *pretesting effect* (also called the *unsuccessful retrieval* and *failed retrieval effect*) provides a good demonstration of how errors can boost memory (Hollins et al., 2023; Kornell et al., 2009; Mera et al., 2022; Pan & Rivers, 2023; Potts & Shanks, 2014; Richland et al., 2009; Seabrooke et al., 2022; Seabrooke, Hollins, et al., 2019; Seabrooke, Mitchell, et al., 2019; Seabrooke, Mitchell, & Hollins, 2021; Seabrooke, Mitchell, Wills, et al., 2021; Tanaka et al., 2019; Zawadzka & Hanczakowski, 2019). In the first demonstration of the pretesting effect, Kornell et al. (2009) had participants study weakly related word pairs (e.g., pond–frog). On pretest trials, participants guessed the target word from a given cue (e.g., pond–?). Because the words in each pair were only weakly related and had not been presented previously, participants' guesses were usually incorrect. These pretest trials were compared with read-only trials, where participants simply studied each word pair, without generating a guess, for a period that matched the pretest trials. In a subsequent cued-recall test, participants recalled significantly more targets from the pretest condition than the read-only condition, even when any targets that were correctly guessed at encoding were removed. Thus, pretesting boosted cued-recall performance, even when participants answered those pretests incorrectly.

Other, related effects also suggest that errors can improve memory. Errors made with high confidence, for example, are more likely to be corrected than those made with low confidence—an effect termed the *hypercorrection effect* (Butterfield & Metcalfe, 2001, 2006; although see Griffiths & Higham, 2018). Similarly, in educational settings, asking learners questions about a topic before they are exposed to it has been shown to improve memory for that information (for an overview, see Carpenter et al., 2023), even though the answers that participants generate are usually wrong. In contrast to the pretesting procedure, participants do not immediately receive corrective feedback but instead study the material after the pretest (for a discussion, see Pan & Carpenter, 2023). This *prequestion effect* has been observed using text passages (Pressley et al., 1990) and video presentations (Carpenter & Toftness, 2017).

Similarly, some studies have focused on the benefits of learning from *incorrect worked examples* (e.g., Große, 2018; Große & Renkl, 2007; Pillai et al., 2020). This body of research indicates that engaging with errors in worked examples can help learners identify misconceptions and refine their problem-solving strategies. For instance, Große and Renkl (2007) explored how finding and fixing errors in worked examples can enhance learning outcomes. Their study suggests that when learners are prompted to analyze incorrect solutions, they develop a better understanding of the underlying concepts and procedures, which can lead to improved performance in problem-solving tasks.

These examples of errorful learning show that making errors can improve memory, which raises the question of whether *deliberately* committing errors would also help learning. Wong and Lim (2022b) set out to answer this question by having participants study unfamiliar term–definition concepts using one of three learning methods. In an error-cancel condition, participants were asked to copy a term–definition concept but deliberately generate a conceptually plausible error before striking through it. For example, when given the term–definition concept *Cocktail party effect is the selective enhancement of attention to filter out distractions*, an appropriate response would be *Cocktail party effect is the selective enhancement of attention ~~making sense of~~ in distractions*. An error-correction condition followed the procedure of the error-cancel condition, except that participants also corrected their error in parentheses, for example, *Cocktail party effect is the selective enhancement of attention ~~in making sense of~~ (to filter out) distractions*. Finally, in a copy condition, participants simply wrote the correct term–definition concept correctly, underlined a key idea within the concept, and rewrote that key idea in parentheses, for example, *Cocktail party effect is the selective enhancement of attention to filter out distractions (selective enhancement of attention)*. In a subsequent cued-recall test of the definitions, both deliberate error conditions outperformed the errorless copy condition, and the error-correction method was especially beneficial. Thus, the authors concluded that deliberately generating conceptually plausible errors during study enhanced subsequent cued recall of the correct answers, an effect termed the *derring effect*.

Subsequent experiments by Wong and Lim (2022b) compared the benefits of derring with other effective study techniques. In two further experiments, the authors found that the error-correction condition also improved subsequent cued recall compared with generating synonyms of the definitions and generating examples of each concept, both of which required participants to elaborate on the definition. Together, these findings suggest that the derring effect is a robust and noteworthy phenomenon. Moreover, the derring effect leads to a clear suggestion that, far from avoiding errors in the classroom, deliberate errors should be encouraged and facilitated.

Three additional sets of studies examined the generalizability of the derring effect. First, Wong and Lim (2022a) had participants study educational texts in one of three conditions. In the error-correction condition, participants copied the text while generating incorrect but conceptually plausible errors for each concept, before striking through the error and correcting it. This condition was compared with a copy condition, which involved copying the text and underlining the key concepts, and a concept-map condition, where participants drew diagrams that linked the concepts together. Participants then completed a free-recall test, where they recalled as many concepts as possible from each educational text, and an application test, which required them to apply the knowledge learned from the educational text to answer questions about a news article on the same topic. Participants from the error-correction condition outperformed those from the copy and concept-map conditions on both tests. Wong and Lim (2023) extended these findings to show that deliberate erring promoted far transfer of knowledge to concepts across different domains. Moreover,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Yap and Wong (2024) recently found that, in the domain of mathematics, deliberate erring improved problem solving and transfer to new and more difficult problems. Together, these results suggest that the derring effect is robust, replicable, and general. In turn, deliberate erring can be regarded as an efficient and effective study technique thus far.

The Present Experiments

Wong and Lim's (2022a, 2022b), Wong's (2023), and Yap and Wong's (2024) research on generating and correcting conceptually plausible errors (derring) present a promising technique to improve learning of both concept–definition terms and information studied in educational passages of text. Moreover, the benefits of derring reportedly appear not only in rote memory tests but also in tests that assess near and far transfer of knowledge and during mathematical problem-solving practice. To our knowledge, however, these findings have not yet been replicated by other researchers. Consequently, in the present work, we report two attempts to replicate and extend the original derring effect reported by Wong and Lim (2022b). Experiment 1 was a laboratory experiment that provided a close replication of the original methodology, while Experiment 2 was an online experiment with a slightly modified procedure. We note that, in both of our experiments, the participants resided in the United Kingdom, while Wong and Lim's (2022b) participants were Singapore students. While we did not anticipate a cultural difference that would affect the magnitude of the derring effect, we note that the samples are not from the same population.

Experiment 1

In Experiment 1, participants studied a series of term–definition concepts, as in Wong and Lim (2022b), in one of two study conditions: error-correction and copy. In the error-correction condition, participants were asked to copy each term–definition concept and generate a conceptually plausible error before striking through it and correcting it. In the copy condition, participants simply copied each term–definition concept, underlined a key idea within it, and rewrote that key idea in parentheses. In Wong and Lim's (2022b) study, the error-correction and copy conditions were the most and least effective study conditions, respectively. Unlike Wong and Lim (2022b), we did not include an error-cancel condition because this condition produced intermediate memory performance and we aimed to replicate the largest effect.[1] Participants completed an immediate cued-recall test for half of the term–definition concepts within each study condition. This test provided the first opportunity to independently replicate the derring effect. Two days later, memory for the remaining term–definition concepts was assessed in a second cued-recall test to assess the longevity of the derring effect.

Following Wong and Lim (2022b), we expected to replicate the derring effect in the immediate test. That is, we predicted that participants would show better cued recall of definitions in the error-correction condition than the copy condition. We did not have strong predictions with respect to the delayed test, but there are effects in which effortful learning techniques are more potent with delayed testing (e.g., the retrieval practice effect; Roediger & Karpicke, 2006). Thus, there was reason to anticipate that the benefits of deliberate erring might be exaggerated with delay.

Method

Transparency and Openness

Both experiments were preregistered. All preregistrations, materials, data, analytic code, and additional online materials for the experiments are publicly available on the Open Science Framework at https://osf.io/j4unw/?view_only=85fb30253d354f4493236461d005cbb5 (Mera et al., 2024). Ethical approval was granted by the University of Southampton Faculty of Environmental and Life Sciences Ethics Committee (No. 72129). We report the rationale for the sample sizes and all data exclusions, manipulations, and measures.

Participants

Forty-six students from the University of Southampton participated in the experiment (36 females, nine males, one who indicated "other," $M = 20.30$ years, $SD = 7.73$). We opted for this sample size to provide sufficient power to replicate the smallest effect comparing error-correction to errorless learning reported by Wong and Lim (2022b; d = 0.43 for a two-tailed pairwise comparison) with 80% power and an α of .05 (G*Power 3.1; Faul et al., 2007). One participant was excluded because English was not their first language. The remaining 45 participants reported English as their first language and received course credit or cash compensation for their participation.

Design

The experiment had a 2 (learning condition: error-correction vs. copy) × 2 (retention interval: immediate vs. delayed test) within-subjects design. The primary dependent variable was the number of correctly recalled definitions on each cued-recall test.

Materials

Forty biology and neuroscience term–definitions were extracted from Wong and Lim's (2022b) study. The concepts were presented in a term–definition "A-is-B" sentence format, with the key term presented in bold font. Four concepts were selected for practice purposes; the remaining 36 concepts were randomly allocated to one of two 18-item lists. The pairing of learning conditions (error-correction or copy method) and item lists, the pairing of retention intervals (immediate or delayed) and item lists, and the order in which participants completed the learning conditions were counterbalanced. Participants completed the experiment in an individual laboratory cubicle.

Procedure

The procedure was similar to Wong and Lim's (2022b) Experiment 1. All participants provided informed consent and were told that they would study scientific term–definition concepts for a later test. They then provided their age, gender, and first language before progressing to the four main phases of the experiment: practice, study, immediate test, and delayed test.

Practice Phase. Participants completed a practice phase to familiarize themselves with the error-correction and copy

[1] In both the current experiment and Wong and Lim's (2022b) experiments, the experimental conditions were blocked, thereby minimizing the potential for trial-by-trial influences between conditions.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

conditions. They were given correct and incorrect example responses for each condition, before practicing each with two different term–definition concepts.

In the error-correction condition, participants were instructed to write each concept such that it contained a conceptual error in its definition (i.e., an error in understanding or interpreting a concept's definition) before striking through and correcting the error. They were encouraged to generate plausible, conceptual errors that were factually incorrect but still believable, for example, ***Proprioception*** *is information about the position and movement of the ~~eyes~~ (body) that is sent to the brain*. In the copy condition, participants were asked to write each concept exactly as it was presented and then identify and underline a key idea in each concept before writing that idea again in parentheses, for example, ***Inattentional blindness*** *is the failure to perceive non-attended stimuli that seem so obvious as to be impossible to miss (failure to perceive non-attended stimuli)*. Upon completion, the experimenter checked that the practice exercises had been completed correctly and discussed any issues with the participants before allowing them to move on to the study phase.

Study Phase. In each learning condition, participants were presented with a printed study list sheet containing 18 term–definition concepts. They were given 2 min to initially read the list before using the error-correction or copy method to study the list for a further 22 min. The learning conditions were blocked, with the order of conditions counterbalanced across participants. Participants were told to continue reviewing the materials if they finished early.

After each learning condition, participants provided (a) a global judgment of learning (JOL) on an 11-point scale ranging from 0% to 100% to predict how much of the material from the study list they would remember later, (b) a rating for how interesting the study list was on a scale ranging from 1 (*not at all*) to 7 (*extremely*), (c) a rating for how understandable the study list was on a scale ranging from 1 (*not at all*) to 7 (*extremely*), and (d) a rating for how well they knew the concepts in the list prior to studying them on a scale ranging from 1 (*not at all*) to 7 (*very well*).

Test Phase. Participants completed two cued-recall tests. Half of the items from each learning condition were tested immediately, while the other half were tested 2 days later for a total of 18 questions per test. In these tests, the terms from the study phase were presented individually on a computer, and participants were asked to type the correct definition in as much detail as they could remember. Response time was not limited, and participants could omit answers. The concepts were tested in blocks by study lists (either the first or the second half of the list in each encoding condition, as per the counterbalancing condition) corresponding to the order in which the lists had been presented during the study phase. The terms within each block were presented in a fixed randomized order. Finally, participants rated how effective each of the two learning methods was on a scale ranging from 1 (*not at all*) to 7 (*extremely*), after both the immediate and delayed tests. All participants were debriefed and thanked for their participation after the delayed test.

Differences Between the Present Study and Wong and Lim's (2022b) Experiment 1

There were some minor differences between the present study and Wong and Lim's (2022b) Experiment 1, which we detail here because this is a replication study. First, memory for half of the items was tested after 2 days, unlike in Wong and Lim (2022b) where memory for all items was tested immediately. While this change may affect the size of the derring effect in the delayed test, we did not anticipate that it would substantially affect the size of the effect in the immediate test. Second, whereas the original study compared three learning conditions (error-cancel, error-correction, and copy), our study compared only two (error-correction and copy, the most and least effective learning conditions in Wong & Lim's, 2022b, study, respectively). Third, our participants studied 18 term–definition concepts with each learning condition compared with 10 in Wong and Lim's (2022b) Experiment 1. We made this change because we had only two conditions and we had both immediate and delayed tests. Finally, participants in our study were given 2 min to initially study each item list, compared with 1 min in the original study, and 22 min to subsequently study the item list in each learning condition, compared with 12 min in the original study. We gave participants this additional time because there were more term–definition concepts to study.

Results

Data Analysis

All analyses were conducted in RStudio (R Core Team, 2023). Bayes factors (BF), which allow us to estimate the empirical evidence in favor of the alternative (H1) and null (H0) hypotheses, were calculated using Version 0.9.12-4.7 of the *BayesFactor* package (Morey & Rouder, 2023). We interpreted the BF according to the evidence categories proposed by Jeffreys (1961) and their corresponding interpretations by Lee and Wagenmakers (2013).

Scoring

Participants' responses were scored as correct if they maintained the meaning of the definition. For each response, a score of no credit (0), partial credit (.5), or full credit (1) was assigned, depending on how well the response demonstrated the essential elements of the concept.[2] There was a maximum possible score of 9 for each learning condition per test, but the total score out of nine was converted to percent correct prior to analyses (i.e., [total score]/9 × 100).

The first two authors individually evaluated 14 of the 45 participants' set of responses. The agreement between them was high, with an intraclass correlation value of .91, 95% CI [.89, .93], determined using a two-way random-effects model. Any inconsistencies between the scorers were examined and discussed until they achieved complete agreement on all responses. Considering the high level of interrater reliability observed, the remaining set of responses were scored by only the first author.

[2] Wong and Lim (2022b) awarded their participants' answers with either full credit (1) or no credit (0), whereas we awarded partial credit (.5) for some answers. We opted for partial credit to better reflect participants' recall accuracy. To check if this affected our results, we reanalyzed the data, replacing the partial credit scores (.5) with both full credit (1) and with no credit (0), and found no statistical differences in the pattern of results. Thus, partial credit scoring was maintained for all analyses.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Preliminary Checks

Table 1 shows participants' mean (and standard deviation) questionnaire ratings and JOLs. Participants had low prior knowledge of the concepts, with no significant differences between learning conditions, $t(44) = 0.88$, $p = .38$, $d = 0.14$, $\mathrm{BF}_{10} = 0.23$. Furthermore, no significant differences were found across learning conditions regarding participants' ratings of the level of interest, $t(44) = 1.16$, $p = .25$, $d = 0.19$, $\mathrm{BF}_{10} = 0.30$, or perceived level of understandability, $t(44) = 0.53$, $p = .60$, $d = 0.08$, $\mathrm{BF}_{10} = 0.18$, of the concepts.

Cued-Recall Performance

Figure 1 shows the mean percent correct, per learning condition, in each cued-recall test. A 2 (learning condition: error-correction, copy) × 2 (retention interval: immediate test, delayed test) within-subject analysis of variance (ANOVA) revealed no significant main effect of learning condition, $F(1, 44) = .01$, $p = .93$, $\eta^2_g < .01$, $\mathrm{BF}_{10} = 0.16$, with participants performing similarly in the error-correction (M = 38.27%, SD = 22.65%) and copy (M = 38.09%, SD = 25.91%) conditions. There was a significant main effect of retention interval, $F(1, 44) = 28.18$, $p < .001$, $\eta^2_g = .06$, $\mathrm{BF}_{10} > 100$, with participants performing better in the immediate test (M = 44.26%, SD = 25.42%) than the delayed test (M = 32.10%, SD = 21.52%). Finally, there was no significant interaction between learning condition and retention interval, $F(1, 44) = 0.08$, $p = .78$, $\eta^2_g < .01$, $\mathrm{BF}_{10} = 0.24$. Thus, Experiment 1 did not replicate the derring effect, and the BF for both the main effect of learning condition and the interaction indicated moderate evidence for the null.

Metacognitive Judgments

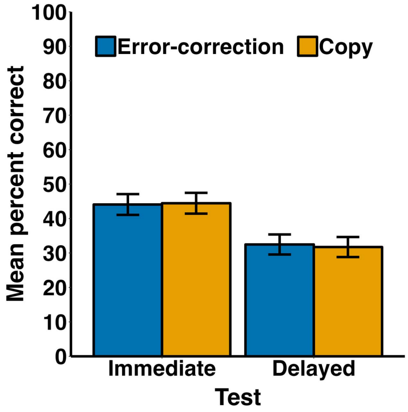
Table 1 shows participants' mean (and standard deviation) JOLs per learning condition and their perceived effectiveness of the learning methods. Participants' JOLs for the two methods did not significantly differ, $t(44) = 1.28$, $p = .21$, $d = 0.17$, $\mathrm{BF}_{10} = 0.35$. Thus, participants' predictions about the effectiveness of the study methods aligned with their actual performance on the cued-recall test, in the sense that cued-recall performance did not differ between the study conditions. Furthermore, a two-way within-subject ANOVA involving the learning condition (copy, error-correction) and retention interval (immediate, delay) factors was conducted to analyze participants' perceived effectiveness of the learning methods. The analysis revealed a significant main effect of learning condition, $F(1, 44) = 4.42$, $p = .04$, $\eta^2_g = .04$, $\mathrm{BF}_{10} = 10.02$, with participants perceiving the copy method (M = 4.01, SD = 1.58) to be more effective than the error-correction method (M = 3.43, SD = 1.36). There was also a significant main effect of retention interval, $F(1, 44) = 21.41$, $p < .001$, $\eta^2_g = .04$, $\mathrm{BF}_{10} = 14.80$, with participants judging the effectiveness of the two methods higher after the immediate test (M = 4.02, SD = 1.51) than the delayed test (M = 3.42, SD = 1.42). No significant interaction was observed between the learning condition and retention interval factors, $F(1, 44) = 0.60$, $p = .44$, $\eta^2_g < .01$, $\mathrm{BF}_{10} = 0.30$.

Table 1
Mean Questionnaire Scores and Metacognitive Judgments in Experiment 1

Variable	Error correction		Copy	
	M	*SD*	*M*	*SD*
Prior knowledge of concepts	2.73	0.99	2.89	1.23
Concept interestingness	4.16	1.45	4.42	1.36
Concept understandability	4.71	1.47	4.82	1.42
JOL (%)	42.22	18.94	45.56	19.60
Method effectiveness				
Immediate	3.69	1.40	4.36	1.57
Delay	3.18	1.28	3.67	1.52

Note. N = 45. JOL = judgment of learning.

Figure 1
Mean Percent Correct in Each Cued-Recall Test of Experiment 1



Note. Error bars represent difference-adjusted, within-subject 95% confidence intervals (Baguley, 2012). See the online article for the color version of this figure.

Error Type Analysis

Following Wong and Lim (2022b), we investigated the types of errors that participants made in the immediate and delayed tests for each learning condition. Four participants were excluded from this analysis because they did not make any errors in at least one condition of one test. The remaining 41 participants' errors were split into four categories: (a) commission errors (inadequate or incorrect responses that were different from participants' deliberate errors in the study phase), (b) omission errors (no response at test), (c) confusions (responses that provided the definition of a different concept term), and (d) intrusions (responses that repeated the same deliberate errors that participants made during the study phase, which are only possible in the error-correction condition). Table 2 shows the mean proportion and standard deviation of each error type.

Most incorrect responses on both the immediate and delayed tests were omission errors. Furthermore, only 1% of incorrect responses were intrusion errors. This result suggests that committing errors deliberately during learning did not cause significant interference. Importantly, the rate of intrusion errors was not significantly different between the immediate and delayed tests, $t(40) = 0.22$, $p = .82$, $d = 0.05$, $\mathrm{BF}_{10} = 0.17$.

We conducted three two-way within-subjects ANOVAs to examine the effects of learning condition (error-correction, copy) and retention interval (immediate, delay) on the proportion of all other

Table 2

The Distribution of Errors at Test Across the Commission, Omission, Confusion, and Intrusion Categories in Experiment 1

	Immediate test				Delayed test			
	Error correction		Copy		Error correction		Copy	
Error category	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Commission	0.24	0.24	0.27	0.22	0.34	0.19	0.29	0.29
Omission	0.65	0.33	0.61	0.29	0.60	0.24	0.68	0.28
Confusion	0.10	0.21	0.12	0.22	0.05	0.14	0.04	0.07
Intrusion	0.01	0.04			0.01	0.04		

Note. Intrusion errors were not applicable to the errorless copy condition. $N = 41$.

error types (commissions, omissions, and confusions). There was no significant main effect of learning condition on the proportion of commission errors, $F(1, 40) = 0.13$, $p = .72$, $\eta_g^2 < .01$, $\text{BF}_{10} = 0.18$, with participants producing a similar proportion of commission errors in the error-correction ($M = 0.29$, $SD = 0.22$) and copy ($M = 0.28$, $SD = 0.26$) conditions. There was, however, a significant main effect of retention interval, $F(1, 40) = 4.72$, $p = .04$, $\eta_g^2 = .02$, $\text{BF}_{10} = 1.14$. Specifically, participants made more commission errors in the delayed test ($M = 0.32$, $SD = 0.24$) than in the immediate test ($M = 0.26$, $SD = 0.23$). There was no significant interaction between these variables, $F(1, 40) = 2.22$, $p = .14$, $\eta_g^2 = .01$, $\text{BF}_{10} = 0.54$.

Regarding omission errors, there were no significant main effects of learning condition, $F(1, 40) = 0.31$, $p = .58$, $\eta_g^2 < .01$, $\text{BF}_{10} = 0.19$, or retention interval, $F(1, 40) = 0.10$, $p = .75$, $\eta_g^2 < .01$, $\text{BF}_{10} = 0.17$. Participants generated a similar proportion of omission errors in the error-correction ($M = 0.62$, $SD = 0.28$) and copy ($M = 0.64$, $SD = 0.29$) conditions and in the immediate ($M = 0.63$, $SD = 0.31$) and delayed ($M = 0.64$, $SD = 0.26$) tests. Finally, no significant interaction between the variables was observed, $F(1, 40) = 3.89$, $p = .06$, $\eta_g^2 = .01$, $\text{BF}_{10} = 1.14$.

Regarding confusion errors, there was no significant main effect of learning condition, $F(1, 40) = 0.04$, $p = .85$, $\eta_g^2 < .01$, $\text{BF}_{10} = 0.17$, with participants generating a similar proportion of confusion errors in the error-correction ($M = 0.07$, $SD = 0.18$) and copy ($M = 0.08$, $SD = 0.17$) conditions. There was, however, a significant main effect of retention interval, $F(1, 40) = 10.35$, $p = .003$, $\eta_g^2 = .04$, $\text{BF}_{10} = 19.31$. Participants made more confusion errors in the immediate test ($M = 0.11$, $SD = 0.21$) than in the delayed test ($M = 0.04$, $SD = 0.11$). Finally, there was no significant interaction between the variables, $F(1, 40) = 1.06$, $p = .31$, $\eta_g^2 < .01$, $\text{BF}_{10} = 0.32$.

Discussion

The present study aimed to replicate and extend the derring effect first reported by Wong and Lim (2022b). Both studies compared the effectiveness of deliberately generating and correcting errors when studying term–definition concepts to simply copying and underlining such concepts. Wong and Lim (2022b) originally found that the error-correction condition led to better subsequent performance on an immediate cued-recall test than the copy condition. We, by contrast, observed no benefit of error-correction over copying in either an immediate or delayed cued-recall test.

We consider our null memory results to be meaningful for several reasons. First, the BF revealed moderate evidence for the null hypothesis for both the main effect of learning condition and the interaction between learning condition and retention interval. Second, as expected, participants performed better in the immediate test than the delayed test, which is consistent with normal memory deterioration over time (Ebbinghaus, 1964). This result suggests that our experiment had appropriate methodology to replicate other established memory effects in the literature. Finally, cued-recall performance was not obviously subject to either ceiling or floor effects. Hence, there was ample opportunity to detect a derring effect, and other aspects of memory (decay over time) followed a typical pattern. Yet, derring did not confer a memory advantage, even in the immediate test, where Wong and Lim (2022a, 2022b), Wong (2023), and Yap and Wong (2024) consistently observed such benefits.

Our experiment had power equal to 80% to replicate the smallest effect comparing error-correction to errorless learning reported by Wong and Lim (2022b; $d = 0.43$), which is the conventional power level used in modern psychological research. Nevertheless, some critics might argue that this level of power is insufficient for a replication study. However, if our experiment was underpowered, the BF would have likely been anecdotal rather than moderate, with the former indicating equal or almost-equal odds for both the null and the alternative hypotheses (Rosenfeld & Olson, 2021; Schönbrodt & Wagenmakers, 2018). Consequently, we do not believe that low power is a legitimate concern.

Aside from memory performance, several other aspects of the data warrant elaboration. Notably, errors in the cued-recall tests were most likely to be omission errors (i.e., where participants provided no answer). This pattern contrasts with Wong and Lim's (2022b) study, where participants' incorrect test responses were mostly commission errors. These findings could indicate a more cautious and conservative response pattern among our sample. One possibility, therefore, is that our participants were more reluctant to respond during the test unless they were confident about the correct answer, thereby limiting their ability to gain credit for partial knowledge.

Given these different results, it would be useful to compare the quality of participants' responses, both at encoding and retrieval, in our study and in Wong and Lim's (2022b) experiments. However, Wong and Lim's (2022b) data are not publicly available nor were they available upon request. Thus, we can only speculate as to the potential differences in participants' responses.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Finally, participants' JOLs—their predictions of how much material they would remember—also did not significantly differ between the error-correction and copy conditions. These predictions aligned with participants' objective recall performance. These findings differ from Experiment 1 of Wong and Lim (2022b), where participants incorrectly predicted that their performance would be better in the copy condition than in the error-correction condition. However, the results are more consistent with Experiments 2 and 3 of Wong and Lim (2022b), where participants predicted no significant differences in performance between the learning conditions employed in those experiments. Furthermore, when participants judged the effectiveness of the study methods after each test, they judged the copy method as the more effective learning method, whereas in Wong and Lim's (2022b) study, both methods were perceived as equally effective. This preference for the copy method in our study might be due to its straightforward nature, where participants simply copied the term–definition concepts and underlined and rewrote the key ideas within them, potentially leading to a perception of better comprehension and retention of the material. In general, learning conditions that allow for fluent processing are often perceived by learners as being more effective than conditions that induce "desirable difficulties" (Bjork & Bjork, 2011). This pattern has also been observed in related work, including studies on the pretesting effect (Huelser & Metcalfe, 2012; Yang et al., 2017), the retrieval practice effect (Kornell & Son, 2009), and fluent versus disfluent lecture styles (Carpenter et al., 2013). However, conditions that present challenges during learning often boost long-term retention (Bjork & Bjork, 2011). Our results are thus in line with the often-seen metacognitive illusion that fluent study conditions feel, but are not objectively, superior for learning.

Experiment 2

There are several potential reasons why we failed to replicate the derring effect in Experiment 1, but we reserve elaborating on these reasons until the General Discussion section. Before drawing strong conclusions from a single null result, we felt it prudent to conduct an additional replication attempt to further explore the boundary conditions of the derring effect. In Experiment 2, we therefore once again pitted the error-correction and copy conditions against each other, this time in an online experiment. We note that there were some procedural differences (noted below) between this study and Wong and Lim's (2022b) experiments. As such, this study should be treated as a conceptual, rather than direct, replication.

We suspected that the biology and neuroscience materials were quite challenging for our psychology student participants in Experiment 1. This suspicion was supported by the low prior knowledge that participants reported for the materials. One likely consequence of having low prior knowledge of the materials is that participants may have struggled to generate conceptually plausible errors. Because generating such errors is the key component of the error-correction condition, it might explain why the error-correction condition did not boost cued-recall performance in Experiment 1.[3] In Experiment 2, we therefore changed the materials to trivia facts that we selected so that it would be straightforward for participants to generate plausible errors, even if they did not have prior knowledge of the trivia fact. For example, given the fact *The* ***sunfish*** *can produce more eggs than any other known vertebrate*, participants should have easily been able to generate another vertebrate.

We also drew inspiration from studies on the pretesting effect, where having participants generate erroneous guesses during learning (when the correct answers are not provided to them) produces robust improvements in cued recall for related word pairs (e.g., Grimaldi & Karpicke, 2012; Kornell et al., 2009; Seabrooke, Mitchell, & Hollins, 2021) and trivia questions (Kornell, 2014) relative to a study-only control condition (see Mera et al., 2022, for a review). In these experiments, word pairs and trivia questions are usually presented on separate, experimenter-paced trials to ensure that exposure to each item is comparable. We adopted this approach in Experiment 2, such that the trivia questions in each learning condition were presented on a trial-by-trial basis, meaning that each trivia question was presented individually, one after the other, rather than presenting them all together in a combined list.

Thus, participants initially studied a series of trivia questions, with a keyword (representing the answer to the question) presented in bold font. Each trivia fact was also presented a second time, but with the keyword missing. On error-correction trials, participants were encouraged to generate a plausible error for the keyword before copying the correct keyword. On copy trials, participants simply copied the correct keyword twice. The key question was whether the error-correction condition would lead to better performance than the copy condition on a subsequent cued-recall test.

Method

Participants

Forty-eight participants were recruited via Prolific (https://www.prolific.com) and paid £6 per hour. Two participants were excluded for failing to follow experimental instructions.[4] The remaining sample ($N = 46$) had sufficient power (81%) to detect the smallest reported effect size in Wong and Lim's (2022b; $d = 0.43$) study. The sample consisted of 28 females and 18 males, who were aged between 19 and 60 years ($M = 38.93$ years, $SD = 12.68$ years). All participants spoke English as their first language and were located in the United Kingdom.

Design

A within-subjects design was used to compare cued-recall performance between the error-correction and copy learning conditions. The primary dependent variable was the number of correctly recalled answers in the cued-recall test.

Materials

The experiment was programmed using jsPsych (Version 6.3.1; de Leeuw, 2015) and hosted on a JATOS (https://www.jatos.org)

[3] To test this possibility, we conducted a multiple linear regression with learning condition, retention interval, and prior knowledge as predictors and cued-recall performance as the dependent variable. The model explained approximately 14% of the variance, $F(7, 172) = 4.01$, $p < .001$. Only the main effect of retention interval ($b = 12.75$, $p = .01$) was significant (see Supplemental Table S1).

[4] The excluded participants provided no response to over 50% of test trials against the instruction to provide a "best guess." The results of the experiment were comparable regardless of whether these participants were excluded or not.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

server (Lange et al., 2015). Device restrictions were applied on Prolific, which suggested that participants access the experiment through a computer. Ninety-six trivia questions and their associated answers were selected from Fastrich et al.'s (2018) study. Each question was reworded to be a trivia fact consisting of one short sentence with the answer included in bold (e.g., *Joseph Priestley discovered* ***oxygen*** *in 1774*). Six trivia facts were allocated to the practice task. The remaining facts were randomly allocated to the two learning conditions such that there were 45 facts in each condition for each participant. The learning conditions were blocked, and their order was counterbalanced between participants.

Procedure

Participants completed an encoding phase, distractor task, and test. Before the experiment, participants were screened through Prolific to ensure they were between 18 and 60 years old and English was their native language. Participants first read an information sheet, provided informed consent, and confirmed their age and English fluency. They were also asked to confirm that they would not use additional memory aids to boost their performance. This "cheat check" was repeated at the end of the experiment. No participants admitted using additional memory aids. In addition, participants answered three simple attention checks during the experiment, which aimed to exclude participants who were not paying attention to the task (i.e., failing two or more of three attention checks). Here, participants viewed a 4 × 4 grid, with each cell containing a different letter. Their task was to identify the letter in red, while all the others were presented in black. No participant was excluded based on this criterion.

Encoding Phase. Participants completed the two learning conditions in a blocked, counterbalanced fashion. Each condition began with instructions on how to use the specified learning method, followed by a short practice opportunity. In the copy condition, a trivia fact was presented at the top of the screen with the answer written in bold (e.g., *The* ***sunfish*** *can produce more eggs than any other known vertebrate*). The trivia fact was also presented beneath, but with the answer missing (e.g., *The ____ can produce more eggs than any other known vertebrate*). Participants had 12 s to copy the correct answer (e.g., *sunfish*) into an onscreen textbox. The textbox was then cleared and participants had 6 s to copy it again. In the error-correction method, trivia facts were presented as in the copy condition, but participants had 12 s to first generate a plausible error (e.g., carp) before correcting themselves by copying the correct answer (e.g., sunfish) for 6 s. For each condition, participants completed three practice trials followed by 45 main trials. The trials were presented in a fresh random order for each participant, and they were separated by 500 ms intervals. After each block, participants provided a global JOL on an 11-point scale ranging from 0% to 100%, predicting what proportion of the facts they would remember at test. Participants then rated their familiarity with the facts on a 5-point scale, ranging from 1 (*not at all familiar*) to 5 (*extremely familiar*).

Distractor Task. The distractor task lasted approximately 2.5 min and involved a continuous performance task occurring in two blocks. In each block, 128 colorful triangles were presented, each shown for 450 ms, pointing up, down, left, or right. In the first block, participants were required to press the spacebar whenever the triangle pointed upward. In the second block, participants had to press the spacebar whenever the triangle was both red and pointing upward. Each block began with 32 practice trials, each shown for 550 ms.

Test Phase. Participants completed a cued-recall test. Each trivia fact was presented individually, in a random order, with the answer missing. Participants were asked to type the missing word into an onscreen textbox in a self-paced manner. Participants could omit answers, although they were instructed to provide a best guess if they could not recall the correct answer. No feedback was provided during this phase. The main test was preceded by a practice test with the six trivia facts from the practice encoding phase. Upon completion of the cued-recall test, participants were asked to rate how effective they found each of the learning methods for learning the trivia facts on a scale from 1 (*not at all*) to 7 (*extremely*).

Results

Scoring

Participants' test responses were classified as correct (1) or incorrect (0) according to whether they matched the correct associated answer. Minor spelling errors, for which the meaning was clear (e.g., "dalmtian" instead of "dalmatian") were scored as correct.[5]

Prior Knowledge

Table 3 shows the mean and standard deviation of participants' prior knowledge ratings and metacognitive judgments. Prior knowledge of the trivia facts was low. Surprisingly, participants reported significantly more prior knowledge of the trivia facts studied using the copy method than the error-correction method, $t(44) = 2.12$, $p = .04$, $d = 0.24$, although the BF evidence for this difference was anecdotal, $\mathrm{BF}_{10} = 1.23$. We did not anticipate this difference (because the facts were randomly allocated to the conditions across participants), but we offer a potential explanation for it in the Discussion section.

Cued-Recall Performance

Table 3 shows the mean percentage of correctly recalled targets (and corresponding standard deviations) for each learning condition in the cued-recall test. A paired-samples t test revealed that cued-recall performance did not significantly differ between the learning conditions, $t(45) = 0.60$, $p = .55$, $d = 0.06$, and the BF moderately supported the null, $\mathrm{BF}_{10} = 0.19$.

Metacognitive Judgments

Concerning participants' JOLs (see Table 3), participants predicted that they would recall more trivia facts from the copy condition than the error-correction condition, $t(44) = 2.63$, $p = .01$, $d = 0.41$, $\mathrm{BF}_{10} = 3.42$. Similarly, after the cued-recall test, participants evaluated the copy method as more effective than the error-correction method, $t(45) = 3.53$, $p = .001$, $d = 0.72$, $\mathrm{BF}_{10} = 30.08$. Together, these results

[5] Unlike Experiment 1, the correct answer in Experiment 2 was a single word. Participants therefore had less opportunity to partially answer the questions, making the partial credit procedure from Experiment 1 unnecessary.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3

Mean Prior Knowledge of Facts, Cued-Recall Performance, and Metacognitive Judgments in Experiment 2

Variable	Error correction		Copy	
	M	*SD*	*M*	*SD*
Prior knowledge of facts	1.91	0.76	2.15	0.99
Cued-recall performance (%)	70.99	18.62	72.08	18.72
JOL (%)	50.67	22.40	59.57	17.63
Perceived method effectiveness	3.96	1.46	4.91	1.19

Note. N = 46 for "perceived method effectiveness" and "cued-recall performance." N = 45 for "prior knowledge of facts" and "JOL" due to one participant not providing a response for these measures. JOL = judgment of learning.

demonstrate a discrepancy between participants' judgments and their memory performance.

Error Type Analysis

As in Experiment 1, participants' cued-recall errors were categorized into commission, omission, confusion, and intrusion errors. Table 4 shows the mean proportion (and standard deviation) of each error type. Most incorrect responses were commission errors, with participants making significantly more commission errors for facts studied using the copy method than the error-correction method, $t(45) = 3.76$, $p < .001$, $d = 0.61$, $\mathrm{BF}_{10} = 56.34$. There was no significant difference between conditions in the proportion of omission, $t(45) = 0.95$, $p = .35$, $d = 0.07$, $\mathrm{BF}_{10} = 0.24$, or confusion, $t(45) = 1.54$, $p = .13$, $d = 0.32$, $\mathrm{BF}_{10} = 0.48$, errors. Intrusion errors accounted for 12% of incorrect responses for items studied in the error-correction condition.

Cross-Experimental Analysis

Experiments 1 and 2 produced no evidence of a derring effect, and indeed the BF provided moderate support for the null in both cases. It is worth noting, however, that the sample size for both experiments was chosen to provide 80% power. While this is the conventional level of power for psychological research, it does allow for 20% possibility of Type II error. To more broadly assess the degree of evidence for the derring effect (across minor methodological changes) with greater power, we reanalyzed the combined cued-recall data from Experiments 1 and 2. Only the immediate test was included from Experiment 1, and the same exclusion criteria were applied as in the original analyses, leaving 45 participants from Experiment 1 and 46 participants from Experiment 2.

Table 4

The Distribution of Errors at Test Across the Commission, Omission, Confusion, and Intrusion Categories in Experiment 2

Error category	Error correction		Copy	
	M	*SD*	*M*	*SD*
Commission	0.55	0.24	0.71	0.29
Omission	0.11	0.19	0.12	0.23
Confusion	0.22	0.17	0.17	0.17
Intrusion	0.12	0.13		

Note. Intrusion errors were not applicable to the errorless copy condition. N = 46.

The mean percentage of correctly recalled targets per condition in Experiments 1 and 2 is shown in Figure 1 and Table 3, respectively. A 2 (group: Experiment 1 vs. Experiment 2) × 2 (condition: error-correction vs. copy) mixed ANOVA on the percentage of correctly recalled items revealed a significant main effect of group, $F(1, 89) = 39.33$, $p < .001$, $\eta_g^2 = .28$, $\mathrm{BF}_{10} > 100$. Average cued-recall performance was poorer in Experiment 1 (M = 44.26%, SD = 25.42%) than Experiment 2 (M = 71.53%, SD = 18.58%). More importantly, there was no significant main effect of condition, with the error-correction (M = 57.68%, SD = 25.30%) and copy (M = 58.41%, SD = 26.90%) conditions producing comparable cued-recall performance, $F(1, 89) = 0.18$, $p = .68$, $\eta_g^2 < .01$. Furthermore, the BF provided moderate support for the null hypothesis even with the greater power that the combined participant sample provided, $\mathrm{BF}_{10} = 0.17$. Finally, there was no significant interaction between the group and condition factors, $F(1, 89) = 0.04$, $p = .84$, $\eta_g^2 < .01$, with the BF again providing moderate support for the null hypothesis, $\mathrm{BF}_{10} = 0.24$.

Discussion

As in Experiment 1, the error-correction condition did not produce better subsequent cued-recall performance than the copy condition. This null result replicated a similar null result in Experiment 1 using different materials and minor procedural variations. Interestingly, participants judged the copy condition to be more effective than the error-correction condition. This belief was reflected both in their JOLs and in their perceived effectiveness ratings. The latter judgment mirrors the pattern seen in Experiment 1, further suggesting that participants preferred the more fluent copy condition.

Participants also reported greater familiarity with the trivia facts in the copy condition than in the error-correction condition. This was an unexpected result because the trivia facts were randomly allocated to the encoding conditions for each participant. Nevertheless, because the familiarity ratings took place *after* each encoding condition, we suspect that the result may also reflect participants' sense of increased fluency in the copy condition.

Finally, some differences were observed in participants' errors in the cued-recall test compared with Experiment 1. Participants made mostly omission errors in Experiment 1 but mostly commission errors in Experiment 2 (as in Wong & Lim's, 2022b, study). We suspect that this difference may first reflect the use of trivia questions in Experiment 2, which provided more context to make a reasonable guess (e.g., *The ____ can produce more eggs than any other known vertebrate* clearly indicates that the answer is a vertebrate). We also suspect that the cued-recall instructions, which emphasized guessing where necessary in Experiment 2 only, may also have encouraged more overt guesses. Participants also made significantly more commission errors in the copy condition than the error-correction condition in Experiment 2. This pattern may have arisen because participants could not make intrusion errors in the copy condition, meaning that their errors from the copy condition were classified into three rather than four categories. Intrusion errors were relatively rare in both experiments, suggesting that participants' guesses did not interfere much with the retrieval of the correct answers in either experiment.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

General Discussion

In two experiments, we failed to observe the derring effect. In each experiment, participants studied biological term–definition (Experiment 1) or trivia (Experiment 2) facts, either by correctly copying the facts twice or by copying the information with a conceptually plausible error before correcting that error. In a subsequent cued-recall test that took place either immediately (Experiments 1 and 2) or after 2 days (Experiment 1), cued recall did not significantly differ between the conditions.

There are several possible reasons why we might have failed to replicate the derring effect. One notable difference between our study and those conducted by Wong and Lim (2022a, 2022b), Wong (2023), as well as Yap and Wong (2024), is the sample characteristics; they recruited university students from Singapore, while we recruited university students from England.[6] Educational practices in Asian communities often encourage students to initially attempt to solve problems on their own—an error-based learning approach—instead of starting with teacher-directed explanations that are more common in Western communities (Metcalfe, 2017; Schleppenbach et al., 2007; Stevenson & Stigler, 1994; Stigler & Hiebert, 1999). Teachers from Asian cultures may also react differently to teachers from Western cultures when students *do* make errors. Schleppenbach et al. (2007), for instance, compared teachers from Chinese and U.S. schools. While both sets of teachers tended to ask follow-up questions when students made errors, Chinese teachers were significantly more likely to ask students follow-up questions after an error than U.S. teachers. Moreover, Chinese teachers were also more likely than U.S. teachers to openly make students feel comfortable about making errors. It is therefore possible that Wong and Lim's (2022b) participants were more familiar with error-based learning than our British participants, thereby leading them to approach the task more enthusiastically and thereby generate "better" kinds of errors. In other words, the preconceptions that participants bring to a study on error-based learning may influence the effectiveness of the technique. In this regard, it would be very useful to compare the types of deliberate errors that our participants made (which are publicly archived) with the errors that Wong and Lim's (2022b) participants made. Unfortunately, this is not possible at present because Wong and Lim's (2022b) data are not publicly archived and were not available upon request.

While our methodology, particularly in Experiment 1, was mostly comparable with that of Wong and Lim (2022b), there were some procedural differences. One difference is that we compared only the most effective error-correction condition with the least effective copy condition and omitted the intermediate error-cancel condition. It is, in principle, possible that the omission of the error-cancel condition is responsible for our failure to observe the derring effect, but we see no a priori reason why this would be the case, and we therefore consider it an unlikely explanation. In both Wong and Lim's (2022b) experiments and ours, the encoding conditions were blocked, and the order of these conditions was counterbalanced across participants. If the derring effect results from carryover effects between the error-cancel condition on the one hand and the error-correction and copy condition on the other, we would expect that influence to be much more pronounced when the conditions are presented in a randomly intermixed order. Moreover, Wong and Lim (2022b) also observed a derring effect in one experiment that did not include an error-cancel condition, which suggests that the error-cancel condition does not play a critical role in obtaining the effect.

A third possibility is that our findings reflect a Type II error (false negative). We of course cannot rule out this possibility. However, we note that we have two preregistered experiments in which the Bayesian analyses moderately supported the null. Experiment 1 was a close laboratory replication of Wong and Lim's (2022b) Experiment 1, which we expected to provide maximal opportunity to obtain the derring effect. The laboratory nature of the experiment, coupled with participants completing the encoding conditions with pen and paper (which then must be hand-scored), made for a time-consuming and resource-dependent experiment. To our knowledge, we are the only group of researchers who have sought to independently replicate the derring effect. We therefore think that it is important to document our struggles to prevent futile attempts to obtain the derring effect in other laboratories.

A fourth possibility is that the original effects reported by Wong and Lim (2022b) reflect a Type I error (false positive). This possibility is noteworthy for several reasons which, when taken together, raise significant concerns. First, as noted earlier, Wong and Lim's (2022b) data are not publicly available nor were they available upon request; the data of Wong and Lim's (2022a) and Wong's (2023) studies are also not publicly available, and only the aggregated (but not raw) data are publicly available for Yap and Wong (2024). Second, Wong and Lim (2022b) excluded 10% of their participants on the vague and non-preregistered grounds that they "failed to conform to the experimental instructions" (p. 28), providing no further details on the matter. Third, the original derring effect that compared the error-correction method with the copy method in Wong and Lim's (2022b) Experiment 1 yielded a surprisingly large effect size of $d = 1.30$, which is considerably larger than some of the most robust effects in the literature. Indeed, it is 2.6 times the meta-analytic effect size of the testing effect ($g = 0.50$; Yang et al., 2021).[7] Overall, despite Wong and Lim's (2022b) experiments being reasonably powered, their lack of data transparency, ambiguous and unplanned exclusion criteria, and unusually large effect size make it pertinent to consider the possibility that the original effects are false positives until further evidence comes to light.

Theoretical Implications

Although we did not observe a derring effect in our work, our findings may have theoretical implications for related phenomena in which errors later affect memory. Wong and Lim (2022b) offered three theoretical accounts of the derring effect. One suggestion derived from an ironic rebound effect of thought suppression (see Wegner et al., 1987). The idea here is that, when participants attempt to generate a plausible error, they suppress the correct answer, which then rebounds and becomes particularly available in memory. Wong and Lim (2022b) argued against this account because their error-correction condition produced better subsequent cued-recall performance than the error-cancel condition, where participants generated plausible errors but were not required to correct them. The

[6] We did not record how many international students participated in the study because we did not anticipate that the derring effect would be moderated by nationality.

[7] Note that Cohen's d and Hedge's g are largely comparable, especially with medium and large sample sizes (Goulet-Pelletier & Cousineau, 2018).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

authors argued that correcting the error would satisfy participants' need to express it, thereby reducing the rebound effect and the availability of the answer in memory. Thus, they argued that this "rebound" account predicts that correcting deliberate errors should be less beneficial than *not* correcting them, which is opposite to their observed pattern.

The second theoretical account posited by Wong and Lim (2022b) can broadly be categorized as a "semantic elaboration" account that is akin to the mediator effectiveness hypothesis of the testing effect (e.g., Carpenter, 2011; Pyc & Rawson, 2010) and semantic accounts of the pretesting effect (e.g., Grimaldi & Karpicke, 2012; Zawadzka & Hanczakowski, 2019). The idea here is that generating a plausible error requires greater elaboration of the correct answer than in the errorless copy condition. In turn, greater elaboration should produce additional retrieval routes, or mediators, to boost recall. Wong and Lim (2022b) also argued against this theory for two reasons. First, they argued that correcting a deliberate error at encoding should not affect the capacity of that error to later serve as a mediator, and yet the error-correction condition boosted cued recall relative to the error-cancel condition.[8] Second, deliberate erring produced better subsequent cued-recall performance than generating synonyms or examples of the term–definition concepts. Generating synonyms and examples both require participants to elaborate on the answer and would be expected to foster a semantic network that could aid subsequent cued-recall performance. The fact that deliberate erring improved cued-recall performance, beyond generating synonyms and examples, suggests that the derring effect reflects more than just semantic elaboration.

Our failure to observe the derring effect clearly provides no positive evidence for these theories and instead suggests that these mechanisms are unlikely to be in operation in the error-correction condition. Wong and Lim's (2022b) favored theory was that derring encourages participants to pay more attention to correct answers, thereby improving episodic recollection of the encoding event (Tulving, 1985). However, we see some problems with this account. First, it is difficult to explain why the error-correction condition would improve cued recall in Wong and Lim's (2022b) experiments, but not in our experiments. Second, and more generally, it is possible that any attentional facilitation requires psychological factors that the error-correction condition does not afford. One prominent theory of the *pretesting* effect, for example, suggests that generating errors boosts attention to the correct answers because the act of guessing induces a sense of curiosity (Potts et al., 2019; Seabrooke, Mitchell, et al., 2019) about the answer. As Wong and Lim (2022b) noted, it is difficult to understand how deliberate erring would induce curiosity because participants already know the correct answer when they generate an error. Likewise, the *hypercorrection* effect has also been attributed to enhanced attentional processing that comes when participants are surprised to receive feedback after generating a high-confidence error (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009). Again, it is difficult to see how participants could be surprised in the derring paradigm because they know the correct answer when they generate an error.

Thus, it is possible that some element of curiosity or surprise is necessary for errors to be beneficial. This suggestion speaks not only to the derring effect but to the question of whether generating errors *in general* is likely to be beneficial. Asking learners to generate guesses, or erroneous alternative answers, may only be beneficial when there is an information gap (Potts et al., 2019) that can be filled with subsequent corrective feedback. As Wong and Lim (2022b) noted, however, it is not clear why their participants would have been curious or surprised to learn the answers given that the answers were already present in their error-correction and error-cancel conditions.

Pedagogical Recommendations

Given the struggles that we have documented in obtaining the derring effect, we cannot recommend deliberate erring as an effective study technique. Instead, we encourage students and educators to adopt other better established study strategies, such as retrieval practice, to exploit the backward (Roediger & Butler, 2011; Rowland, 2014) and forward (Yang et al., 2018) testing effects and successive relearning to obtain the benefits of spaced retrieval practice (Higham et al., 2022, 2023). Deliberate erring appeared to be highly effective for Wong and Lim's (2022a, 2022b), Wong's (2023), and Yap and Wong's (2024) participants, who were students from Singapore, and so there may be cultural differences at play. At present, this is an empirical question that awaits further research.

This is not to say that we encourage students to adopt a wholly *errorless* approach to learning. For example, studies on the prequestioning effect have shown that asking participants questions before they watch an educational video can improve their learning of the material (Carpenter et al., 2023). In such studies, participants are asked to answer questions about a topic before they have studied it, meaning that their answers are almost inevitably wrong, and yet such guessing still improves subsequent cued-recall performance. Similar findings have also been shown when participants guess the answers to questions before studying educational passages of text (Richland et al., 2009) and when learning trivia facts (Kornell, 2014). Moreover, the act of attempting to retrieve previously studied information from memory in general is beneficial for learning, regardless of whether the retrieval attempt is ultimately successful (Kornell et al., 2015). Similar findings have also been observed in studies on the generation effect, where participants were tasked with completing sentences that had open endings (e.g., *The executive went to shop for a new* ___; Kane & Anderson, 1978). Thus, we recommend that students regularly attempt to retrieve information from memory, even when uncertain, because successful retrieval is beneficial and unsuccessful retrieval is not usually harmful (although see Seabrooke, Hollins, et al., 2019, Experiment 5 for an exception). Regular retrieval attempts that are followed by corrective feedback, rather than deliberate error generation when the correct answer is already provided, remains to be the more effective study strategy.

Conclusion

We report two preregistered experiments in which we failed to observe the derring effect first reported by Wong and Lim (2022b).

[8] One possibility, however, is that the error is inhibited by correcting it, thereby reducing its capacity to act as a mediator. This possibility would be in line with research demonstrating apparent inhibition of memory representations that are not the direct focus of the current task. For example, representations of ignored stimuli are potentially inhibited in selective attention tasks (e.g., negative priming; Tipper, 1985), and representations of words related to retrieved items are potentially inhibited during retrieval (e.g., retrieval-induced forgetting; Anderson et al., 1994).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

We observed no benefit of generating and correcting errors relative to copying terms without errors, in either an immediate or a delayed cued-recall test. Our findings suggest that committing and correcting deliberate errors during learning does not reliably enhance cued-recall performance. Further research is needed to thoroughly evaluate the effects of generating errors—both deliberate and accidental—on learning considering diverse conditions and participant populations.

Constraints on Generality

Following Simons et al. (2017), we consider the generality of our findings in a constraints on generality statement. Our results contrast with those reported by Wong and Lim (2022a, 2022b), Wong (2023), and Yap and Wong (2024). Before commencing this line of research, we expected the derring effect to be a robust phenomenon that would generalize beyond Wong and Lim's (2022a, 2022b), Wong's (2023), and Yap and Wong's (2024) participants. This appears not to be the case. Their participants were all Singapore undergraduates, while our participants were located in the United Kingdom (undergraduates in Experiment 1 and Prolific participants in Experiment 2). It is possible that the derring effect is a replicable effect but that it is limited to certain samples such as Singapore students. If this is the case, it would be worth testing whether the effect also extends to age groups that are different from the university undergraduates that Wong and Lim (2022a, 2022b), Wong (2023), and Yap and Wong (2024) tested (e.g., younger children and older adults). Specifying the target populations, and the boundary conditions, of the derring effect will be critical if deliberate erring is to be used in real educational environments.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158–175. https://doi.org/10.3758/s13428-011-0123-7

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494. https://doi.org/10.1037/0278-7393.27.6.1491

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69–84. https://doi.org/10.1007/s11409-006-6894-z

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K., King-Shepard, Q., & Nokes-Malach, T. (2023). The prequestion effect: Why it is useful to ask students questions before they learn. In C. Overson, C. Hakala, L. Kordonowy, & V. Benassi (Eds.), *In their own words: What scholars and teachers want you to know about why and how to apply the science of learning in your academic setting* (pp. 74–82). Society for the Teaching of Psychology.

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350–1356. https://doi.org/10.3758/s13423-013-0442-z

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). Dover. (Original work published 1885).

Fastrich, G. M., Kerr, T., Castel, A. D., & Murayama, K. (2018). The role of interest in memory for trivia questions: An investigation with a large-scale database. *Motivation Science*, *4*(3), 227–250. https://doi.org/10.1037/mot0000087

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

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88–92. https://doi.org/10.3758/PBR.16.1.88

Goulet-Pelletier, J. C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's *d* family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Griffiths, L., & Higham, P. A. (2018). Beyond hypercorrection: Remembering corrective feedback for low-confidence errors. *Memory*, *26*(2), 201–218. https://doi.org/10.1080/09658211.2017.1344249

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Große, C. S. (2018). "Copying allowed—But be careful, errors included!"—Effects of copying correct and incorrect solutions on learning outcomes. *Learning and Instruction*, *58*, 173–181. https://doi.org/10.1016/j.learninstruc.2018.06.004

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Higham, P. A., Fastrich, G. M., Potts, R., Murayama, K., Pickering, J. S., & Hadwin, J. A. (2023). Spaced retrieval practice: Can restudying trump retrieval? *Educational Psychology Review*, *35*(4), Article 98. https://doi.org/10.1007/s10648-023-09809-2

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

Hollins, T. J., Seabrooke, T., Inkster, A., Wills, A., & Mitchell, C. J. (2023). Pre-testing effects are target-specific and are not driven by a generalised state of curiosity. *Memory*, *31*(2), 282–296. https://doi.org/10.1080/09658211.2022.2153141

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*(4), 514–527. https://doi.org/10.3758/s13421-011-0167-z

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology*, *70*(4), 626–635. https://doi.org/10.1037/0022-0663.70.4.626

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 106–114. https://doi.org/10.1037/a0033699

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915

Lange, K., Kühn, 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*(6), Article e0130834. https://doi.org/10.1371/journal.pone.0130834

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modelling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Melton, A. W., & Irwin, J. M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *The American Journal of Psychology*, *53*(2), 173–203. https://doi.org/10.2307/1417415

Mera, Y., Modirrousta-Galian, A., Thomas, G., Higham, P. A., & Seabrooke, T. (2024, March 21). *Erring on the side of caution: Two failures to replicate the derring effect*. Open Science Framework. https://osf.io/j4unw

Mera, Y., Rodríguez, G., & Marin-Garcia, E. (2022). Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories. *Psychonomic Bulletin & Review*, *29*(3), 753–765. https://doi.org/10.3758/s13423-021-02022-8

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Morey, R. D., & Rouder, J. N. (2023). *BayesFactor: Computation of Bayes factors for common designs* (R package Version 0.9.12-4.7) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and pretesting effects: A review of empirical research, theoretical perspectives, and implications for educational practice. *Educational Psychology Review*, *35*(4), Article 97. https://doi.org/10.1007/s10648-023-09814-5

Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition*, *51*(6), 1461–1480. https://doi.org/10.3758/s13421-022-01392-1

Pillai, R. M., Loehr, A. M., Yeo, D. J., Hong, M. K., & Fazio, L. K. (2020). Are there costs to using incorrect worked examples in mathematics education? *Journal of Applied Research in Memory and Cognition*, *9*(4), 519–531. https://doi.org/10.1016/j.jarmac.2020.06.007

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, *1*(1), 19–40. https://doi.org/10.3758/BF03198064

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1023–1041. https://doi.org/10.1037/xlm0000637

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology*, *15*(1), 27–35. https://doi.org/10.1016/0361-476X(90)90003-J

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), Article 335. https://doi.org/10.1126/science.1191465

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rosenfeld, J. P., & Olson, J. M. (2021). Bayesian data analysis: A fresh approach to power issues and null hypothesis interpretation. *Applied Psychophysiology and Biofeedback*, *46*(2), 135–140. https://doi.org/10.1007/s10484-020-09502-y

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

Schleppenbach, M., Flevares, L. M., Sims, L. M., & Perry, M. (2007). Teachers' responses to student mistakes in Chinese and U.S. mathematics classrooms. *The Elementary School Journal*, *108*(2), 131–147. https://doi.org/10.1086/525551

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Seabrooke, T., Hollins, T. J., Kent, C., Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, *104*, 70–82. https://doi.org/10.1016/j.jml.2018.10.001

Seabrooke, T., Mitchell, C. J., & Hollins, T. J. (2021). Pretesting boosts item but not source memory. *Memory*, *29*(9), 1245–1253. https://doi.org/10.1080/09658211.2021.1977328

Seabrooke, T., Mitchell, C. J., Wills, A. J., & Hollins, T. J. (2021). Pretesting boosts recognition, but not cued recall, of targets from unrelated word pairs. *Psychonomic Bulletin & Review*, *28*(1), 268–273. https://doi.org/10.3758/s13423-020-01810-y

Seabrooke, T., Mitchell, C. J., Wills, A. J., Inkster, A. B., & Hollins, T. J. (2022). The benefits of impossible tests: Assessing the role of error-correction in the pretesting effect. *Memory & Cognition*, *50*(2), 296–311. https://doi.org/10.3758/s13421-021-01218-6

Seabrooke, T., Mitchell, C. J., Wills, A. J., Waters, J. L., & Hollins, T. J. (2019). Selective effects of errorful generation on recognition memory: The role of motivation and surprise. *Memory*, *27*(9), 1250–1262. https://doi.org/10.1080/09658211.2019.1647247

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Skinner, B. F. (1953). *Science and human behavior*. Simon & Schuster.

Skinner, B. F. (1958). Teaching machines; from the experimental study of learning come devices which arrange optimal conditions for self instruction. *Science*, *128*(3330), 969–977. https://doi.org/10.1126/science.128.3330.969

Stevenson, H. W., & Stigler, J. W. (1994). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. Simon & Schuster.

Stigler, J. W., & Hiebert, J. (1999). Understanding and improving classroom mathematics instruction: An overview of the TIMSS video study. In Australian Council for Educational Research (Ed.), *Raising Australian* *standards in mathematics and science: Insights from TIMSS (1997 conference proceedings)* (pp. 52–65). Australian Council for Educational Research. https://research.acer.edu.au/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1000&context=research_conference_1997#page=59

Tanaka, S., Miyatani, M., & Iwaki, N. (2019). Response format, not semantic activation, influences the failed retrieval effect. *Frontiers in Psychology*, *10*, Article 599. https://doi.org/10.3389/fpsyg.2019.00599

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology*, *37*(4), 571–590. https://doi.org/10.1080/14640748508400920

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Wegner, D. M., Schneider, D. J., Carter, S. R., III, & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*(1), 5–13. https://doi.org/10.1037/0022-3514.53.1.5

Wong, S. S. H. (2023). Deliberate erring improves far transfer of learning more than errorless elaboration and spotting and correcting others' errors. *Educational Psychology Review*, *35*(1), Article 16. https://doi.org/10.1007/s10648-023-09739-z

Wong, S. S. H., & Lim, S. W. H. (2022a). Deliberate errors promote meaningful learning. *Journal of Educational Psychology*, *114*(8), 1817–1831. https://doi.org/10.1037/edu0000720

Wong, S. S. H., & Lim, S. W. H. (2022b). The derring effect: Deliberate errors enhance learning. *Journal of Experimental Psychology: General*, *151*(1), 25–40. https://doi.org/10.1037/xge0001072

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

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1073–1092. https://doi.org/10.1037/xlm0000363

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning*, *3*(1), Article 8. https://doi.org/10.1038/s41539-018-0024-y

Yap, J. B. K., & Wong, S. S. H. (2024). Deliberately making and correcting errors in mathematical problem-solving practice improves procedural transfer to more complex problems. *Journal of Educational Psychology*, *116*(7), 1112–1128. https://doi.org/10.1037/edu0000850

Zawadzka, K., & Hanczakowski, M. (2019). Two routes to memory benefits of guessing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1748–1760. https://doi.org/10.1037/xlm0000676

Received March 20, 2024
Revision received August 27, 2024
Accepted October 23, 2024 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.