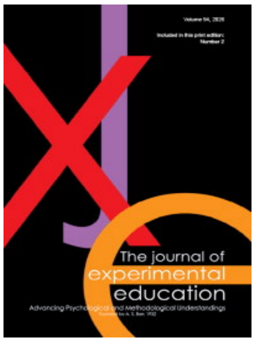

The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Generative Processing Strengthens Learning Across Errorful and Non-Errorful Contexts

Thet Hnin Khin Khin & Steven C. Pan

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2026.2636103

Routledge
Taylor & Francis Group

Check for updates

LEARNING, INSTRUCTION, AND COGNITION

Generative Processing Strengthens Learning Across Errorful and Non-Errorful Contexts

Thet Hnin Khin Khin[a,b] and Steven C. Pan[a]

[a]Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Singapore City, Singapore; [b]Harvard Graduate School of Education, Harvard University, Cambridge, MA, USA

ABSTRACT

Errorful generation—making (often) incorrect guesses about unfamiliar information—can enhance learning if accompanied by corrective feedback. Four experiments investigated how feedback processing influences the retention of obscure word definitions. Participants processed feedback via *non-generative* (reading, copying), *elaborative* (studying examples or mnemonics), or *generative* (rephrasing, example generation) methods. In Experiments 1, 2, and 4, rephrasing feedback consistently yielded better learning than reading or copying. In Experiment 3, rephrasing and example generation outperformed keyword mnemonics. Experiment 4 revealed that rephrasing improved learning even without errorful generation. Evidence for studying provided examples was mixed. Overall, these findings demonstrate that generative processing facilitates learning, and for errorful generation, the impact of feedback is determined not just by its presence, but how it is engaged with.

KEYWORDS
Corrective feedback; errorful generation; feedback processing; generative learning; learning from errors; pretesting

Introduction

In the learning sciences, *errorful generation*—wherein learners attempt to answer questions about previously unlearned material, often resulting in incorrect guesses—has attracted growing interest for its potential to enhance learning, although it is still relatively uncommon in real-world educational settings (Pan et al., 2020). Importantly, the memorial benefit of errorful generation, which is also known as the *errorful generation benefit* or the *pretesting effect*, is contingent on learners having the opportunity to engage with correct answer feedback following an error. For example, a student might be asked to guess the meaning of the unfamiliar word "guerdon", respond incorrectly, and then receive the correct answer ("to provide with a reward or compensation") as feedback. Provided that the student has the opportunity to learn from that feedback—which is usually accomplished by reading it—then they are likely to remember the meaning of the word more successfully than if they had simply studied the word and its definition from the outset. Benefits of errorful generation with corrective feedback have been demonstrated relative to non-errorful methods (e.g., reading or studying without any prior guessing) with such stimuli as word-definition pairs (Potts et al., 2019), facts (Kornell et al., 2009), and prose (Richland et al., 2009).

Various theoretical accounts for this phenomenon exist (for reviews, see Metcalfe, 2017; Kornell & Vaughn, 2016; Mera et al., 2022; Pan & Carpenter, 2023). The semantic mediator hypothesis proposes that guessing activates related mental concepts, one of which may serve as mediator linking the cue and target (Potts et al., 2019); once feedback is provided, the correct

CONTACT Steven C. Pan scp@nus.edu.sg Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, 9 Arts Link, Singapore City 117572, Singapore

answer becomes associated with this mediator, facilitating later retrieval. Search set theory suggests that presenting a cue triggers a memory search, activating candidate answers, with feedback strengthening the association between the correct answer and the relevant cue, facilitating later retrieval (Grimaldi & Karpicke, 2012). The prediction error account posits that when learners' responses are flagged as erroneous, they become aware of knowledge gaps, creating a disequilibrium that induces greater attention and motivation to learn from corrective feedback (Knight et al., 2012; cf. Seabrooke et al., 2022), yielding deeper encoding and increased retention relative to passive studying. Finally, the memory-for-change account (related to recursive reminding theory, Hintzman, 2011) proposes that the act of errorful generation and learning the correct answer can form a distinct episodic memory (Jacoby & Wahlheim, 2013; Wahlheim & Jacoby, 2013), which enhances later recall of the correct response.

Most research on errorful generation has focused on the effects of making incorrect guesses, leaving a notable gap regarding how learners process and internalize corrective feedback (with a notable exception being the timing of such feedback, which appears to be most potent when it is immediate; Hays et al., 2013). In fact, in most studies of errorful generation, participants are expected to study feedback without any detailed instructions or procedures for how to engage with it (Metcalfe, 2017). Yet feedback processing, which is crucial for the benefits of errorful generation, can take on a multitude of forms beyond reading correct answers (e.g., Park & Ahn, 2022). In classroom and other real-world learning contexts, exactly what form(s) of feedback processing should occur following errorful generation remains unclear. To address that gap, the present study investigated the impacts of different feedback processing strategies—including *generative* versus *non-generative* approaches—on learning through errorful generation.

Feedback processing strategies and learning

Several prominent learning theories and frameworks imply that processing corrective feedback in a generative manner—that is, where the learner must produce something meaningful and separate from information provided to them (Brod, 2021)—could be beneficial for learning. In fact, generative learning theories posit that producing new content, as opposed to passively receiving information, can improve comprehension, understanding, and other learning outcomes (Fiorella & Mayer, 2016). If so, then having learners not just read feedback but generate something new in response to that feedback—such as rephrasing feedback in one's own words—may be beneficial. Similarly, the interactive–constructive–active–passive framework (Chi & Wylie, 2014) suggests that when learners engage in constructive (i.e., creating new content) or interactive (i.e., integrating new with old knowledge) activities, learning can benefit. Active learning activities, which require manipulating content in some fashion, are thought to be less beneficial but still more effective than passive learning activities. The well-established *generation effect* (Slamecka & Graf, 1978; see also Bertsch et al., 2007; Crutcher & Healy, 1989), in which memory for verbal materials improves when participants produce rather than merely read them (e.g., completing the word fragment "ap_le" rather than reading the intact word "apple"), further underscores the learning benefits of generating content over passive reception.

Although the effects of generative versus non-generative feedback processing on errorful generation have yet to be investigated, related findings offer relevant insights. For instance, Park and Ahn (2022) examined reading, copying, and explaining (describing the rationale behind feedback from the learner's understanding) as feedback processing strategies after students received written corrective feedback on their second language writing errors. Copying and explaining yielded increased feedback implementation during subsequent writing activities compared to the reading condition. Although the students' mistakes were natural occurrences rather than the product of deliberate errorful generation, this finding supports the notion that different feedback processing strategies and at least one type of generative strategy (explaining) can yield different levels of cognitive engagement, affecting learning. Similarly, Siegler (1995) found that encouraging young children to explain the reasoning behind corrective feedback received, beyond merely reading it, enhanced their learning of number conservation. The generative process of providing justification for the correct answer benefited learning even when it conflicted with the child's initial response, possibly through greater depth in cognitive processing (Craik & Lockhart, 1972) and elaborative thinking (Levin, 1988). Further, Lahtinen et al. (1997) demonstrated that generative strategies such as summarization and concept mapping outperformed verbatim notetaking (i.e., copying) and underlining, which themselves surpassed passive reading. Although this study examined strategies employed autonomously by students to assimilate lecture content rather than feedback processing methods, it provides empirical support for the advantages of generative processing of knowledge in educational contexts.

In some cases, feedback processing does not necessarily need to be generative to enhance learning. The copying condition in Park and Ahn (2022) is one such example. In another case, Pan, Hutter, et al. (2018) had participants engage in retrieval practice with feedback on biology concepts, and on each trial, compare corrective feedback with their own response and copy that feedback verbatim. That approach to feedback processing (comparison and copying) yielded better learning outcomes than simply studying feedback.

Finally, it should be emphasized that feedback processing strategies in the absence of errorful generation vary across real-world pedagogical contexts. In some instances, students may be simply asked to read or study feedback, as commonly implemented by U.S. undergraduate educators (e.g., Fyfe et al., 2021), whereas more elaborate practices are used in other contexts, such as schools in Hong Kong and Malaysia, where the practice of engaging in "corrections," where students copy and reflect on correct answer feedback, is common (e.g., Lee, 2013; Sidhu & Fook, 2010). Similarly, in countries such as Japan—where critically reflecting on failure is regarded as integral for self-improvement (i.e., *hansei*)—feedback from educators is often viewed as infallible, resulting in a greater tendency for learners to engage in verbatim copying and modeling of correct answer feedback (e.g., Kozato et al., 2023; Stephens, 2020).

Feedback types and learning

The research literature on different types of corrective feedback further reveals that not all forms of feedback are equally effective (for a taxonomy, see Kulhavy & Stock, 1989). For example, Pashler et al. (2005) compared the effects of correct answer feedback with right-or-wrong feedback on the learning of word translations, finding that correct answer feedback was more effective. Crucially, that pattern was specific to instances of incorrect responding; if a learner's response was correct, then feedback type was inconsequential. Other feedback types involving more than just a correct answer, such as explanations or examples—which can be classified as *elaborative* feedback (Butler et al., 2013)—can also benefit learning. In the context of concept learning, for example, elaborative feedback can yield more transferrable learning than simple correct answer feedback (e.g., Butler et al., 2013; Pan, Hutter, et al., 2018).

This study

Across four experiments, we investigated the impact of different feedback processing strategies on learning the definitions of unfamiliar, low-frequency English words through errorful generation. In each experiment, participants underwent a learning phase where they engaged in errorful generation with correct answer feedback to learn definitions (a stimulus type that has been a focus of prior studies; e.g., Potts & Shanks, 2014) and then took a posttest. Each experiment compared generative (i.e., *rephrasing*, *example generation*) and non-generative (i.e., *reading*, *copying*, *studying keyword mnemonics*) feedback processing strategies. These strategies were drawn from prior research on evidence-based learning techniques (Cummings et al., 2023; Dunlosky et al., 2013)

Table 1. Feedback processing strategies.

Expt.	Strategy	Definition	Generative?	Elaborative?	Active?
1, 2, 3, 4	Rephrasing	Rewriting the definition in one's own words	Yes	Yes	Yes
1, 2, 4	Reading	Reading the word and its definition	No	No	No
2	Copying	Transcribing the definition verbatim	No	No	Yes
3	Example generation	Create a sentence that uses the word in a meaningful way	Yes	Yes	Yes
3	Keyword mnemonic	Using a provided mnemonic to help learn the word's meaning	No	Yes	Yes

Note. Generative refers to whether new content is being created or constructed; elaborative refers to mental processing that involves more than just encoding the exact provided information; and active refers to manipulating content as opposed to passive learning.

and represent ways of processing correct answer feedback that differ qualitatively and on other dimensions.

As shown in Table 1, we classified rephrasing (i.e., rewriting the definition in one's own words) and example generation (i.e., using the word in a meaningful sentence) as generative, elaborative, and active learning strategies. Both strategies require learners to produce semantically or contextually relevant ideas that extend beyond the information provided in corrective feedback, and the process involved is not passive. In contrast, copying (i.e., transcribing the definition verbatim) is a non-generative, non-elaborative, but active processing strategy, as it involves the physical act of typing but does not engage learners in creating new content or require added mental operations. Studying keyword mnemonics, which are designed to make a word and its definition or translation more memorable, is also a non-generative strategy, as the mnemonic is provided directly to the learner. Although it does not involve creating new material, the keyword mnemonic engages active mental visualization processes and involves mental elaboration on the to-be-learned materials. Finally, reading (i.e., the word and its definition) is non-generative, arguably non-elaborative, and passive, as it involves simply receiving the feedback without overt manipulation or cognitive effort. It was hypothesized that generative strategies, which engage learners in higher levels of cognitive and constructive effort, would lead to better learning outcomes compared to non-generative strategies.

Additionally, we investigated elaborative feedback through the provision of examples. In Experiments 1 and 2, participants received correct answer feedback (i.e., target words and their definition) that did or did not also contain an example sentence which used the target word in a manner appropriate for its meaning. Analogous to the benefits of explanation feedback (Butler et al., 2013), it was hypothesized that example-enhanced feedback would yield greater word learning accuracy by providing learners with more detailed context for learning the target word definitions. The provision of keyword mnemonics in Experiment 3 also constituted a form of elaborative feedback.

Experiment 1

Experiment 1 compared, in the context of errorful generation, the effectiveness of two feedback processing strategies: the generative strategy of rephrasing and the non-generative strategy of reading. It also examined whether providing examples as a form of elaborative feedback confers learning benefits. A non-errorful generation control condition was further included for comparison purposes. Ultimately, this experiment addressed the question: Is generative feedback processing (*via* rephrasing) more effective than reading, and as a secondary question, is providing example sentences as feedback also beneficial?

Method

Participants

An a priori power analysis using G*Power (Faul et al., 2009) was conducted to determine the minimum sample size needed to detect a small-to-medium effect size (cf. Cohen, 1988) of Cohen's $f = 0.20$ with 80% power at $\alpha = 0.05$ for a within-subjects design with two independent variables: Processing Condition (3 levels: rephrasing, reading, and copying) and Example Feedback (2 levels: yes and no). The copying condition served as the non-errorful generation control, in which participants did not guess the word meaning but simply copied the definition. For reference, Pashler et al. (2005) observed a large effect size for correct answer feedback over right-or-wrong feedback of Cohen's $d \approx 1.0$; we sought to power for potentially smaller effect sizes. In the context of a repeated-measures 3×2 ANOVA, the analysis was designed to detect a main effect of one factor, such as Processing Condition or Example Feedback. The study was not specifically powered to detect an interaction between the two factors, as we had no a priori reason to expect an interaction and were primarily interested in the main effect of Processing Condition, followed by any main effect of Example Feedback (it is acknowledged that this experiment may have been underpowered to detect an interaction). A minimum sample size of 28 participants was indicated but given uncertainty about how different feedback processing strategies might impact learning, to provide a buffer in case of possible data exclusions, and given conventional minimum sample sizes in psychological research (Marszalek et al., 2011), we targeted a larger sample size of at least 40 participants (which applied to Experiments 1–2).

Fifty-one participants were recruited from the participant pool at a large research university in Southeast Asia in exchange for course credit or a remuneration of SGD$10. Participants had to be at least 18 years of age, fluent in English, and to be included in the analysis, demonstrate low familiarity with the target words (i.e., guessing performance being three standard deviations higher than mean performance). Because the experiment was conducted in person, an experimenter confirmed participants' ability to follow instructions and complete the tasks, and none disclosed any developmental or mental conditions or medication use that would affect performance. One participant was removed prior to formal analysis due to evidence of prior familiarity with the stimuli. The final sample had 50 participants with a mean age of 21.6 years and was 72% female.

All experiments in this study were conducted at the National University of Singapore under ethics board approval (Psych-DERC Reference Code: 2024-June-05). All participants provided informed consent before participating and were treated in accordance with the principles outlined in the Declaration of Helsinki.

Design

Experiment 1 used a 3 (Processing Condition: rephrasing/reading/copying) $\times$ 2 (Example Feedback: Yes/No) within-subjects design. The dependent variable was the proportion of correct answers on the posttest. For each participant, target words were randomly assigned to one of the three Processing Conditions, with half of the words in each condition also randomly assigned to receive example feedback while the remaining half received no examples alongside feedback. The order of the three Processing Conditions was counterbalanced across participants and the order of target words was randomized within each condition.

Materials

Thirty-six low-frequency English verbs and their definitions were used as the learning stimuli (e.g., *asperse: attack or criticize another's image*). The words were extracted from the SUBTLEX-US (Brysbaert & New, 2009) database and the Thorndike and Lorge (1944) list, with each word

Table 2. Example word-definition stimuli.

Target word	Definition	Example Feedback (Experiments 1–2)	Keyword Mnemonic (Experiment 3)
aberrate	deviate from what is normal or correct	Her behavior began to aberrate from her usual calm demeanor after the intense argument.	Imagine: The airplane flew an **AB**NORMAL route with a high **RATE** of descent.
cavil	make petty or unnecessary complaints	Rather than appreciating the team's progress, the boss continued to cavil about minor setbacks.	Imagine: The **CA**TTY lady kept making UNCI**VIL** comments about little details.
kibitz	provide someone with unwelcome advice	The woman liked to kibitz about her relatives' life choices during every family reunion.	Imagine: The **KI**D could not resist the **IT**CH to give tips that no one asked for.
poleaxe	bring about incredible shock to someone	The unexpected announcement about their leader's passing seemed to poleaxe the citizens.	Imagine: The crowd gasped when the seemingly harmless **POLE** turned a sharp **AX**.
wassail	drink and celebrate in a joyous manner	As the clock struck midnight, the crowd gathered to wassail and welcome the New Year.	Imagine: The celebrities S**WA**YED to the music and set **SAIL** to the party island.

being 5–8 letters long, containing 1–3 syllables, and having a frequency of one or fewer occurrences per million words. The definitions, each between 4 and 8 words long, were adapted from the Oxford University Press (2025) and Collins Online Dictionary (2025). For the Yes level of the Example Feedback condition, example sentences of between 11 and 19 words in length were generated (e.g., "The gossip columnist was known to *asperse* celebrities with sensational and false stories."). Examples of the stimuli are presented in Table 2, whereas the entire stimulus set and other study materials are archived on the Open Science Framework (https://osf.io/c5f6g/?view_only=1bca1000941f42f08b75277cd61aefb4).

Throughout the study, every target word was presented alongside a five-option multiple-choice question during both learning and posttest phases, requiring selection of the correct definition apart from four distractor definitions. Each word had unique distractor options that were semantically distinct yet plausible enough to not be easily dismissed (e.g., mirroring meanings of other words with phonetic or orthographic properties similar to the target word). To maintain consistency with the correct definition, distractor definitions were also kept between 4 and 8 words long. The sole exception was the non-errorful generation control condition, wherein the target word and correct definition were presented simultaneously with no additional information.

It should be noted that we selected obscure English verbs and their definitions because they were well-suited for the feedback-based learning methods investigated in this study. The mnemonic benefits of errorful generation with such materials—where participants have little to no prior knowledge or preexisting semantic associations—are however sometimes weaker than those observed with materials that are more familiar or semantically linked (e.g., associated word pairs; cf. Potts et al., 2019; Seabrooke et al., 2022).

Procedure

Experiment 1 used a general procedure that was largely maintained across subsequent experiments. Participants completed a 60-minute session in a psychology laboratory room, working individually at privacy-walled computer stations under supervision. After giving consent and answering demographic questions, they proceeded to the learning phase, with the three Processing Conditions presented as three separate sections, each with 12 target words and their corresponding definitions (i.e., as feedback), presented one word at a time, with random assignment and counterbalancing as described previously. The entire procedure is depicted in Figure 1 and details of the different sections, each of which took around 15 min (except where noted), are as follows.

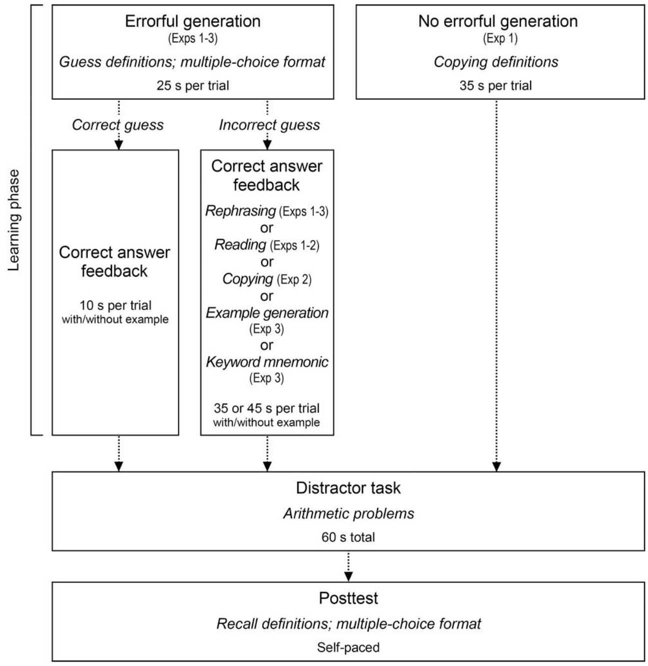

Figure 1. Overview of procedure from Experiments 1–3. In each experiment, participants learned definitions of obscure words via errorful generation with correct answer feedback. If they guessed correctly, they simply viewed that feedback, but if not, they rephrased it (Experiments 1–3), read it (Experiments 1–2), copied it (Experiment 2), or generated an example or used a provided keyword mnemonic (Experiment 3). A control condition that involved copying definitions with no errorful generation beforehand was included in Experiment 1.

Learning phase. At the start of each section, instructions describing what participants were about to experience, including how they were to process any feedback, were presented. In both feedback Processing Conditions, participants were told they would guess word meanings *via* multiple-choice questions before receiving corrective feedback. It was stated that the feedback would indicate whether their choice guess was correct or incorrect. In the control condition, words and definitions would be presented simultaneously for participants to copy. Before each condition, participants further completed a practice trial demonstrating the procedure for that section.

Rephrasing. For each of the 36 words, participants were first shown a five-option multiple-choice question and asked to select the option that they believed best represented the definition of the target word. After 25 s had elapsed, correct answer feedback was shown. If answered accurately, the correct definition of the word was shown for 10 s, with or without an example sentence, depending on Example Feedback condition (if Yes, then the sentence was preceded by the instruction: "Example of [word] used in a sentence."). If answered wrongly, the same information (including an example sentence, if Yes under Example Feedback condition) was shown along with instructions to paraphrase the definition of the target word in the participants' own words, for an extended duration of 35 s instead of 10 s. Once the timer had elapsed, participants automatically proceeded to the next target word (Note: if a participant had not selected an option within the initial 25-s limit for a multiple-choice question, they were given an additional 25 s to submit a response; as described in the Scoring section, data from words involved in such trials were removed from analysis).

Reading. The procedure for this condition mirrored the rephrasing condition, including the provision of correct answer feedback with or without an example sentence. However, when participants answered incorrectly, rather than processing the feedback in any overt way, they were simply asked to mentally read the provided definition of the word (and any example sentence if it was provided) without engaging in any typing, and as in the rephrasing condition, were given 35 s to do so.

Copying (no errorful generation). For each word, participants were presented with that word along with its definition and given 35 s to type out the definition exactly as shown, matching every single letter and spacing. An example sentence was displayed or not depending on Example Feedback condition. It should be noted that the time allotted equated exposure to the correct answer across the errorful generation (rephrasing or reading) and copying (no errorful generation) conditions, consistent with standard design practices in errorful generation research (e.g., Huelser & Metcalfe, 2012). Furthermore, prior studies that provided additional exposure to correct answers in non-errorful generation conditions have found that the benefits of errorful generation are either preserved or somewhat reduced, but not reversed (e.g., Kornell et al., 2009).

Distractor task. Immediately after the learning phase, participants answered up to 60 basic arithmetic questions as accurately as possible within one minute.

Posttest. Thirty-six five-option multiple-choice questions (one per learned word, identical to the learning phase format) were presented individually in random order. Participants had to choose the correct definition of each word, without time constraints. Afterwards, they were debriefed and dismissed.

Scoring

All experiments used the same scoring system. Each word was scored 0 or 1 based on correct definition selection. Feedback processing responses (e.g., rephrasing) were also checked by the experimenter for adherence to instructions; incomplete or imprecise responses were accepted if procedures were followed. Trials were excluded if participants exceeded the 25-s limit without responding (0.44%) or violated feedback instructions (0.12%), resulting in the removal of 20 trials, hence affecting 10 words across eight participants (representing 0.56% of all the participants' trials in Experiment 1).

Results

Learning phase performance

The mean proportion of correct initial guesses in the learning phase was $M = 0.18$, SD $= 0.15$ in the rephrasing condition and $M = 0.14$, SD $= 0.15$ in the reading condition. Since each multiple-choice question had five answer options, a guessing accuracy rate below chance (0.20) indicates that participants were largely unfamiliar with the words, providing a baseline for evaluating how Processing Condition and Example Feedback influenced word learning. That pattern was observed across subsequent experiments.

A paired-samples t-test revealed that guessing performance was higher for rephrasing, $t(99) = 1.99$, $p = .0497$, $d = 0.20$ (95% CI = [0.00, 0.40]). This pattern, which occurred unexpectedly despite random assignment of target words to condition and was not replicated in subsequent experiments, implies that the words randomly assigned to the rephrasing condition were somewhat easier to guess. To address this imbalance, we conducted a supplementary posttest analysis that statistically controlled for guessing accuracy and was restricted to the rephrasing and reading conditions (as the copying condition did not involve guessing). The results of this analysis are reported in the Supplementary Analysis section that follows.

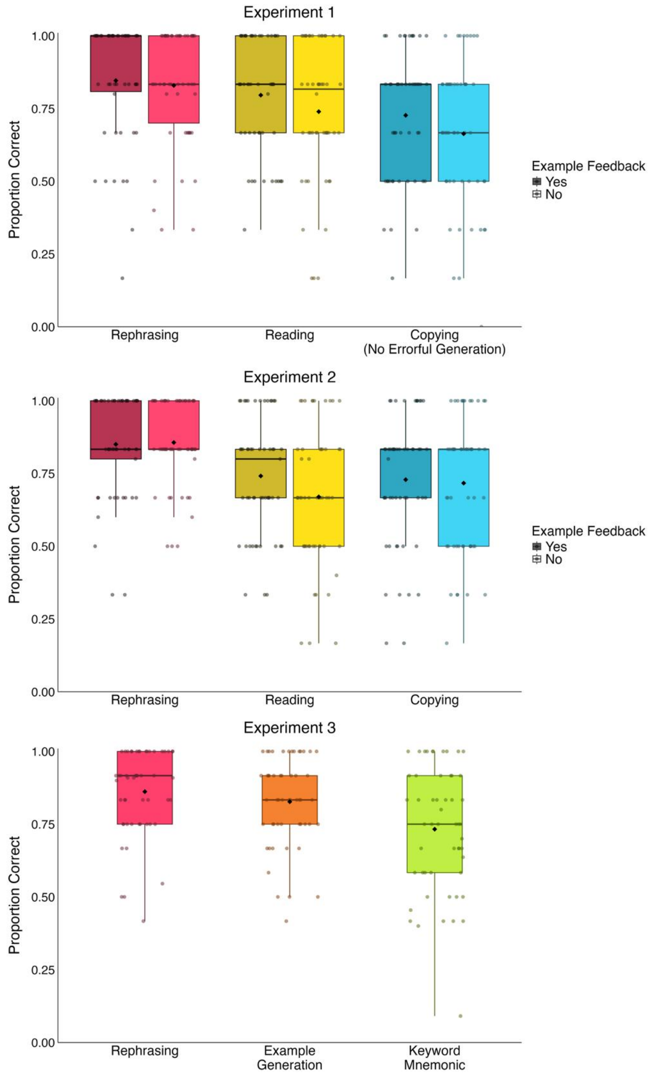

Figure 2. Posttest results of Experiments 1–3.

Posttest performance

Posttest results for each condition are depicted in the upper panel of Figure 2 (including data from all posttest trials). In all experiments, participant-level mean posttest accuracy scores for each condition were analyzed using Analysis of Variance (ANOVA). Consistent with the errorful generation literature (Kornell et al., 2009), analyses for Experiments 1–3 were performed on posttest data that, for both Feedback Processing conditions, participants had guessed incorrectly during the learning phase (as those words were uniquely subject to generative or non-generative feedback processing). It is noted, however, that in supplementary analyses not reported here, analyses performed on all posttest data yielded identical patterns. Greenhouse-Geisser corrections were applied for violations of Mauchly's sphericity test (ε range = 0.83–0.99).

Effect of Processing Condition and example feedback. A 3 (Processing Condition) × 2 (Example Feedback) within-subjects ANOVA on posttest accuracy scores revealed a significant main effect of Processing Condition, $F(1.84, 90.26) = 16.25$, $p < .001$, $\eta_p^2 = 0.25$, indicating significant differences in word learning accuracy across the rephrasing, reading, and copying conditions; a significant main effect of Example Feedback, $F(1, 49) = 7.52$, $p = .008$, $\eta_p^2 = 0.13$, indicating higher accuracy when target words were presented with example sentences than when presented alone; and no significant interaction, $F(1.82, 89.30) = 0.23$, $p = .771$, $\eta_p^2 = 0.005$. Follow-up pairwise post-hoc comparisons with Tukey adjustment revealed that the rephrasing condition produced significantly higher accuracy than both the reading, $t(49) = 2.96$, $p = .0129$, $d = 0.42$ (95% CI = [0.13, 0.71]), and copying conditions, $t(49) = 6.76$, $p < .001$, $d = 0.97$ (95% CI = [0.62, 1.30]). However, the difference between the reading and copying conditions was not significant, $t(49) = 2.37$, $p = .0563$, $d = 0.34$ (95% CI = 0.05, 0.62]), although the p value was marginal. This pattern is broadly consistent with the weaker benefits of errorful generation for obscure word definitions that are not familiar or semantically linked (cf. Potts & Shanks, 2014).

Supplementary analysis involving guessing accuracy. To address the possibility that items in the rephrasing condition may have been easier (as implied by initial guessing performance), we conducted a supplementary analysis restricted to the rephrasing and reading conditions. A linear mixed-effects model was fit with Processing Condition (rephrasing versus reading) as a fixed effect, initial guessing accuracy as a covariate, and random intercepts for participants to account for within-subject variability. That analysis, which was not preregistered and can be regarded as exploratory, found a significant main effect of Processing Condition, $F(1, 148.62) = 14.67$, $p < .001$, suggesting that rephrasing still yielded greater posttest accuracy than reading even after controlling for initial guessing.

Experiment 2

Experiment 1 revealed that a generative feedback processing strategy, rephrasing, outperformed a non-generative strategy, reading. In fact, an errorful generation benefit versus a read-only control was only observed following rephrasing. Further, the presence of example sentences as a form of elaborative feedback also enhanced learning. That experiment did not address, however, whether the advantage of rephrasing stemmed from the act of generating new content *per se* or simply the act of typing, which occurred with rephrasing but not reading. To adjudicate this issue, Experiment 2 compared rephrasing against not just reading but also copying (after errorful generation). If performing an active feedback processing strategy such as typing is sufficient to enhance learning, then transcribing information *via* copying after an incorrect guess should yield similar learning as rephrasing. On the other hand, if generative processing is required, then rephrasing should be more effective. The effects of examples in feedback were also reexamined.

Method

This experiment was preregistered at: https://aspredicted.org/76xw-g922.pdf.

Participants

Fifty-four participants were recruited in the same manner as for Experiment 1. As in that experiment, participants were fluent in English, demonstrated low familiarity with the target words, and were confirmed by the experimenter to be able to follow instructions and complete the tasks; none disclosed any conditions or medication use that would affect performance. One person was excluded due to their guessing performance being three standard deviations higher than mean performance, resulting in a final sample of 53 participants that had a mean age of 23.3 years and was 66% female.

Design, materials, procedure, and scoring

The design, materials, and procedure remained largely identical to Experiment 1, with the primary modification being the addition of a copying condition in place of the prior (non-errorful generation) control condition. This condition included a 25-s guessing period for each learning phase trial, consistent with the rephrasing and reading conditions. If the guess was correct, feedback was shown for 10 s; if not, participants had 35 s to transcribe the correct definition exactly as shown, matching every single letter and spacing.

In addition, participants completed an exit survey after the posttest where they ranked feedback processing methods by usage, preference, and perceived effectiveness, rated the frequency of receiving standard corrective feedback versus example-enhanced corrective feedback, and rated the perceived effectiveness of example inclusion. Survey data are considered after the results for Experiment 4.

Exclusions were applied to trials that exceeded the time limit for initial guessing (0.21%), violated feedback processing instructions (0.26%) or encountered technical failures (0.05%), resulting in the removal of 20 trials (0.52%) across all participants' trials and affecting 10 words across eight participants.

Results

Learning phase performance

A one-way within-subjects ANOVA revealed no significant difference in mean initial guessing accuracy across the rephrasing ($M = 0.19$, SD $= 0.15$), reading ($M = 0.16$, SD $= 0.15$), and copying ($M = 0.16$, SD $= 0.15$) conditions, $F(1.97, 102.68) = 1.25$, $p = .290$, $\eta_p^2 = 0.02$, indicating similar item difficulty across conditions.

Posttest performance

Posttest results for each condition are depicted in the middle panel of Figure 2 (including data from all posttest trials). A 3 (Processing Condition) × 2 (Example Feedback) within-subjects ANOVA analogous to that performed for Experiment 1 revealed a significant main effect of Processing Condition, $F(1.98, 102.81) = 35.43$, $p < .001$, $\eta_p^2 = 0.41$. Unlike Experiment 1, there was no significant main effect of Example Feedback, $F(1, 52) = 0.85$, $p = .361$, $\eta_p^2 = 0.02$, suggesting that the presence of examples was not beneficial. There was also no significant interaction, $F(1.95, 101.14) = 1.38$, $p = .255$, $\eta_p^2 = 0.03$. Post-hoc Tukey comparisons revealed that rephrasing produced significantly higher accuracy than reading, $t(52) = 7.14$, $p < .001$, $d = 0.99$ (95% CI = [0.66, 1.32]), and copying, $t(52) = 7.43$, $p < .001$, $d = 1.03$ (95% CI = [0.69, 1.36]). The reading and copying conditions did not differ from each other, $t(52) = -0.393$, $p = .919$, $d = -0.05$ (95% CI = [−0.33, 0.22]). These results suggest that the advantage of rephrasing stems specifically from its generative nature and not simply from any added engagement or other learning benefits due to typing.

Experiment 3

Experiments 1 and 2 demonstrated that generative feedback processing, as occurred during rephrasing, yields better learning. Experiment 3 addressed whether that advantage is specific to rephrasing or other generative strategies more generally. It compared rephrasing to another generative strategy, example generation, as well as a non-generative but elaborative strategy, the keyword mnemonic. In the case of example generation, participants were required to produce an original sentence that used the target word in an appropriate and meaningful way. The keyword mnemonic, which is commonly used to learn the meaning of foreign vocabulary words, is an evidence-based learning technique that involves forming a mental image between a keyword (i.e., a similar sounding word or fragment of the target word) and the meaning of that word (Atkinson & Raugh, 1975; Cummings et al., 2023; Dunlosky et al., 2013). Given the fact that participants were generating their own examples in the example generation condition, the Example Feedback factor was dropped.

Method

This experiment was preregistered at: https://aspredicted.org/3fbp-5br7.pdf.

Participants

An a priori power analysis was conducted using G*Power (Faul et al., 2009) to determine the minimum sample size required to detect a small-to-medium effect size of Cohen's $f = 0.20$ with 80% power at $\alpha = 0.05$ for a within-subjects design with one independent variable consisting of three levels. The analysis indicated that a minimum of 42 participants would be needed. To account for potential attrition or data exclusions, a target sample size of 50 participants was set. Fifty-six participants were recruited in the same manner as in the preceding experiments, with fluency in English, low familiarity with the target words, and verification by the experimenter that they could follow instructions and complete the tasks without any conditions or impairments affecting performance. Data were removed from two participants for noncompliance with instructions and from one participant due to their guessing performance exceeding three standard deviations above the mean, resulting in a final sample of 53 participants that had a mean age of 21.2 years and was 70% female.

Design and materials

Experiment 3 employed a within-subjects single-factor design with Processing Condition (rephrasing/example generation/keyword mnemonic) as the single factor. Apart from the example sentences from the discontinued Example Feedback factor, all materials from the prior experiments were retained. Additionally, for each word assigned to the keyword mnemonic condition, a keyword mnemonic was created (drawing on Cummings et al., 2023; Qu et al., 2024). Specifically, two keywords sharing letter sequences with the target word were embedded in a sentence between 8 and 16 words long. The mnemonic was carefully designed to form a memorable, semantically relevant scenario that functioned as an associative cue for the word (e.g., "The reviewer **AS**SAULTED the novelist's **PERS**ONAL beliefs.", with the bold letters representing shared letters with the target word *asperse*). Examples of the keyword mnemonics are presented in Table 2.

Procedure

The overall procedures remained unchanged, including instructions for the rephrasing condition, whereas several novel procedures were introduced for the two new Processing Conditions.

Example generation. Participants who made an incorrect initial guess were tasked to construct their own sentence with vivid and elaborative details to demonstrate correct contextual usage of the target word (as if helping a hypothetical learner better comprehend the word). Participants were further instructed to refrain from simply embedding the target word in a random sentence that did not properly illustrate the word's application.

Pilot testing revealed that participants required additional time to generate examples, prompting an extension of the allotted time from 35 s (as in prior experiments) to 45 s. To maintain equal time on task, this adjustment was implemented across all feedback processing conditions.

Accurate guesses received the same 10-second feedback display reiterating the word's correct definition, as in prior experiments.

Keyword mnemonic. Participants who guessed incorrectly were provided with the keyword mnemonic and instructed to imagine or mentally visualize the sentence described, without engaging in any typing. Participants were given 45 s to do so, as in the other processing conditions.

Scoring

Time limit violations for initial guessing led to the removal of 38 trials (1.00%) across all participants, impacting 19 words among 11 participants.

Results

Learning phase performance

A one-way within-subjects ANOVA revealed no significant difference in mean initial guessing accuracy across the rephrasing ($M = 0.15$, SD = 0.11), example generation ($M = 0.16$, SD = 0.12), and keyword mnemonic ($M = 0.16$, SD = 0.11) conditions, $F(1.98, 102.99) = 0.20$, $p = .813$, $\eta_p^2 = 0.004$, indicating similar item difficulty across conditions.

Posttest performance

Posttest results for each condition are depicted in the bottom panel of Figure 2 (including data from all posttest trials). A one-way within-subjects ANOVA on posttest accuracy scores found a significant main effect of Processing Condition, $F(1.66, 86.43) = 14.41$, $p < .001$, $\eta_p^2 = 0.22$. In post-hoc Tukey comparisons, no significant difference emerged between rephrasing and example heneration, $t(52) = 0.263$, $p = .963$, $d = 0.04$ (95% CI = [−0.24, 0.31]), suggesting that both forms of generative feedback processing were similarly effective. Rephrasing was more effective than the keyword nnemonic condition, $t(52) = 4.19$, $p < .001$, $d = 0.58$ (95% CI = [0.28, 0.87]), and example generation also outperformed the keyword mnemonic condition, $t(52) = 4.22$, $p < .001$, $d = 0.59$ (95% CI = [0.29, 0.88]). Hence, generative feedback processing yielded better learning than a potentially elaborative, active, but non-generative approach.

Experiment 4

In prior experiments, rephrasing was always paired with errorful generation. To further characterize the contributions of rephrasing to learning, the final experiment examined its efficacy alongside errorful generation and, for the first time, in its absence as well. This approach allowed for separate assessment of the effects of rephrasing and, similar to Experiment 1, compared errorful generation with a non-errorful generation control condition.

Method

Experiment 4 was preregistered at: https://aspredicted.org/r9t3-6mkk.pdf.

Participants

An a priori power analysis using G*Power (Faul et al., 2009) was conducted to determine the minimum sample size needed to detect an effect size of Cohen's $f = 0.20$ with 80% power at $\alpha = 0.05$ for a within-subjects design with two factors: Acquisition Method (2 levels: errorful generation and no errorful generation) and Processing Condition (2 levels: rephrasing and reading).

Consistent with earlier experiments, the study was powered to detect main effects (although an interaction would be theoretically interesting, we did not advance a specific hypothesis regarding one, and adequately powering the design to detect such an effect would have required substantially larger samples beyond the scope of the present investigation). The analysis indicated that a minimum of 52 participants was required, and to account for potential attrition or data removal, we targeted a larger sample size of at least 60 participants.

Seventy-four participants were recruited online from Prolific Academic, a crowdsourcing platform known for relatively good data quality (Palan & Schitter, 2018), in exchange for a payment of at least USD$8.00. All participants had to be at least 21 years of age and fluent in English. Sixteen participants were removed for evidence of off-task activity (i.e., time spent away from the experiment window, in violation of instructions), two participants were removed due to technical errors, and one participant was removed due to their posttest performance falling three standard deviations below the mean. The final sample included 55 participants—thus exceeding the minimum number needed per the aforementioned power analysis—with a mean age of 34.6 years and 60% of the sample being female. Forty percent of the participants were from the United States, 34.5% were from the United Kingdom, 18.2% were from Canada, and the remainder were from other countries; 18.2% reported answered in the affirmative to a screening question about whether they were a student.

Design, materials, procedure, and scoring

As depicted in Figure 3, Experiment 4 built on the prior experiments but included several modifications for research purposes and other design considerations, including appropriate task difficulty and maintaining attention for online participants. A 2×2 within-subjects factorial design was used in which each participant learned the same 36 English verbs and their definitions as previously but under four conditions created by crossing the factors of Acquisition Method and Processing Condition: Errorful generation + rephrasing, errorful generation + reading, no errorful generation + rephrasing, and no errorful generation + reading. Unlike the prior experiments, these conditions were not presented in separate sections; rather, items from all four conditions were randomly intermixed during the learning phase (Gupta et al., 2024). Specifically, for each trial, participants encountered a word item from one of the four conditions, with trials from different

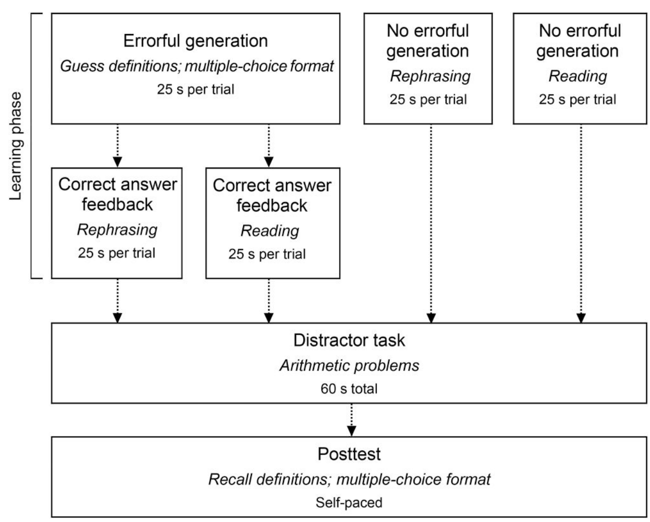

Figure 3. Overview of procedure from Experiment 4. In Experiment 4, participants learned definitions of obscure words via errorful generation with correct answer feedback or no errorful generation (which served as the control condition). Crucially, they did or did not engage in rephrasing of the definitions in the errorful generation and no errorful generation conditions.

conditions randomly intermixed rather than presented in blocked sequences, while maintaining an overall equal distribution of nine items per condition. To minimize carryover effects and possible confusion across trials, condition-specific instructions written in capital letters were provided on each trial to emphasize the strategy to be used on the current trial. Moreover, due to the online setting, the extent of experimenter monitoring level and control differed from the prior experiments; to help ameliorate this issue, a monitoring tool was programmed into the experiment to detect cases where participants strayed away from the experiment window (Permut et al., 2019).

In the errorful generation conditions, learning occurred very much as in the prior experiments, with participants being given 25 s to guess the definition of a given word in multiple-choice format, followed by feedback processing *via* rephrasing (errorful generation + rephrasing) or reading (errorful generation + reading). To reduce difficulty and study duration, however, only three multiple-choice answer options were provided rather than five options, and feedback processing time was reduced from 35 to 25 s. Additionally, all participants engaged in feedback processing regardless of whether they had guessed correctly or not, thus ensuring equal amounts of such processing across errorful generation conditions. In the no errorful generation conditions, participants were also given 25 s to examine a given word and its definition, and during that time, to rephrase the definition (no errorful generation + rephrasing) or just read it (no errorful generation + reading). Examples were not provided in any condition.

After the learning phase, the distractor task and posttest occurred just as in the prior experiments, including five answer options per word on the posttest. Scoring also occurred identically as in the prior experiments.

Results

Learning phase performance

A paired-samples *t*-test revealed no significant difference in mean initial guessing accuracy across the errorful generation + rephrasing ($M = 0.45$, SD = 0.26), and errorful generation + reading ($M = 0.46$, SD = 0.22) conditions, $t(54) = -0.31$, $p = .757$, $d = -0.04$ (95% CI = [−0.31, 0.22]), indicating similar item difficulty across conditions. Guessing accuracy was higher than in prior experiments, reflecting the fewer number of answer options per item. Those accuracy levels constitute a departure from typical patterns observed in studies of errorful generation and are further addressed in supplementary analyses described below.

Posttest performance

Posttest results for each condition are depicted in Figure 4 (including data from all posttest trials). Analyses were performed on all posttest items for this experiment given that all items in the errorful generation conditions received feedback processing (similar to Experiment 1, a supplementary analysis also examined posttest performance separately for items initially guessed incorrectly versus correctly). A 2 (Acquisition Method) × 2 (Processing Condition) within-subjects ANOVA on posttest accuracy scores revealed a significant main effect of Acquisition Method, $F(1, 54) = 21.59$, $p < .001$, $\eta_p^2 = 0.29$, indicating an overall benefit of errorful generation, and a significant main effect of Processing Condition, $F(1, 54) = 16.63$, $p < .001$, $\eta_p^2 = 0.24$, reflecting an overall benefit of rephrasing. The interaction was not significant, $F(1, 54) = 1.02$, $p = .318$, $\eta_p^2 = 0.02$. Because each factor had only two levels, the main effects correspond directly to pairwise comparisons on data collapsed across the other factor. Collapsing across feedback, the mean accuracy for Errorful Generation groups combined ($M = 0.83$, SD = 0.20) was higher than the No Errorful Generation groups combined ($M = 0.73$, SD = 0.23) ($d = 0.63$, 95% CI = [0.34, 0.92]), and collapsing across acquisition method, the mean accuracy for Rephrasing groups combined

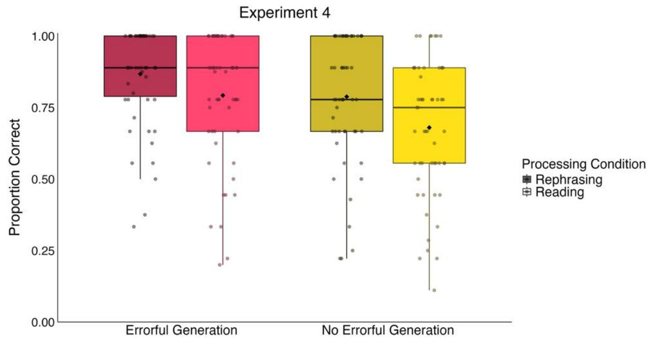

Figure 4. Posttest results of Experiment 4.

($M = 0.83$, SD $= 0.20$) was higher than the reading groups combined ($M = 0.74$, SD $= 0.23$) ($d = 0.55$, 95% CI = [0.27, 0.84]).

Supplementary analysis involving guessing accuracy

In an exploratory analysis, we examined how guessing accuracy might interact with feedback processing methods in the errorful generation conditions (to reiterate, unlike the prior experiment, all items in the errorful generation conditions received additional feedback processing). Posttest scores were not statistically different for items initially guessed correctly versus incorrectly in the errorful generation + rephrasing condition ($M = 0.87$, SD $= 0.25$ versus $M = 0.87$, SD $= 0.20$, $t(50) = 0.01$, $p = .989$, $d = 0.002$, 95% CI = [−0.28, 0.28]). In the errorful generation + reading condition, however, posttest scores were significantly higher for items initially guessed correctly than incorrectly ($M = 0.86$, SD $= 0.26$ versus $M = 0.76$, SD $= 0.27$, $t(52) = 2.37$, $p = .0217$, $d = 0.33$, 95% CI = [0.05, 0.61]). Together, these results suggest that when feedback is processed in a generative manner, both correctly and incorrectly guessed items are learned to a similar extent, yielding comparable posttest results, whereas when feedback is processed only superficially through reading, an advantage for initially correct guesses persists (perhaps reflecting better prior knowledge).

Survey of feedback experiences and beliefs

In the exit survey for Experiments 2–4, participants rank-ordered the processing strategies they experienced across three dimensions—frequency of usage, personal preference, and perceived effectiveness—both in school (i.e., formal educational settings) and external self-regulated learning contexts (cf. Huelser & Metcalfe, 2012). For Experiments 2–3, these rankings reflect educational experiences in Singapore, where the study took place, whereas for Experiment 4, the rankings reflected the experiences of the more diverse online sample (including participants from the USA, the UK, and Canada).

Among Experiment 2 respondents—who rated rephrasing, reading, and copying—copying was most frequently performed in school (49%), whereas reading was most common during self-regulated learning (53%). A plurality of participants preferred copying in school (36%) and rephrasing in self-regulated learning (38%), with rephrasing also ranked as most effective in both settings (school: 62%; self-regulated learning: 51%). Experiment 2 respondents also rated the frequency and effectiveness of example feedback on a 7-point Likert scale. A plurality rated it as frequently received (school: 28%; self-regulated learning: 34%) and very effective (school: 40%; self-regulated learning: 34%).

Among Experiment 3 respondents—who rated rephrasing, example generation, and the keyword nnemonic—rephrasing was most frequently used in school (47%), but example generation was most common in self-regulated learning (51%), and it was also the top-rated strategy for both preference (school: 59%; self-regulated learning: 51%) and perceived effectiveness (school: 55%; self-regulated learning: 57%). Among Experiment 4 respondents, no errorful generation + reading was reported to be most frequently performed in both school (62%) and self-regulated learning contexts (53%). In terms of participants' personal preferences, no errorful generation + reading was rated highest in school (47%), whereas no errorful generation + rephrasing was rated highest in self-regulated learning (33%). For perceived effectiveness, there was an effective tie between no errorful generation + rephrasing and no errorful generation + reading for school (33%), while no errorful generation + reading was highest for self-regulated learning (38%).

Overall, the exit survey results indicate that the prevalence of feedback processing strategies varies by setting. Rephrasing and example generation were seen as the most effective in Experiments 2 and 3, respectively—judgments that matched actual outcomes in Experiment 2 but underestimated Rephrasing in Experiment 3. In Experiment 4, where participants were asked to rate errorful generation *per se,* it is particularly interesting that it was not rated less highly than no errorful generation in terms of frequency, preference, and perceived effectiveness (regardless of feedback processing method), which contrasts with the actual outcomes of the experiment. That disparity echoes prior research findings (e.g., Huelser & Metcalfe, 2012; see also Pan et al., 2020).

Discussion

The present study reveals that the manner with which corrective feedback is processed can influence learning outcomes following errorful generation. Across four experiments, generative feedback processing yielded the best learning outcomes. Specifically, in Experiments 1 and 2, generative feedback processing *via* rephrasing produced significantly higher posttest accuracy compared to the non-generative strategies of transcription *via* copying ($d = 1.03$) and passive reading ($d = 0.99$), the latter being the most common approach to feedback processing in the errorful generation literature. Further, in Experiment 3, an active and elaborative but non-generative processing strategy, using the keyword mnemonic, still paled in comparison to the generative strategies of rephrasing ($d = 0.58$) and example generation ($d = 0.59$).

Experiment 4 additionally revealed that the benefit of rephrasing was comparable in magnitude in both errorful and non-errorful learning, and the lack of a significant interaction indicates that the benefits of errorful generation and generative feedback are additive rather than multiplicative. More broadly, the present findings demonstrate that feedback processing methods can exert a substantial impact on learning outcomes that, in earlier experiments, approached the effect size magnitude found by Pashler et al. (2005) for correct answer feedback versus right-or-wrong feedback ($d \approx 1.0$). Moreover, Experiment 4 revealed that the benefit of rephrasing versus no rephrasing ($d = 0.55$) was comparable in magnitude to the benefit of errorful generation versus no errorful generation ($d = 0.63$). Overall, these results suggest that much of the errorful generation literature, in which corrective feedback is processing solely by reading it, is not capitalizing fully on the learning opportunities that such feedback presents. Indeed, in the two experiments that specifically established an effect of errorful generation versus no errorful generation (Experiments 1 and 4), generative feedback processing clearly enhanced performance in the errorful generation condition.

The power of generative feedback processing

The finding that both rephrasing and example generation yielded similar performance suggests that the advantage of generative feedback processing is not limited to either strategy, even when they require learners to produce qualitatively different types of information. Unlike copying, reading, or the keyword mnemonic, both generative strategies did not just afford learners the opportunity to study the correct definition. Instead, rephrasing and example generation required learners to use the provided definition as a basis for generating new content—that is, a paraphrased definition and a new sentence using the target word in a way that was consistent with its definition. It is likely that doing so (a) caused learners to process the correct definition more thoroughly (as a basis for the new content they would generate), and (b) led them to mentally manipulate or elaborate upon that definition (in order to transform it into a paraphrased form or translate it into an example), resulting in "deeper" and more substantial levels of learning. In contrast, engaging in non-generative but active strategies such as copying almost certainly did not require leaners to engage in more thorough processing of the definition or mentally manipulate it in any way.

The patterns observed across the four experiments are fully consistent with predictions of generative learning theory, with feedback processing strategies that required learners to produce new content being uniquely beneficial. Our findings also align with Fiorella and Mayer (2016) characterization of generative strategies—such as summarizing and self-explaining—that enhance learning by promoting active sense-making, as well as with ICAP theory, which predicts that constructive strategies such as rephrasing and example generation yield greater learning gains than active but non-constructive ones. In addition, the patterns are broadly consistent with prior findings on feedback processing, including the benefits of explanation reported by Park and Ahn (2022) and Siegler (1995), and are broadly consistent of findings for the generation effect. Notably, however, the null effect of copying in Experiment 2 diverges from Park and Ahn's finding of a copying advantage in a different feedback context (see also Pan, Hutter, et al., 2018). This discrepancy suggests that although active—but non-generative—strategies may sometimes outperform passive approaches, their benefits may be less robust or context-dependent. Returning to the tiered classification of feedback processing approaches outlined in the Introduction (see Table 1), the present results suggest that solely passive or active approaches are least effective, with elaborative approaches yielding improvements, and crucially, generative approaches (e.g., the results of Experiment 3) yielding the best overall performance.

Processing elaborative feedback

Our findings for a form of elaborative feedback, namely example-enhanced feedback—which yielded significant learning improvements in Experiment 1 ($d = 0.39$) but not in Experiment 2 ($d = 0.13$)—were inconclusive, implying that the benefit of providing examples for definition recall is modest or not reliable. That result contrasts with at least one study involving participants learning from definitions with examples (but not as corrective feedback), namely Finn et al. (2018), which showed improved recall when examples were studied. However, it is broadly consistent with Rawson et al. (2015), who found that studying examples did not improve recall of previously studied definitions. Additionally, Butler et al. (2013) found that elaborative (explanation-based) feedback did not enhance performance on test items requiring recall of previously studied ideas, but did support performance on novel inference questions. These findings, along with a separate result from Rawson et al. (2015) showing that studying examples did enhance performance when a subsequent test required linking new examples to the appropriate concepts, suggest that the benefits of example-based feedback may depend on the demands of a subsequent test.

Interestingly, the elaborative advantage of contextual example learning seemed to be more fully leveraged when participants themselves generated the example sentences in Experiment 3. This result implies that simply reading feedback may not fully translate its intended benefits, at least for definition recall. Rather, learners' constructive effort may serve as a bridging link for a more pronounced and meaningful effect.

Additionally, the provision of mnemonics as feedback in the keyword mnemonic condition of Experiment 3 represented another form of elaborative feedback, but one that was also processed in an active and elaborative manner (i.e., participants were specifically instructed to do so and given practice, albeit limited, in doing so). Therefore, the effects of the feedback itself and the deliberate processing of that feedback must be considered together. Unlike example-enhanced feedback, the keyword mnemonic not only elaborates on to-be-learned content but also introduces separate details that are intentionally designed to make the material more memorable. Overall, the present results suggest that using the keyword mnemonic method as a form of feedback does not impair learning of definitions following errorful generation but may not enhance it to the same degree as generative approaches. It should also be noted that although the keyword mnemonic can be beneficial for vocabulary learning due to the use of associative imagery (Atkinson & Raugh, 1975; Pressley et al., 1982), its overall utility is considered by some learning researchers as moderate (e.g., Dunlosky et al., 2013).

Study limitations and future directions

Although the present study provides valuable insights into the effectiveness of generative feedback processing following errorful generation, it has several limitations that warrant further exploration. First, due to the single-session experimental design, it remains unclear whether the advantage of generative feedback processing over non-generative approaches persists in the long term. A study with a longer retention interval could address this issue. Second, the posttest assessment did not include alternative measures, such as transfer tasks, which might offer additional insights (Pan, Lovelett, et al., 2019). In particular, example-enhanced feedback may be more beneficial if the posttest involved a different task. Third, the inherent variability of generative responses—compared to the uniformity of non-generative tasks such as reading and copying—raises the possibility of individual differences in the efficacy of rephrasing and example generation (cf. Pan et al., 2025). Future research could involve qualitative analysis of how learners construct information and explore whether extensive pre-training influences the effectiveness of feedback processing (which could also apply to learning from example feedback or using the keyword mnemonic). The role of time-on-task—including potential differences in the relative balance of time allotted for guessing versus feedback processing across errorful and non-errorful generation conditions—could also be further explored. Finally, future work could further investigate how mechanisms purported to occur in the act of errorful generation—e.g., those implicated in semantic mediator theory, search set theory, the prediction error account, and recursive reminding theory—might interact with (generative) feedback processing. Current theoretical perspectives suggest that mechanisms involved in errorful generation and feedback processing occur in sequential, interconnected steps that contribute to learning benefits (Metcalfe, 2017), but how they interact remains to be fully understood.

Practical implications

Our findings have important implications for the effective use of errorful generation and feedback processing more generally. Although generative feedback and errorful generation have been individually shown to enhance learning in various educationally relevant circumstances (e.g., Siegler, 1995; Soderstrom & Bjork, 2023), errorful generation is still relatively uncommon in real-world educational practice (Pan et al., 2020) and further research is needed on its educational applications, highlighting the practical relevance of identifying strategies that maximize its benefits. Remarkably, based on exit survey data, learners in Experiments 2–3 collectively agreed that generative strategies such as rephrasing and example generation were the most effective, yet their preferences and usage frequency did not always align with those beliefs. A similar behavioral pattern was observed in Lahtinen et al. (1997), where a larger proportion of participants chose to utilize non-generative (e.g., underlining, notetaking) over generative strategies (e.g., summarizing, concept mapping), even though the latter methods yielded higher final test scores. These discrepancies point to gaps between learners' metacognitive understanding and their actual learning behaviors, which future research could explore (along with attendant cultural factors such as experience with feedback processing strategies in formal schooling).

The present results further indicate that when errorful generation is employed, pairing it with generative feedback processing yields additive benefits, providing clear guidance for maximizing learning outcomes in applied settings. More broadly, they also raise the prospect that real-world learning outcomes could be enhanced by adopting more effective feedback processing methods, even in educational settings where actively processing corrective feedback is common. Ultimately, the potency of errorful generation as a learning technique appears to depend not only on providing corrective feedback but also on how that feedback is processed, with generative engagement of feedback serving as a key mechanism that adds to its mnemonic benefits.

Acknowledgments

The authors thank Tan Jia An Valerie and Oh Wen Xi Megan for assistance with data collection. This work was supported by a Thesis Support Fund grant awarded to Thet Hnin Khin Khin and a Faculty of Arts & Social Sciences grant awarded to Steven C. Pan.

Disclosure statement

The authors report there are no competing interests to declare.

ORCID

Thet Hnin Khin Khin http://orcid.org/0009-0006-2911-2841
Steven C. Pan http://orcid.org/0000-0001-9080-5651

Data availability statement

The data set for the study is archived at the Open Science Framework (OSF) at: https://osf.io/vq82h/?view_only=c0f08023e6f14bc390d90d34621aec7f.

References

Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning & Memory*, *1*(2), 126–133. https://doi.org/10.1037/0278-7393.1.2.126

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/bf03193441

Brod, G. (2021). Predicting as a learning strategy. *Psychonomic Bulletin & Review*, *28*(6), 1839–1847. https://doi.org/10.3758/s13423-021-01904-1

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290–298. https://doi.org/10.1037/a0031026

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Collins Online Dictionary (2025). *Collins online dictionary*. https://www.collinsdictionary.com/

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 669–675. https://doi.org/10.1037/0278-7393.15.4.669

Cummings, E. L., Reeb, A., & McDaniel, M. A. (2023). Do not forget the keyword method: Learning educational content with arbitrary associations. *Journal of Applied Research in Memory and Cognition*, *12*(1), 70–81. https://doi.org/10.1037/mac0000031

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

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

Finn, B., Thomas, R., & Rawson, K. A. (2018). Learning more from feedback: Elaborating feedback with examples enhances concept learning. *Learning and Instruction*, *54*, 104–113. https://doi.org/10.1016/j.learninstruc.2017.08.007

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., Alford, L. K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, M., Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … Motz, B. A. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, *4*(3), 1-24. https://doi.org/10.1177/25152459211027575

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

Gupta, M. W., Pan, S. C., & Rickard, T. C. (2024). Interaction between the testing and forward testing effects in the case of cued-recall: Implications for theory, individual difference studies, and application. *Journal of Memory and Language*, *134*, 104476. https://doi.org/10.1016/j.jml.2023.104476

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(1), 290–296. https://doi.org/10.1037/a0028468

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(3), 253–271. https://doi.org/10.1177/1745691611406

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

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition*, *41*(5), 625–637. https://doi.org/10.3758/s13421-013-0298-5

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731–746. https://doi.org/10.1016/j.jml.2011.12.008

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., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning & Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Kozato, A., Shikino, K., Matsuyama, Y., Hayashi, M., Kondo, S., Uchida, S., Stanyon, M., & Ito, S. (2023). A qualitative study examining the critical differences in the experience of and response to formative feedback by undergraduate medical students in Japan and the UK. *BMC Medical Education*, *23*(1), 408. https://doi.org/10.1186/s12909-023-04257-6

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308. https://doi.org/10.1007/BF01320096

Lahtinen, V., Lonka, K., & Lindblom-Ylänne, S. (1997). Spontaneous study strategies and the quality of knowledge construction. *British Journal of Educational Psychology*, *67*(1), 13–24. https://doi.org/10.1111/j.2044-8279.1997.tb01223.x

Lee, I. (2013). Research into practice: Written corrective feedback. *Language Teaching*, *46*(1), 108–119. https://doi.org/10.1017/S0261444812000390

Levin, J. R. (1988). Elaboration-based learning strategies: Powerful theory = powerful application. *Contemporary Educational Psychology*, *13*(3), 191–205. https://doi.org/10.1016/0361-476X(88)90020-3

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

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

Oxford University Press (2025). *Oxford english dictionary*. https://www.oed.com/

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., & Rickard, T. C. (2018). In search of transfer following cued recall practice: The case of biology concepts. Applied Cognitive Psychology, 33(4), 629–645. https://doi.org/10.1002/acp.3506

Pan, S. C., Lovelett, J., Stoeckenius, D., & Rickard, T. C. (2019). Conditions of highly specific learning through cued recall. *Psychonomic Bulletin & Review*, *26*(2), 634–640. https://doi.org/10.3758/s13423-019-01593-x

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), 97. https://doi.org/10.1007/s10648-023-09814-5

Pan, S. C., Sana, F., Samani, J., Cooke, J., & Kim, J. A. (2020). Learning from errors: Students' and instructors' practices, attitudes, and beliefs. *Memory (Hove, England)*, *28*(9), 1105–1122. https://doi.org/10.1080/09658211.2020.1815790

Pan, S. C., Yu, L., Wong, M. J., Selvarajan, G., & Teo, A. Z. (2025). Do individual differences in working memory capacity, episodic memory ability, or fluid intelligence moderate the pretesting effect? *Journal of Memory and Language*, *142*, 104608. https://doi.org/10.1016/j.jml.2025.104608

Park, J.-H., & Ahn, S. (2022). L2 learners' cognitive and behavioral engagement with written corrective feedback. *English Teaching*, *77*(3), 133–152. https://doi.org/10.15858/engtea.77.3.202209.133

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Permut, S., Fisher, M., & Oppenheimer, D. M. (2019). Taskmaster: A tool for determining when subjects are on task. *Advances in Methods and Practices in Psychological Science*, *2*(2), 188–196. https://doi.org/10.1177/2515245919838479

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., Levin, J. R., & Delaney, H. D. (1982). The mnemonic keyword method. *Review of Educational Research*, *52*(1), 61–91. https://doi.org/10.3102/00346543052001061

Qu, K., Liu, T., Qiao, Y., & Wang, P. (2024). The facilitative effect of the keyword mnemonic on L2 vocabulary retrieval practice. *Heliyon*, *10*(3), e25212. https://doi.org/10.1016/j.heliyon.2024.e25212

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review*, *27*(3), 483–504. https://doi.org/10.1007/s10648-014-9273-3

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

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

Sidhu, G. K., & Fook, C. Y. (2010). Organization of homework: Malaysian teachers' practices and perspectives. *Journal of International Studies*, *13*, 63–78.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology*, *28*(3), 225–273. https://doi.org/10.1006/cogp.1995.1006

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Soderstrom, N. C., & Bjork, E. L. (2023). Pretesting enhances learning in the classroom. *Educational Psychology Review*, *35*(3), 88. https://doi.org/10.1007/s10648-023-09805-6

Stephens, M. (2020). Transferring literacy and subject knowledge between disparate educational systems. In M. L. Cook & L. G. Kittaka (Eds.), *Intercultural families and schooling in Japan: Experiences, issues, and challenges* (pp. 118–148). Candlin & Mynard. https://doi.org/10.47908/12/6

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Bureau of Publications, Teachers Co.

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition*, *41*(1), 1–15. https://doi.org/10.3758/s13421-012-0246-9