

EMPIRICAL ARTICLE

User-Generated Digital Flashcards Yield Better Learning Than Premade Flashcards

Steven C. Pan^{1, 2}, Inez Zung², Megan N. Imundo³, Xuxin Zhang³, and Yunning Qiu³
¹ Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Singapore
² Department of Psychology, University of California, San Diego, United States
³ Department of Psychology, University of California, Los Angeles, United States







Digital flashcard users typically must choose between creating their own flashcard content or using freely available flashcard sets. The latter is more convenient and saves time, but is it more effective for learning? We conducted six experiments, each involving the use of *user-generated* or *premade* flashcards to learn material drawn from educational text passages, followed by a 48-hr delayed criterial test. Different approaches to generating content and variations in the quality of premade content were also examined. Across experiments, user-generated flashcards improved memory relative to premade flashcards (an estimated advantage of $d = 0.45$, 95% CI [0.25, 0.66]), and in most cases, enhanced performance on application questions (an estimated advantage of $d = 0.29$, 95% CI [0.12, 0.45]). These results suggest that generating one's own flashcards enables productive learning processes that enhance memory and comprehension. Accordingly, digital flashcard users may benefit from eschewing premade versions in favor of making their own.

General Audience Summary

When using digital flashcards—which rank among the most popular e-learning tools available today—there is typically the option of making one's own flashcards (i.e., by manually adding content) or using flashcards that have already been made by somebody else. Currently, the latter option, which is also known as premade flashcards, is more popular than the former option, which is also known as user-generated flashcards. Using premade rather than user-generated flashcards is more convenient, saves time, and takes advantage of the millions of premade flashcards sets that are freely available online. However, with premade flashcards, users miss out on learning experiences that might occur when making one's own flashcards, and moreover, the quality of premade flashcards cannot be guaranteed. In this study, we investigated the learning of facts and concepts from premade versus user-generated flashcards. In five out of six experiments, using user-generated flashcards improved learning relative to using premade flashcards. These benefits were especially pronounced for flashcards made via paraphrasing or copying-and-pasting materials and were observed relative to premade flashcards of high and low quality. Thus, making user-generated flashcards can trigger productive learning processes. Given a fixed amount of time, adding content to digital flashcards prior to using them is potentially more beneficial for learning than using flashcards made by someone else.

Keywords: digital and computer flashcards, online learning technologies, retrieval practice, generative learning, Quizlet

Supplemental materials: <https://doi.org/10.1037/mac0000083.supp>

Steven C. Pan  <https://orcid.org/0000-0001-9080-5651>
 Inez Zung  <https://orcid.org/0000-0002-0947-2309>
 Megan N. Imundo  <https://orcid.org/0000-0003-4599-4777>
 Yunning Qiu  <https://orcid.org/0000-0002-6099-4612>

The authors thank Michelle Kaku for assistance with running the experiments, Guo Hengyi for scoring assistance, Wanxin Xie for programming suggestions, Erik Brockbank for data analysis consultation, and lab members for assistance with pilot testing. They also thank Andrew Butler for providing

useful suggestions (including with respect to the theoretical interpretation of the results), and Shana Carpenter and Tim Rickard for helpful comments on an earlier version of this article. The authors have no conflict of interest to disclose.

Steven C. Pan played a lead role in conceptualization, formal analysis, investigation, methodology and writing of original draft. Inez Zung played a supporting role in conceptualization, data curation, formal analysis, methodology, and writing of review and editing. Megan N. Imundo played a supporting role in conceptualization, methodology, and writing of review and editing. Xuxin Zhang played a lead role in formal analysis

continued

Today, millions of students use *digital flashcards* (Glottzbach, 2019). Also called computer flashcards, electronic flashcards, or virtual flashcards, digital flashcards duplicate the functions of paper flashcards, including the capacity to engage in retrieval practice (self-testing), plus offer extra features, including the use of freely available flashcard sets, or *premade flashcards*, that address virtually every conceivable topic. For instance, the website Quizlet offers over 500 million premade flashcard sets (Quizlet, 2022). These sets are commonly created by students, publishers, and even instructors.

As an alternative to premade flashcards, students might manually add content to flashcards. These *user-generated flashcards* may involve word-for-word transcription (e.g., of a definition or fact), copying-and-pasting, adding content in one's own words, and other methods. Given the time and effort involved, however, it is unsurprising that user-generated flashcards are less popular. Indeed, a recent survey found that 56% of U.S. undergraduate students prefer premade over user-generated flashcards, with convenience and saving time as common reasons (Zung et al., 2022; see also Green & Bailey, 2010). In contrast, among the 44% that favored user-generated flashcards, reasons included content control, greater accuracy, higher quality, and intriguingly, the belief that creating flashcards benefits learning.

Prior Research on Premade Versus User-Generated Flashcards

Some researchers have speculated that user-generated flashcards confer learning benefits that premade flashcards do not. Dodigovic (2013) and Wilkinson (2020b), for example, theorized that creating flashcards increases depth of processing (Craik & Lockhart, 1972), whereas Cihon et al. (2012) suggested that the increased exposure to course materials that might result benefits learning. The extant research comparing premade versus user-generated flashcards (see Table 1), however, has yielded inconsistent results.

Dodigovic (2013), Sage et al. (2019), and Wilkinson (2020a; see also Wilkinson, 2020b) had undergraduate students use premade or user-generated flashcards to learn vocabulary words, then take a recall test. The generated content varied from synonyms to definitions and example sentences. On an immediate test, performance was higher in the premade condition (Sage et al., 2019; Wilkinson, 2020a). On a test after several weeks of flashcard use, however, results ranged from a premade flashcard advantage (Dodigovic, 2013) to a user-generated advantage (Wilkinson, 2020b). Results involving more complex materials (e.g., anthropology texts; psychology course content) have also been mixed. Lin et al. (2018; Experiment 1) found no significant differences between flashcard types as measured on a 20-min delayed test, whereas Cihon et al. (2012) reported inconsistent results across two experiments as measured on weekly unit quizzes.

Although the evidence to date suggests no clear advantage for user-generated or premade flashcards, design differences across

studies—including methods for generating and practicing with flashcards (e.g., self-testing, studying, or both), controls for time on task (or lack thereof), retention interval, to-be-learned materials, and the content in the premade versus user-generated conditions—complicate interpretation. Moreover, there does not appear to be a single design feature that is responsible for the disparate results. For instance, Lin et al. (2018) suggested that their study instructions may have been suboptimal, whereas in Cihon et al. (2012), only the premade condition studied the exact information to be tested, thus leaving the user-generated condition at a disadvantage.

Does Generating Flashcard Content Elicit Productive Learning Processes?


Several prominent theories and related findings imply that user-generated flashcards may yield better learning than premade versions. Based on levels of processing theory (Craik & Lockhart, 1972), creating flashcards may improve memory if doing so involves attending to semantic or contextual details (Dodigovic, 2013; Wilkinson, 2020a) or otherwise mentally manipulating information (Nation, 2001). In addition, generative learning theories suggest that “active” learning activities that require selecting, organizing, integrating, or producing material, any of which might occur with user-generated flashcards, can enhance learning over more “passive” methods (Chi, 2009; Fiorella & Mayer, 2016; see also Foos et al., 1994; Pan et al., 2021). Creating flashcards may also yield a generation effect, wherein memory is improved for information that is mentally produced rather than read (Slamecka & Graf, 1978; see also Bertsch et al., 2007; Crutcher & Healy, 1989; Foos et al., 1994).


Conversely, there exist reasons to doubt the efficacy of user-generated flashcards. For instance, Sage et al. (2019) observed that user-generated flashcards are disadvantaged in terms of time available for practice. If learners have a fixed period to use premade flashcards or make and use user-generated flashcards, then that entire time can be devoted to practicing in the case of premade flashcards, whereas it must be divided between creating and practicing in the case of user-generated flashcards. It is also well established that flashcard users benefit from repeated learning opportunities that are spaced apart in time (i.e., distributed practice; Kornell, 2009; Wissman et al., 2012) and involve retrieval practice with correct answer feedback (Glover, 1989; Kulhavy & Stock, 1989), as can occur when flashcards are used repeatedly for self-testing and “flipped over” to check answers. With premade flashcards, users can immediately engage in distributed practice with retrieval practice and feedback, whereas with user-generated flashcards, there are fewer opportunities to do so. It has further been suggested that generating original content may not always yield sufficient semantic elaboration to improve memory (Sage et al., 2019).


and a supporting role in data curation. Yunning Qiu played a lead role in software.

Data and materials for this study are archived at the Open Science Framework at <https://osf.io/k9q8t/>.

Links to the data and materials, which are available for download, as well as the preregistered design and analysis plans, are provided in this article. Deviations from the preregistered analysis plan are also indicated in the article.

 The data are available at <https://osf.io/k9q8t/>.

 The experimental materials are available at <https://osf.io/k9q8t/>.

 The preregistered design (transparent changes notation) is available at https://aspredicted.org/BYR_7YG; https://aspredicted.org/NK5_P5D.

Correspondence concerning this article should be addressed to Steven C. Pan, Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, 9 Arts Link, Singapore 117572, Singapore. Email: scp@nus.edu.sg

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 1
Studies of User-Generated Versus Premade Flashcards

Reference	Materials	User-generated flashcard instructions	Premade flashcard content and source	Practice method	Controlled time-on-task	Retention interval	Result
Cihon et al. (2012; Experiment 1)	Terms, definitions (psychology)	Choose and then write terms and definitions	Terms and definitions; from instructor	Retrieval practice	No	<1 week	Premade advantage
Cihon et al. (2012; Experiment 2)	Terms, definitions (psychology)	Choose and then write terms and definitions	Terms and definitions; from instructor	Retrieval practice	No	<1 week	User-generated advantage ^a
Dodigovic (2013)	Vocabulary (academic)	Unspecified (at user's discretion)	Semantic, phonetic, contextual, and other details; from instructor	Retrieval practice or study	No	≤2 months	Premade advantage
Lin et al. (2018; Experiment 1)	Terms, definitions (anthropology)	Fill out cards focusing on details	Terms and definitions; from textbook publisher	Retrieval practice or study	No ^b	20 min	No clear advantage
Lin et al. (2018; Experiment 1)	Terms, definitions (anthropology)	Fill out cards focusing on concepts	Terms and definitions; from textbook publisher	Retrieval practice or study	No ^b	20 min	No clear advantage
Sage et al. (2019)	Vocabulary (GRE)	Add synonym word	Words and synonyms; from instructor	Retrieval practice or study	Yes	Immediate	Premade advantage
Wilkinson (2020a; Experiment 2)	Vocabulary (TOEFL)	Transcribe word, definition, example sentence	Word, definition, example sentence; from instructor	Retrieval practice or study	Yes	Immediate	Premade advantage
Wilkinson (2020a; Experiment 2)	Vocabulary (TOEFL)	Transcribe word, definition, example sentence	Word, definition, example sentence; from instructor	Retrieval practice or study	No	≤8 weeks	No clear advantage
Wilkinson (2020b; Experiment 2)	Vocabulary (TOEFL)	Generate own definition, example sentence	Word, definition, example sentence; from instructor	Retrieval practice or study	No	≤8 weeks	User-generated advantage

Note. GRE = Graduate Record Examination; TOEFL = Test of English as a Foreign Language.

^a Comparison was between a mix of user-generated and premade flashcards versus premade flashcards alone. ^b Including study time as a covariate did not change the observed patterns. All studies except for Dodigovic (2013) and Sage et al. (2019) relied solely on paper flashcards, and all studies except for Lin et al. (2018) and Sage et al. occurred in the context of an actual undergraduate course.

The literature on question generation, in which learners devise practice questions after viewing a text or lecture, illustrates potential limits on the pedagogical benefits of creating content. In several studies, generating questions has yielded test performance that is better than a restudy condition, but not a retrieval practice-only condition (e.g., Ebersbach et al., 2020; Weinstein et al., 2010; see also Hoogerheide et al., 2019). Potential explanations include the more time-consuming nature of generating questions, reduced learning efficiency, processing of extraneous information, and the need for specialized training beforehand (for discussions, see Bae et al., 2019; Bugg & McDaniel, 2012; Davey & McBride, 1986).

Overall, the literature to date suggests potential benefits and costs of generating flashcard content. Some accounts make specific predictions about the cognitive processes involved. There is, however, no consensus on the relative efficacy of user-generated versus premade flashcards, and it is unclear whether patterns observed in the question generation literature also apply to generating flashcard content. Accordingly, before strong conclusions can be made, further studies with more robust experimental controls are needed.

The Present Study

The present study entailed six experiments. In each experiment, participants read two text passages. After reading a given passage, they spent 20–25 min using premade flashcards or creating and then using user-generated flashcards to practice key terms and concepts from the passage, as digital flashcard users commonly focus on terms and concepts (Zung et al., 2022). Crucially, time on task was controlled, to-be-learned information in each condition was identified for each participant, and there were precise instructions regarding content generation and method of practice. We measured learning on a 48-hr delayed criterial test featuring definition and application questions (measuring memory and transfer of learning, respectively). The delayed test addressed the durability of learning over time and was used instead of an immediate test given the finding that many suboptimal learning techniques (e.g., cramming) yield comparable or even better performance than more effective methods (e.g., distributed practice) on an immediate test, but not on a delayed test (Bjork, 1994).

Across experiments, the most common methods of generating flashcard content according to survey research (Zung et al., 2022) were investigated (Experiment 1: word-by-word transcription; Experiment 2: copying-and-pasting; Experiments 3A, 4A, and 4B: paraphrasing; Experiment 3B: generating examples). These methods varied in apparent depth of processing and engagement in generative learning processes. In Experiments 1–3B, the premade flashcards featured content drawn verbatim from relevant text passages, whereas in the final experiments, the premade flashcards featured high-quality (Experiment 4A) or low-quality (Experiment 4B) content.

Ultimately, each experiment addressed a practical issue facing many students: Given a fixed amount of time, is it better for learning to create digital flashcards before using them or directly use flashcards made by someone else? That issue was investigated in a manner that manipulated flashcard type while keeping other factors carefully controlled or constant.

Experiment 1

Experiment 1 compared premade digital flashcards versus user-generated versions created via *word-for-word transcription* of to-be-learned content.

Method

All experiments in this study involved two sessions. In the first session, participants read a text passage, used digital flashcards to practice content from that passage, and then repeated the procedure with another passage. One passage each was assigned to the user-generated and the premade conditions, respectively. The assignment of passage to flashcard condition, and passage/condition order, was counterbalanced over participants. Two days later, participants completed a second session involving a criterial test on content from both text passages.

All experiments were programmed using the open-source, HTML- and JavaScript-based platform collector (Garcia & Kornell, 2014). Participants completed each experiment online, using an internet browser, and with a computing device of their choice.

Participants

The target sample size for all experiments, 47, was based on an a priori power analysis conducted in G*Power (Faul et al., 2007) involving a two-tailed, one-sample *t* test with the assumptions of a medium effect size of Cohen's $d = 0.42$ (i.e., the effect size difference for word recall with user-generated vs. premade flashcards in Sage et al., 2019), $\alpha = 0.05$, and 80% power. That power analysis was conducted based on our intention to compare a user-generated versus a premade flashcard condition, manipulated within-participants. For Experiment 1, we recruited 62 undergraduate students from the participant pool at a large U.S. public university in exchange for course credit. Data were excluded from three participants that had technical difficulties and two participants that did not follow instructions (i.e., left at least one flashcard blank or did not practice each flashcard at least once), thus leaving 57 participants ($M_{\text{age}} = 20.6$ years, 74% female) in the final sample. All experiments in the study were declared as exempt from review by the university's research ethics committee, and all participants gave informed consent.

Design

All experiments featured a within-participants design wherein each participant experienced both levels of flashcard condition (user-generated vs. premade).

Materials

The materials included two educational text passages (“expressionist art,” “ancient Rome”) adapted from Lippmann et al. (2013) and Magreehan (2016). Both text passages were just over 500 words in length, contained five to six short paragraphs covering different subtopics, and had a Flesch–Kincaid readability score of 15–16 (as measured using readability tools from online-utility.org). Within each passage, there were 10 italicized key terms (e.g., “metaphysical painting”), each with a corresponding one-sentence definition (e.g., “a style of painting using representational but incongruous imagery to produce disquieting effects on the viewer”). The learning

objective in each experiment was to master the meanings of those key terms, each of which represented a fact or concept. Moreover, in Experiment 1, participants practiced with the same exact materials regardless of whether a passage was assigned to the user-generated or premade condition (given word-for-word transcription in the user-generated condition).

For each of the 10 key terms, we developed two types of multiple-choice criterial test questions, each with four possible answer choices (40 questions in total across both passages). The two types were as follows: definition questions, which assessed memory for what the key term meant, and application questions, which presented new information (e.g., a new example) and required participants to relate what they had learned to it (e.g., for metaphysical painting, “Giorgio de Chirico was a metaphysical painter. Which of the following likely describes one of his metaphysical works, *The Disquieting Muses* (1916)?” for which the best answer was “Chirico uses imagery of mannequins set in a claustrophobic space, evoking a sense of irony and enigma and distorting perspective”).

A pilot test involving 10 undergraduate students (who read the passages and then attempted to answer all 40 questions) confirmed that the two sets of materials were comparable in difficulty (mean performance of 0.63 and 0.61 for the “expressionist art” and “ancient Rome” test questions, respectively).

The key terms are listed in Appendix A; all materials are archived at the Open Science Framework (Pan et al., 2022b) and accessible at <https://osf.io/k9q8r/>.

Procedure

Session 1. Upon signing up, participants received the URL to access the first session. The sequence of events for the two flashcard conditions is depicted in Figure 1. Session 1 lasted approximately 1 hr.

Premade Flashcards Condition. Participants first read a statement noting that a set of flashcards had already been prepared for them to use, with each flashcard addressing one of 10 key terms from the passage. Below that statement, the instructions stated:

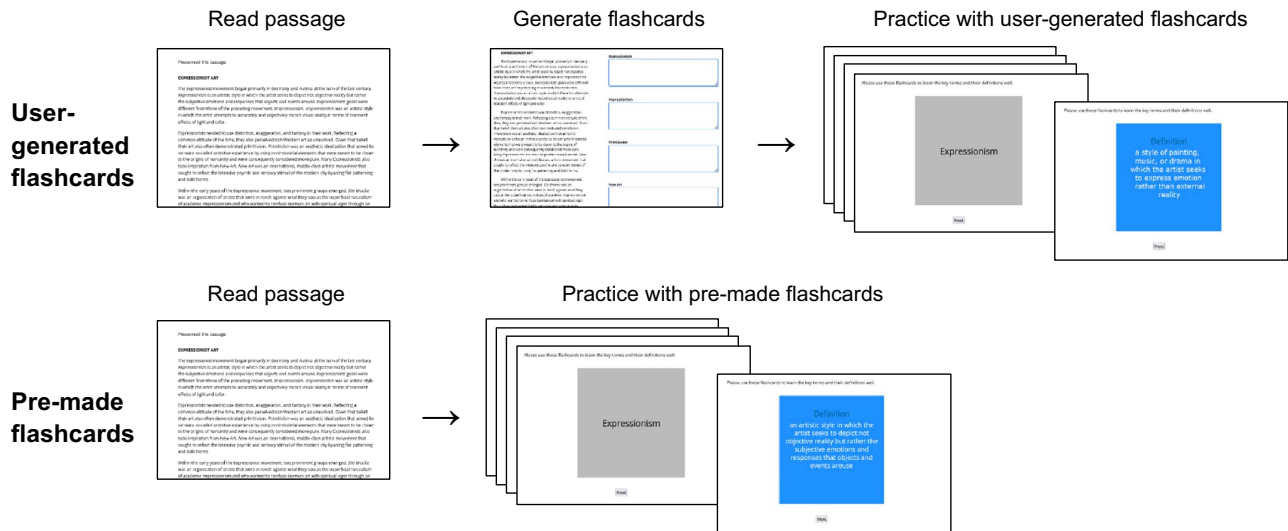
Please learn the content on the flashcards, focusing on the definitions. We suggest that you quiz yourself by trying to recall the definition from memory before clicking to reveal the answer. You will be able to go through the flashcards as many times as you wish in the allotted time.

Next, they used the flashcards for the full 20-min period.

User-Generated Flashcards Condition. Participants first read a statement noting that they would see a screen with 10 key terms listed, each with an accompanying textbox, and the text passage that they had just read. For each key term, they were to scan the passage for the relevant definition, then transcribe it, word-for-word, into the relevant textbox. Afterward, they would practice with their user-generated flashcards (and were instructed to do so just as in the premade condition). The overall amount of time was constant in both conditions, so the time allotted for practicing was 20 min minus the time spent making the flashcards.

Practicing With Digital Flashcards. The flashcard interface displayed one flashcard at a time at the center of the browser window (see Figure 1). The “front” side of the flashcard was shown by default and featured a light gray square within which the key term was displayed. Hovering over it triggered a “flipping” animation that displayed the “back” of the flashcard, which consisted of a blue square containing a premade (i.e., already copied from the passage) or user-generated (i.e., transcribed by the participant) definition, depending on condition. That side remained visible until participants hovered away, at which point the flashcard would “flip” back to the front side. Above each flashcard was a reminder of the instructions (“Please use these flashcards to learn the key terms”).

Figure 1
User-Generated and Premade Digital Flashcard Learning Procedures



Note. In the *user-generated* and *premade* conditions of each experiment, participants first read an educational text passage for 5 min. Next, they used digital flashcards to practice key terms from the passage for 20 min (Experiments 1–2) or 25 min (Experiments 3A–4B). In the user-generated condition (upper row), participants added content to the flashcards prior to using them, whereas in the premade condition (lower row), participants used the provided flashcards for the entire 20–25 min period. For each card, clicking on the front (grey-colored) side would flip it over to reveal the reverse (blue-colored) side. A larger, more detailed version of this figure is available at <https://osf.io/6wfs9>. See the online article for the color version of this figure.

and their definitions well”) and a countdown timer indicating the number of seconds remaining. Below was a button marked “Next” which, when clicked, would advance the screen to the next flashcard in the deck. A video clip of the flashcard interface is available at this study’s Open Science Framework repository (Pan et al., 2022a) and accessible at <https://osf.io/vtacs/>.

In both conditions, participants could go through the flashcards as many times and at whatever pace they wished until the allotted time had elapsed; they could choose to flip or not flip each flashcard to check the correct answers. To avoid confounding effects of differential interitem spacing or large imbalances in exposure frequency between specific flashcards, the presentation order was fixed and “dropping” (discontinuing the use of a flashcard) or “starring” (selecting a flashcard for extra practice) functions were disabled. The ability to highlight and copy text was also disabled.

After Flashcard Practice. In both conditions, after practicing, participants answered a question about their flashcard use (i.e., the percentage of time that they had engaged in self-testing during flashcard practice) and two metacognitive questions (i.e., how well they had learned the definitions and predicting future test performance). Self-testing was defined as “Quiz[zing] yourself by attempting to recall the definition from memory before clicking to reveal the answer.” These questions addressed participants’ practice activities and associated metacognitive thoughts. The second text passage was preceded by a 5-min distractor task involving six personal preference questions (e.g., one’s favorite movies).

Session 2. Forty-eight hr after the first session, participants were emailed the URL for the second session, involving a criterial test with all 40 questions for both text passages, and given 24 hr to complete it. Participants saw each question one at a time and had unlimited time to answer. To avoid contaminating effects of exposure to the information included in the application questions on definition question performance, the questions were grouped such that all 10 definition questions for a passage were presented immediately prior to all 10 application questions for the same passage. The choice of passage assessed first, order of questions within each group of questions, and order of answer choices per question were randomized anew for each participant. The second session typically took 30 min.

Results

The results from all experiments are reported on data collapsed across text passages. Analyses conducted on data restricted to individual passages, not reported here, yielded patterns that did not substantially differ from that for the overall data. As previously

noted, data were analyzed from participants that followed all experiment instructions, including correctly transcribing content in the user-generated condition.

Duration and Amount of Practice

Descriptive statistics for the time spent practicing (and making) cards is included in Table 2. As they were able to use the full 20 min to practice, participants viewed each flashcard significantly more times (about 1.6 more repetitions per card) in the premade versus user-generated conditions, $t(56) = 3.00, p = .0040, d = 0.40$.

Use of Self-Testing and Metacognitive Ratings

Table 3 presents participants’ self-reported use of self-testing and metacognitive judgments for both flashcard conditions. In this and all subsequent experiments, participants reported engaging in comparable amounts of self-testing in both conditions and their metacognitive judgments were also similar. The metacognitive data will be revisited in the Discussion section.

Criterial Test Results

Separately for definition and application questions, we conducted a 2 (order: user-generated first vs. premade first) \times 2 (flashcard condition: user-generated vs. premade) within-participants analysis of variance (ANOVA) on participant-level mean criterial test scores (Note: Analyses involving order were included at the suggestion of a reviewer, and for 14 participants, data from one definition question were unanalyzable due to a programming error). In the ANOVA for definition questions, the main effect of order was not significant, $F(1, 55) = 0.036, p = .85, \eta_p^2 = 0.0066$, indicating that practicing with one flashcard type did not substantially influence learning from the other, subsequently presented, flashcard type. The main effect of flashcard condition was also not significant, $F(1, 55) = 0.91, p = .35, \eta_p^2 = 0.016$, indicating no advantage for either user-generated or premade flashcards, and the interaction between order and flashcard condition was not significant, $F(1, 55) = 0.030, p = .86, \eta_p^2 = 0.00055$.

The same patterns were obtained in the ANOVA for application questions. The main effect of order was not significant, $F(1, 55) = 0.73, p = .40, \eta_p^2 = 0.013$, nor was the main effect of flashcard condition, $F(1, 55) = 0.33, p = .57, \eta_p^2 = 0.0060$, or the interaction, $F(1, 55) = 0.16, p = .69, \eta_p^2 = 0.0029$. The critical finding from both ANOVAs is depicted in the top left panel of Figure 2, in which it is apparent that performance in the user-generated and premade conditions was statistically indistinguishable for both question

Table 2
Mean Duration and Amount of Practice (SD)

Experiment	User-generated flashcards			Premeade flashcards	
	Generating content, in min	Practice time, in min	Repetitions per card	Practice time, in min (fixed)	Repetitions per card
1	8.9 (2.9)	11.1 (2.9)	4.4 (3.7)	20.0	6.0 (3.5)
2	2.5 (1.2)	17.5 (1.2)	3.6 (2.3)	20.0	4.3 (2.4)
3A	9.9 (4.8)	15.1 (4.8)	5.0 (6.6)	25.0	6.2 (7.2)
3B	11.7 (4.5)	13.3 (4.5)	4.7 (3.4)	25.0	6.5 (3.6)
4A	12.5 (5.2)	12.5 (5.3)	4.9 (4.0)	25.0	8.6 (7.7)
4B	12.9 (5.7)	12.1 (5.7)	4.4 (2.4)	25.0	9.3 (8.5)

Table 3
Metacognitive Ratings and Use of Self-Testing, in Mean Percentages (SD)

Experiment	Judgment of learning		Predicted test performance		Use of self-testing with feedback	
	User-generated	Premade	User-generated	Premade	User-generated	Premade
1	73.6 (18.7)	71.0 (22.4)	56.9 (21.5)	56.6 (26.0)	69.6 (22.9)	67.5 (24.3)
2	68.8 (20.3)	66.3 (20.3)	50.7 (25.4)	51.1 (23.6)	59.7 (28.9)	57.4 (28.4)
3A	74.5 (18.6)	67.0 (21.9)	62.6 (19.4)	56.9 (21.3)	65.0 (27.4)	64.1 (26.2)
3B	74.2 (23.1)	70.0 (21.2)	62.9 (22.4)	57.2 (23.4)	66.5 (28.8)	66.6 (28.0)
4A	78.2 (22.9)	72.1 (19.3)	63.6 (27.0)	56.6 (25.0)	65.2 (24.4)	66.8 (19.8)
4B	76.5 (23.1)	78.7 (23.8)	62.6 (24.0)	61.6 (26.9)	64.7 (31.6)	60.4 (31.1)

types, although numerically slightly higher for the user-generated condition.

Experiment 2

In Experiment 1, having users create flashcards via word-for-word transcription did not enhance learning. That result raised the possibility that there is no consistent advantage of generating flashcards. Alternatively, transcription might not have elicited productive learning processes due to shallow depth of processing and/or being a relatively passive activity. Further, transcription required nearly half the allotted time. To address the time issue and investigate another common method of creating flashcards, in Experiment 2, we switched the user-generated task to *copying-and-pasting*, which we surmised would be less time-consuming to carry out.

Method

Participants

Sixty-two undergraduate students were recruited in the same manner as in the first experiment. Data were excluded from five participants that left at least one flashcard blank (i.e., did not copy-and-paste content correctly) or did not practice each flashcard at least once, two participants that had technical problems, and one participant that did not return to complete the second session, thus leaving 54 participants ($M_{\text{age}} = 20.5$ years, 73% female) in the final sample.

Design, Materials, and Procedure

All aspects of the design, materials, and procedure were unchanged except for the user-generated condition, in which we restored copy-paste functionality and directed participants to scan the accompanying text passage for the relevant definition, highlight it, and use the Ctrl + C or Command + C and Ctrl + V or Command + V keyboard shortcuts to copy and paste it into the relevant textbox.

Results

Duration and Amount of Practice

As detailed in Table 2, participants generated cards more quickly and had more time to practice in the user-generated condition than in Experiment 1. Participants still completed less repetitions per card, on average, in the user-generated condition versus the premade condition (approximately 0.7 fewer repetitions on average), $t(53) = 2.09$, $p = .041$, $d = 0.28$, but the mean difference was less than half that observed in the preceding experiment.

Criterion Test Results

Two ANOVAs analogous to those conducted for Experiment 1 were performed on criterion test scores. In the ANOVA for definition questions, the main effect of order was not significant, $F(1, 52) = 3.30$, $p = .075$, $\eta_p^2 = 0.060$, whereas the main effect of flashcard condition was significant, $F(1, 52) = 9.76$, $p = .0029$, $\eta_p^2 = 0.16$. There was also a significant order by flashcard condition interaction, $F(1, 52) = 5.70$, $p = .021$, $\eta_p^2 = 0.099$. In the ANOVA for application questions, the main effect of order was significant, $F(1, 52) = 4.37$, $p = .042$, $\eta_p^2 = 0.078$, as was the main effect of flashcard condition, $F(1, 52) = 4.89$, $p = .032$, $\eta_p^2 = 0.086$, whereas the order by flashcard condition interaction was not significant, $F(1, 52) = 0.77$, $p = .39$, $\eta_p^2 = 0.015$.

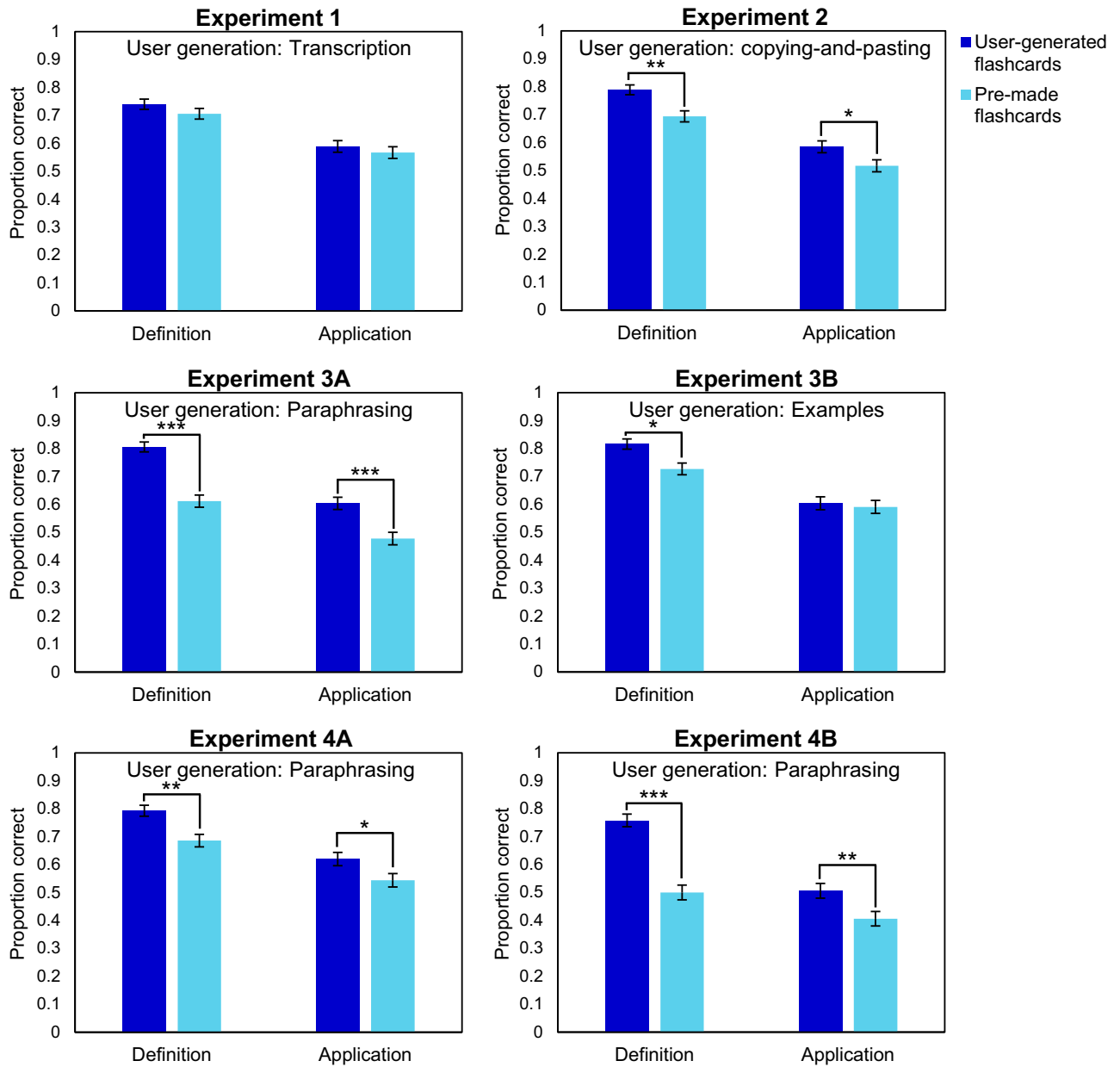
The finding of significant main effects of flashcard condition corresponds with examination of the top right panel of Figure 2, in which it is evident that, unlike Experiment 1, the user-generated condition outperformed the premade condition for both question types. Indeed, in two follow-up t tests, there was a significant user-generated advantage for definition questions, $t(53) = 2.99$, $p = .0042$, $d = 0.41$, and for application questions, $t(53) = 2.22$, $p = .031$, $d = 0.30$.

Uniquely in this experiment, the order of flashcard type during practice appeared to affect criterion test performance: Using user-generated flashcards before premade flashcards yielded numerically larger user-generated advantages for definition questions, on average, and lower average scores, overall, on application questions. Those effects, however, do not survive correction for multiple comparisons, which raise the possibility of a Type I error, and given that no corresponding patterns were observed in any other experiment, condition order is not discussed further.

Experiments 3A and 3B

Experiment 2 demonstrated that copying-and-pasting content onto flashcards, which took two-thirds less time than that required for transcription, can enhance learning. We will further consider both transcription and copying-and-pasting in the Discussion section. For Experiments 3A and 3B, we explored the possibility that creating original content—which might elicit benefits of generative learning activities and/or the generation effect—also improves learning (for related approaches, see Appleby, 2013; Senzaki et al., 2017). These experiments entailed two user-generated conditions: *paraphrasing* (Experiment 3A) and *generating examples* (Experiment 3B),

Figure 2
Delayed Critical Test Results for Experiments 1–4B



Note. Each panel displays the efficacy of user-generated versus pre-made digital flashcards as evident on a 48-hr delayed criterial test featuring definition (recall) and application (transfer) questions. Results are shown for the case of *user-generated* flashcards involving word-for-word transcription (Experiment 1); copying-and-pasting of content (Experiment 2); paraphrasing (Experiments 3A, 4A, 4B); or generation of an example sentence (Experiment 3B), relative to *premade* flashcards that featured the same wording as in the preceding text passage (Experiments 1–3B) or different wording that was of high quality (Experiment 4A) or low quality (Experiment 4B). Error bars represent standard error of the mean. See the online article for the color version of this figure. * $p < .05$. ** $p < .01$. *** $p < .001$.

which we compared against pre-made conditions identical to those used in the prior experiments.

Method

Experiments 3A and 3B were preregistered at <https://aspredicted.org/bx35n.pdf>.

Participants

One hundred seventeen undergraduate students were recruited in the same manner as in the preceding experiments. Data were excluded from seven participants that attempted the first session twice, 13 participants that left at least one flashcard blank or did not practice each flashcard at least once, and two participants that had

technical problems,¹ thus leaving 95 participants (Experiment 3A: $n = 50$, $M_{\text{age}} = 20.4$ years, 80% female; Experiment 3B: $n = 45$, $M_{\text{age}} = 21.6$ years, 84% female) in the final sample.

Design, Materials, and Procedure

All aspects of the design, materials, and procedure were identical to Experiment 1, except for the following changes. First, we conducted both experiments at the same time and with participants randomly assigned to either experiment. That design feature allowed us to analyze data from both experiments in an ANOVA with experiment as a between-participants factor. Second, because we expected that paraphrasing and generating examples would require more time, we increased the flashcard activity period by 5 min in the user-generated and premade flashcard conditions. Third, the user-generated flashcard task was modified as follows.

In Experiment 3A, participants were instructed as to come up with “an accurate and complete definition” of each key term, in their own words, and type it into the corresponding text box. A hypothetical example was presented for reference. In Experiment 3B, participants were instructed to develop and type, for each key term, an “example sentence that uses the key term in a scenario or situation of some kind,” with the sentence reflecting a plausible use of the term given how it was described and including “details about the term without simply being a definition of it” (thus requiring an understanding of the key term but disallowing paraphrasing). An example was also provided. Both conditions were inspired by the literature on generative learning activities (Brod, 2021; Fiorella & Mayer, 2016).

Scoring of User-Generated Flashcard Content

To evaluate the accuracy and completeness of user-generated content, we developed a scoring rubric wherein the correct definition of each key term was divided into between three and seven idea units. To be scored as fully accurate, a user-generated definition (Experiment 3A) had to include all the idea units and describe each correctly. The completeness of the examples in Experiment 3B was also scored according to the same rubric.

Two raters first independently scored a randomly selected 32% of all responses to evaluate the reliability of the rubric, and differences were adjudicated via discussion. As interrater agreement was reasonable (intraclass correlation coefficient = 0.77), the remaining data were scored by a single rater. Descriptive statistics calculated on participant-level average idea unit scores are reported in this article.

Results

The preregistration for Experiments 3A and 3B proposed multiple analyses to be performed on the criterial test data. In these and the next two experiments, after examination of the overall pattern of results, we performed a subset of those proposed analyses (which are described in the following section). The pairwise comparisons involving number of repetitions per card were not preregistered and should be regarded as exploratory.

Duration and Amount of Practice

As detailed in Table 2, in Experiment 3A, participants in the premade condition on average achieved about 1.2 more repetitions

per card than participants in the user-generated condition, $t(49) = 2.29$, $p = .026$, $d = 0.32$. The same was true in Experiment 3B, with participants in the premade condition achieving about 1.8 more repetitions per card, on average, than participants in the user-generated condition, $t(44) = 3.45$, $p = .0013$, $d = 0.51$. Both differences reflect the greater amount of time available for practicing in the premade condition.

Quality of User-Generated Flashcard Content

Scoring of the paraphrased definitions in Experiment 3A yielded a mean completeness rating (SD) of 71% (17%). Scoring of the definitional content of the example sentences in Experiment 3B yielded a mean completeness rating of 62% (19%). Examples of paraphrased definitions that participants created in Experiment 3A, which would play an additional role in the final two experiments, can be found in Appendix B.

Criterial Test Results

Separately for definition and application questions, we conducted a 2 (Experiment: 3A vs. 3B) \times 2 (order: user-generated first vs. premade first) \times 2 (flashcard condition: user-generated vs. premade) mixed-factors ANOVA on participant-level mean criterial test scores. In the analysis for definition questions, there was a significant main effect of flashcard condition, $F(1, 91) = 23.00$, $p < .0001$, $\eta_p^2 = 0.20$, whereas the main effect of experiment, main effect of order, and all other interactions were not significant ($ps > .13$). That analysis is reinforced by inspection of the middle panels of Figure 2, wherein it is apparent that the user-generated condition in both experiments exhibited a similar performance advantage over the premade condition on definition questions. In the analysis for application questions, there was a significant main effect of flashcard condition, $F(1, 91) = 9.41$, $p = .0029$, $\eta_p^2 = 0.094$, and a significant experiment by flashcard condition two-way interaction, $F(1, 91) = 5.66$, $p = .019$, $\eta_p^2 = 0.059$. The main effect of experiment, main effect of order, and all other interactions were not significant ($ps > .15$). Inspection of Figure 2 suggests that, in line with that analysis, there was a sizeable user-generated advantage for application questions in Experiment 3A, but not in Experiment 3B.

Overall, paraphrasing in the user-generated condition of Experiment 3A yielded better performance for both definition and application questions. Generating example sentences in the user-generated condition of Experiment 3B yielded better performance for definition questions only. Thus, although the user-generated condition consistently outperformed the premade condition, performance on definition questions benefited similarly from paraphrasing and generating of example sentences, whereas performance on application questions benefited more from paraphrasing than from generating example sentences.

¹ In the analyses reported here, we excluded such participants on the basis that they were unable to practice on all items (which produces an imbalance in the items per condition). Nevertheless, for each experiment, when we reanalyzed data with those participants included (and including performance on the relatively few items for which there was no practice), the overall patterns were unchanged.

Experiments 4A and 4B

The preceding experiments compared user-generated flashcards against premade flashcards featuring content drawn verbatim from source materials. As previously noted, however, doubts exist about the quality of premade flashcards (Zung et al., 2022). Accordingly, in Experiments 4A and 4B, we compared user-generated (paraphrased) flashcards versus premade flashcards containing separately created content that was of high quality (Experiment 4A) or low quality (Experiment 4B).

Method

Experiments 4A and 4B were preregistered at <https://aspredicted.org/b2t7w.pdf>.

Participants

One hundred two undergraduate students, recruited in the same manner as in the preceding experiments, participated. Data were excluded from 10 participants that did not complete the second session, 13 participants that left at least one flashcard blank or did not practice each flashcard at least once, and one participant that had technical problems, thus leaving 78 participants (Experiment 4A: $n = 42$, $M_{\text{age}} = 20.0$ years, 73% female; Experiment 4B: $n = 36$, $M_{\text{age}} = 20.8$ years, 70% female) in the final sample (although we exceeded sample size recruitment targets, and had comparable attrition rates, the total number of participants before exclusions was smaller than in prior experiments).

Design, Materials, Procedure, and Scoring

To create the materials for the premade conditions, we drew on the user-generated (i.e., paraphrased and/or generated) definitions from Experiment 3A, and for each key term, selected one of the highest scoring (at or near 100%) and one of the lowest scoring definitions (below 25%) from the manual scoring of those responses (with an additional requirement that such responses should not be an exact match to the original source materials and could not contain outright inaccuracies). Those responses (examples of which are shown in Appendix B) comprised the content for the high-quality and low-quality flashcards in the premade conditions of Experiments 4A and 4B, respectively.

All other aspects of the design, materials, procedures, and scoring methods in both experiments were identical to Experiment 3A, including the paraphrasing of definitions in the user-generated condition (which we chose because it appeared to be the most effective, at least numerically, among all investigated instantiations of user-generated flashcards). We also randomly assigned participants to Experiment 4A or 4B in the same manner as in the prior experiments.

Results

Duration and Amount of Practice

As detailed in Table 2, in Experiment 4A, participants in the premade condition achieved about 3.7 more repetitions per card, on average, than participants in the user-generated condition, $t(40) = 3.66$, $p < .001$, $d = 0.57$. In Experiment 4B, participants achieved

about 4.9 more repetitions per card, on average, in the premade condition than participants in the user-generated condition, $t(35) = 3.41$, $p = .0017$, $d = 0.57$. The sizeable disparity in repetitions may reflect faster reading of the simple language and/or brief construction of the premade content. (Note: Due to a technical malfunction, data on repetitions per card were not recorded for one participant, respectively, in both conditions in Experiment 4A.)

Quality of User-Generated Flashcard Content

Scoring of the paraphrased definitions in Experiments 4A and 4B yielded mean completeness ratings (SD) of 63% (16%) and 61% (21%), respectively.

Criterial Test Results

Just as with Experiments 3A and 3B, we conducted a 2 (experiment: 4A vs. 4B) \times 2 (order: user-generated first vs. premade first) \times 2 (flashcard condition: user-generated vs. premade) mixed-factors ANOVA on participant-level mean criterial test scores and separately for definition and application questions. In the analysis for definition questions, there was a significant main effect of experiment, $F(1, 74) = 8.13$, $p = .0057$, $\eta_p^2 = 0.099$, a significant main effect of flashcard condition, $F(1, 74) = 37.16$, $p < .00001$, $\eta_p^2 = 0.33$, and a significant experiment by flashcard condition two-way interaction, $F(1, 74) = 6.75$, $p = .011$, $\eta_p^2 = 0.084$. The main effect of order and all other interactions were not significant ($ps > .12$). That analysis is reinforced by inspection of the bottom panels of Figure 2, wherein it is apparent that overall performance on definition questions was higher in Experiment 4A (wherein the premade condition featured high-quality content) and that the user-generated advantage for definition questions was larger in Experiment 4B (wherein the premade condition featured low-quality content). In the analysis for application questions, there was a significant main effect of experiment, $F(1, 74) = 56.11$, $p < .00001$, $\eta_p^2 = 0.43$, and a significant main effect of flashcard condition, $F(1, 74) = 37.16$, $p < .001$, $\eta_p^2 = 0.17$. The main effect of order and all other interactions were not significant ($ps > .13$). In line with that analysis, inspection of Figure 2 reveals that performance on application questions was higher in Experiment 4A (wherein the premade condition featured high-quality content) than in Experiment 4B. Moreover, user-generated advantages for application questions of similar magnitude were observed in both experiments.

In summary, the user-generated condition exhibited better overall performance in both experiments, but in the case of definition questions, to a larger extent in Experiment 4B. Inspection of Figure 2 also indicates that performance was generally lower in Experiment 4B, evidently driven by worse performance in the premade condition (which was subject to the low-quality flashcard content). Overall, user-generated flashcards were more effective than premade flashcards of both high and low quality, but with a more pronounced advantage relative to premade flashcards of low quality and especially for the case of definition recall.

Analyses of Experiments 3A–4B Involving User-Generated Flashcard Quality

To address whether the quality of the user-generated flashcards in Experiments 3A–4B affected the observed results, we conducted

two sets of supplementary analyses. These analyses relied on participant-level average idea unit scores (where each participant received a single average quality score). We first addressed whether participants that did a better job making flashcards scored better on criterial test questions assessing content addressed by those flashcards. In Experiment 3A, there was a significant positive correlation between quality scores and criterial test performance in the user-generated condition for definition questions ($r = 0.30, p = .035$) and for application questions ($r = 0.37, p = .0084$). In the remaining experiments, however, no significant correlations were observed ($r_s \leq 0.23, p_s \geq .14$). Thus, although the quality of user-generated content was associated with better performance on user-generated test items in Experiment 3A, that relationship was much weaker, if it existed at all, in other experiments. Next, to address whether quality might have influenced the overall patterns of results—as might be the case if the user-generated advantage, or lack thereof, differed among individuals producing different quality flashcards—we conducted an analysis of covariance with quality as the covariate separately for all application and for all definition questions in each of Experiments 3A–4B. Quality was only significantly associated with performance in Experiment 3A ($p = .00058$) and including quality as a covariate did not alter the presence or absence of a user-generated advantage.

Internal Meta-Analyses of Experiments 1–4B

To derive an overall estimate of the learning advantage that is conferred by user-generated flashcards, we conducted two separate internal meta-analyses (Goh et al., 2016), one involving definition questions and the other involving application questions, using the criterial test data from each experiment. Both random-effects meta-analyses were performed using the *metafor* package

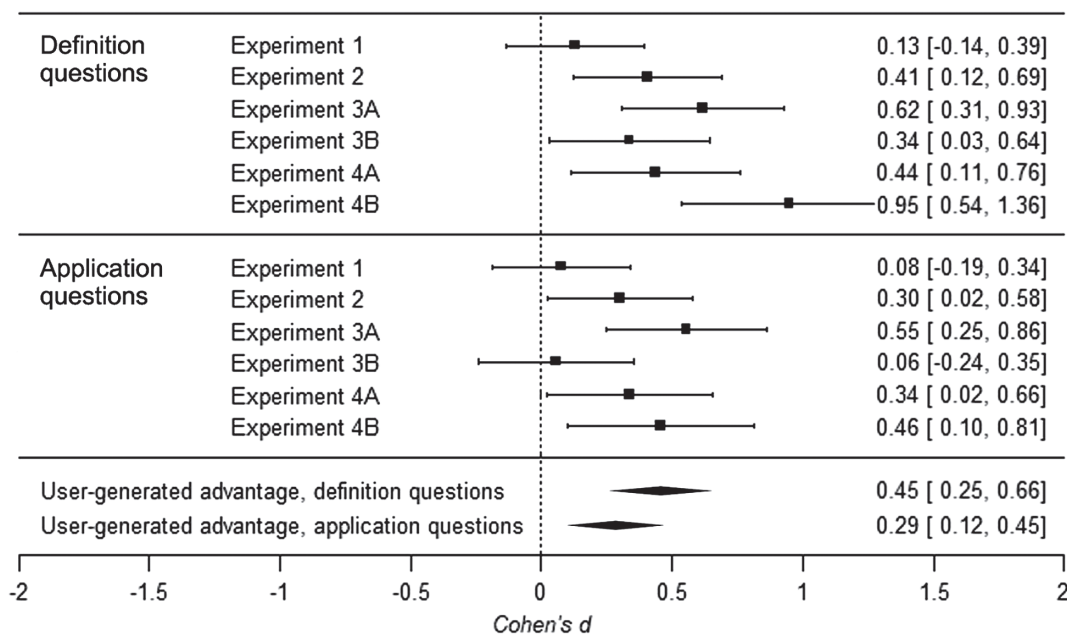
in R (Viechtbauer, 2010), effect sizes in terms of Cohen’s d (wherein each effect size represented the performance difference between the user-generated and premade conditions, with a positive d value reflecting a user-generated advantage and a negative d value reflecting a premade advantage), and the sampling variance for each effect size calculated according to Morris and DeShon (2002). Results are shown as a forest plot in Figure 3. Overall, user-generated digital flashcards yielded better performance than premade flashcards on definition questions by $d = 0.45, 95\% \text{ CI } [0.25, 0.66]$, and better performance than premade flashcards on application questions by $d = 0.29, 95\% \text{ CI } [0.12, 0.45]$.

Discussion

Across experiments, user-generated flashcards yielded significantly better delayed criterial test performance than premade flashcards. The user-generated advantages of $d_s = 0.45$ and 0.29 for definition and application questions, respectively, are educationally meaningful (Hattie, 2009; Kraft, 2020) and contrast with prior studies that did not include the same controls for time-on-task and methods of content generation and practice. The finding that generating content prior to retrieval practice enhanced learning over retrieval practice alone also contrasts with patterns observed in the question generation literature.

Different methods of generating content varied in effectiveness. Paraphrasing and copying-and-pasting enhanced overall performance, generating examples enhanced definition recall only, and word-for-word transcription was the least effective. In addition, when we manipulated the quality of premade flashcard content, a user-generated advantage still occurred, but to a larger extent relative to low-quality premade flashcards. Further, most participants did not exhibit a strong metacognitive awareness of the

Figure 3
Forest Plot of Effect Sizes (Cohen’s d) With 95% Confidence Intervals for the Relative Advantage of User-Generated Versus Premade Digital Flashcards on Criterial Test Performance in Experiments 1–4B



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.

benefits of generating flashcard content even after doing so, which resembles patterns observed with other learning techniques (Bjork et al., 2013; Kornell, 2009). For users that do endorse benefits of generating flashcards, however, this research provides compelling support.

Theoretical Implications

The present results rule out the hypothesis that the extra time available for study and practice with premade flashcards guarantees better learning (Sage et al., 2019). A user-generated advantage was repeatedly observed despite deficits in study time, practice time, and number of repetitions per flashcard. Experiments 4A and 4B further rule out the possibility that stimulus variability (for discussion, see Schmidt & Bjork, 1992), which was always present in the user-generated condition (i.e., passage vs. flashcard content), was a crucial factor. Such variability was also present in the premade conditions of those experiments. Other potential sources of the user-generated advantage include extra exposure to materials (Cihon et al., 2012) and learning processes that occur during content generation (Dodigovic, 2013). Given that extra exposure did not always improve learning, however, the latter possibility seems more likely. We next consider the learning processes that different methods of content generation may elicit.

The two “passive” methods, word-by-word transcription and copying-and-pasting, arguably are not generative learning activities and might not yield generation effects. Yet, copying-and-pasting enhanced learning, whereas transcription did not. That pattern likely stemmed from the more time-consuming, attention-demanding, and effortful nature of transcription (i.e., typing while maintaining focus on the letters to be typed), which prevented extensive reexamination and/or further processing of the text. The limited mnemonic value of transcription has been demonstrated in other circumstances; for instance, transcribing idioms does not enhance memory relative to studying (Stengers et al., 2016). In contrast, with copying-and-pasting, more cognitive resources were available for processing of text (for a related example involving summarization and copying-and-pasting, see Morgan et al., 2008).

The two “active” methods, paraphrasing and generating examples, arguably qualify as generative learning activities and may have yielded generation effects and/or extra processing of text. Yet, generating examples only enhanced recall, whereas paraphrasing enhanced recall and transfer. That result potentially stems from attention being almost exclusively focused on the text in the case of paraphrasing and divided between the text and prior knowledge in the case of generating examples (i.e., in order to devise novel examples). Although such divided attention might have led to integration of new and old information (Fiorella & Mayer, 2016), it may have instead reduced comprehension, which impacted transfer performance. Analogous extraneous processing has also been blamed for the limited efficacy of question generation (Hoogerheide et al., 2019).

The foregoing discussion is consistent with a candidate principle of flashcard learning: Generating flashcard content benefits learning by eliciting extra processing—that is, rereading, mental elaboration, depth of processing, or even improved attention—of to-be-learned information (for related theorizing, see (Cihon et al., 2012; Dodigovic, 2013; Wilkinson, 2020a). That improved learning

may occur during content generation (which, depending on experiment, required between 13% and 52% of the entire first session). Alternatively, or in conjunction, that extra processing may yield better learning during subsequent retrieval practice.

Regarding the latter scenario, it is possible that generating content increases the rate of successful recall during retrieval practice, which in turn improves learning. In that scenario, retrieval practice in the premade condition is less effective, at least initially, due to a low rate of retrieval success. Higher rates of successful recall during practice are indeed associated with larger retrieval practice effects (Pan & Rickard, 2018), although the relationship is less strong when correct answer feedback is provided (Rowland, 2014). In our view, it is conceivable that some methods of generating content, such as paraphrasing, are particularly effective at improving learners’ performance during subsequent retrieval practice, yielding better criterial test performance. If so, then user-generated flashcards should be especially helpful for learners lacking strong mastery of to-be-learned materials.

Finally, the inconsistent user-generated advantages in prior studies might be attributed to the method of content generation that was used, with instructions, learning materials, and flashcard content (see the effects of quality in the user-generated condition of Experiment 3A) as secondary factors. If the aforementioned principle of flashcard learning holds, however, then it should be possible to optimize a variety of content generation methods to yield productive learning processes, efficacious practice activities, and ultimately, a user-generated advantage.

Practical Implications

The present results suggest that the common practice of using freely available flashcard sets—which many learners do for convenience, despite concerns about quality (what Zung et al., 2022, called an “ease–accuracy trade-off”), and with greater frequency than premade paper flashcards—can impair learning efficacy. Accordingly, one of the chief selling points of many digital flashcard platforms, namely the millions of premade flashcard sets (including sets prepackaged with textbooks and other educational products), may not be as compelling as currently thought. Fortunately, the solution is quite simple: use flashcard-making features.

Before the present results can be generalized broadly, however, some caveats apply. It is important to emphasize that digital flashcards are commonly used in many ways—including different platforms, learning environments, and materials—and not under the same controlled circumstances as in the present experiments (cf. Cihon et al., 2012; Golding et al., 2012; Imundo et al., 2021; Reeser & Moon, 2018). Guidance for flashcard use is rare and the match between flashcard and exam content varies (e.g., strong alignment for instructor-provided flashcards and weak alignment for premade flashcards made by students lacking foreknowledge of the content to be tested). If instructor-provided flashcards preview an upcoming exam, then students would be well advised to use them.

Despite those caveats, our findings are directly applicable to a variety of common scenarios. For example, a student may have the option of downloading flashcards from students taking similar classes or using assigned course materials to create their own (i.e., the premade flashcard content is not inherently superior).

Given the same amount of time and the same methods of practice, user-generated flashcards are likely to be more beneficial.

Limitations and Future Research

Future research could address study limitations and different circumstances than investigated to date, including other platforms, learners, and to-be-learned materials. Features that were disabled for the purposes of experimental control, including shuffling and dropping functions (Sage et al., 2016), could be reenabled. It is not necessarily likely, however, that the use of any specific feature differs strongly between flashcard types. To explore the role of retrieval success, future studies could also require overt responses. Participants were not required to type out their recall attempts in the present experiments, yielding no direct measure of practice performance. Improvements in the quality of paraphrased definitions or examples might also be explored. Despite the lack of a consistently observed relationship between quality and learning, prior training or better instructions might improve the efficacy of such content generation. Further, the competitiveness of copying-and-pasting with paraphrasing, which presumably fosters different levels of processing yet yielded comparable learning benefits, remains to be fully explained.

An uninvestigated middle ground between pre-made versus user-generated digital flashcards also exists: modifying existing flashcard sets (Green & Bailey, 2010). Modifying flashcards may enhance learning, perhaps to a lesser extent than generating brand-new sets. Finally, given the continued evolution of flashcard technologies (e.g., Chen & Chan, 2019), further research stands to determine whether the benefits of generating flashcards persist into the future.

References

- Appleby, D. C. (2013). *A flashcard strategy to help students prepare for three types of multiple-choice questions commonly found on introductory psychology tests*. <http://www.apadiv2.org/Resources/Documents/otrp/resources/appleby13flashcard.pdf>
- Bae, C. L., Therriault, D. J., & Redifer, J. L. (2019). Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction, 60*, 206–214. <https://doi.org/10.1016/j.learninstruc.2017.12.008>
- 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>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*(1), 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>
- Brod, G. (2021). Generative learning: Which strategies for what age? *Educational Psychology Review, 33*, 1295–1318. <https://doi.org/10.1007/s10648-020-09571-9>
- Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology, 104*(4), 922–931. <https://doi.org/10.1037/a0028661>
- Chen, R. W., & Chan, K. K. (2019). Using augmented reality flashcards to learn vocabulary in early childhood education. *Journal of Educational Computing Research, 57*(7), 1812–1831. <https://doi.org/10.1177/0735633119854028>
- Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105. <https://doi.org/10.1111/j.1756-8765.2008.01005.x>
- Cihon, T. M., Sturtz, A. M., & Eshleman, J. (2012). The effects of instructor-provided or student-created flashcards with weekly, one-minute timings on unit quiz scores in introduction to applied behavior analysis courses. *European Journal of Behavior Analysis, 13*(1), 47–57. <https://doi.org/10.1080/15021149.2012.11434404>
- Craik, F. I., & 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](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>
- Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology, 78*(4), 256–262. <https://doi.org/10.1037/0022-0663.78.4.256>
- Dodigovic, M. (2013). Vocabulary learning with electronic flashcards: Teacher design vs. student design. *Voices in Asia Journal, 1*(1), 15–33.
- Ebersbach, M., Feierabend, M., & Nazari, K. B. B. (2020). Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning. *Applied Cognitive Psychology, 34*(3), 724–736. <https://doi.org/10.1002/acp.3639>
- 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>
- 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>
- Foos, P. W., Mora, J. J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology, 86*(4), 567–576. <https://doi.org/10.1037/0022-0663.86.4.567>
- Garcia, M. A., & Kornell, N. (2014). *Collector* [Software]. <https://github.com/gikeymarcia/Collector>
- Glotzbach, M. (2019, December 12). *Celebrating 2019 and quizlet's impact*. Quizlet.com. Retrieved November 12, 2021, from <https://quizlet.com/blog/2019-impact-report>
- Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392–399. <https://doi.org/10.1037/0022-0663.81.3.392>
- Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549. <https://doi.org/10.1111/spc3.12267>
- Golding, J. M., Wasarhaley, N. E., & Fletcher, B. (2012). The use of flashcards in an introduction to psychology class. *Teaching of Psychology, 39*(3), 199–202. <https://doi.org/10.1177/0098628312450436>
- Green, T., & Bailey, B. (2010). Digital flashcard tools. *Tech Trends, 54*(4), 16–18. <https://doi.org/10.1007/s11528-010-0415-2>
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hoogerheide, V., Staal, J., Schaap, L., & van Gog, T. (2019). Effects of study intention and generating multiple choice questions on expository text retention. *Learning and Instruction, 60*, 191–198. <https://doi.org/10.1016/j.learninstruc.2017.12.006>
- Imundo, M. N., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2021). Where and how to learn: The interactive benefits of contextual variation, restudying, and retrieval practice for learning. *Quarterly Journal of Experimental Psychology, 74*(3), 413–424. <https://doi.org/10.1177/1747021820968483>

- Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, 23(9), 1297–1317. <https://doi.org/10.1002/acp.1537>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- 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>
- Lin, C., McDaniel, M. A., & Miyatsu, T. (2018). Effects of flashcards on learning authentic materials: The role of detailed versus conceptual flashcards and individual differences in structure-building ability. *Journal of Applied Research in Memory and Cognition*, 7(4), 529–539. <https://doi.org/10.1037/h0101829>
- Lippmann, M., Narciss, S., Schwartz, N. H., & Danielson, R. W. (2013). *Effects of text titles and the timing of keywording tasks on metacognitive monitoring* [Conference session]. Proceedings of the Annual Meeting of the Cognitive Science Society, Berlin, Germany.
- Magreehan, D. A. (2016). *Presentation of electronic glosses: Effects on student learning and monitoring from text* [Doctoral dissertation, Texas Tech University]. <http://hdl.handle.net/2346/73659>
- Morgan, M., Brickell, G., & Harper, B. (2008). Applying distributed cognition theory to the redesign of the 'Copy and Paste' function in order to promote appropriate learning outcomes. *Computers & Education*, 50(1), 125–147. <https://doi.org/10.1016/j.compedu.2006.04.006>
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1), 105–125. <https://doi.org/10.1037/1082-989X.7.1.105>
- Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524759>
- Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, 144(7), 710–756. <https://doi.org/10.1037/bul0000151>
- Pan, S. C., Rickard, T. C., & Bjork, R. A. (2021). Does spelling still matter—And if so, how should it be taught? Perspectives from contemporary and historical research. *Educational Psychology Review*, 33(4), 1523–1552. <https://doi.org/10.1007/s10648-021-09611-y>
- Pan, S. C., Zung, I., & Imundo, M. (2022a). *Digital flashcard interface video*. Open Science Framework. <https://osf.io/vtacs>
- Pan, S. C., Zung, I., & Imundo, M. (2022b). *Optimising the use of digital flashcards as learning devices*. Open Science Framework. <https://doi.org/10.17605/OSF.IO/K9Q8R>
- Quizlet. (2022). *About quizlet*. Retrieved March 22, 2022, from <https://quizlet.com/mission>
- Reeser, V., & Moon, D. (2018). *Digital flashcard study methods: Teacher-Led vs. independent study* [Conference session]. The 25th Korea TESOL-PAC International Conference, Seoul, South Korea.
- 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>
- Sage, K., Krebs, B., & Grove, R. (2019). Flip, slide, or swipe? Learning outcomes from paper, computer, and tablet flashcards. *Technology, Knowledge and Learning*, 24(3), 461–482. <https://doi.org/10.1007/s10758-017-9345-9>
- Sage, K., Rausch, J., Quirk, A., & Halladay, L. (2016). Pacing, pixels, and paper: Flexibility in learning words from flashcards. *Journal of Information Technology Education*, 15, 431–456. <https://doi.org/10.28945/3549>
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–218. <https://doi.org/10.1111/j.1467-9280.1992.tb00029.x>
- Senzaki, S., Hackathorn, J., Appleby, D. C., & Gurung, R. A. (2017). Reinventing flashcards to increase student learning. *Psychology Learning & Teaching*, 16(3), 353–368. <https://doi.org/10.1177/1475725717719771>
- 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>
- Stengers, H., Deconinck, J., Boers, F., & Eyckmans, J. (2016). Does copying idioms promote their recall? *Computer Assisted Language Learning*, 29(2), 289–301. <https://doi.org/10.1080/09588221.2014.937723>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Weinstein, Y., McDermott, K. B., & Roediger, H. L., III. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, 16(3), 308–316. <https://doi.org/10.1037/a0020992>
- Wilkinson, D. (2020a). *Deliberate vocabulary learning from word cards*. *Vocabulary Learning and Instruction*, 9(2), 69–74. <https://doi.org/10.7820/vli.v09.2.wilkinson>
- Wilkinson, D. (2020b). *Effects of word card methodology and testing on vocabulary knowledge and motivation* [Doctoral dissertation, Temple University]. <https://doi.org/10.34944/dspace/292>
- Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, 20(6), 568–579. <https://doi.org/10.1080/09658211.2012.687052>
- Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? *Memory*, 30(8), 923–941. <https://doi.org/10.1080/09658211.2022.2058553>

Appendix A

List of Key Terms and Definitions

Text passage	Key term	Definition
Expressionist art	Apollonian	Described things relating to the God Apollo and representing reason, culture, harmony, and restraint.
	Blauer Reiter	A loosely knit organization of artists that used abstract forms and prismatic colors to explore the spiritual values of art as a counter to what they saw as the corruption and materialism of their age.
	Die Brücke	An organization of artists that were in revolt against what they saw as the superficial naturalism of academic Impressionism and who wanted to reinfuse German art with spiritual vigor through an elemental, highly personal and spontaneous expression.
	Dionysian	Described things relating to the God Dionysus and representing excess, irrationality, lack of discipline, and unbridled passion.

(Appendices continue)

Appendix A (continued)

Text passage	Key term	Definition
	Expressionism	An artistic style in which the artist seeks to depict not objective reality but rather the subjective emotions and responses that objects and events arouse.
	Impressionism	An artistic style in which the artist attempts to accurately and objectively record visual reality in terms of transient effects of light and color.
	Metaphysical painting	A style of painting that used representational but incongruous imagery to produce disquieting effects on the viewer.
	Neue Sachlichkeit	Artists created works executed in a more realistic style that reflected the resignation and cynicism of the post-World War I period in Germany.
	New art	An international, middle-class artistic movement that sought to reflect the intensive psychic and sensory stimuli of the modern city by using flat patterning and bold forms.
	Primitivism	An esthetic idealization that aimed to recreate so-called primitive experience by using nonindustrial elements that were meant to be closer to the origins of humanity and were consequently considered more pure.
Ancient Rome	Domus	A type of private, single-family residence of modest to palatial proportions inhabited primarily by the wealthy upper class.
	Donativum	A donation given to each soldier upon the emperor's accession to secure their loyalty.
	Insula	Tenements providing economically practical housing that were inhabited primarily by the laboring class.
	Palatine Hill	A plateau on which the city was founded and the city's aristocratic quarter.
	Paterfamilias	The oldest male and the head of the family, to whom his wife, his slaves, and possibly several generations of his descendants were subject and to whom title to all property was vested.
	Praetorian Guard	Household troops of Roman emperors that had significant political influence and generally participated in appointing emperors.
	Proconsular imperium	Gave the emperor authority over the Roman army, as well as the power to declare war, ratify treaties, negotiate with foreign leaders, and control senate membership.
	Sacramentum	An oath of allegiance taken by soldiers to their commander that was sworn in a sacred place and using a formula that had a religious connotation.
	Salutatio	The daily morning ritual of paying their respects in the houses of senators, who were obligated to protect them.
	Tribunicia Potestas	Vested in the emperor authority over Rome's civil government, including the power to preside over and to control the Senate, made him personally inviolable, and gave him the power to veto measures freely, summon the organs of government, and propose decrees and legislation.

Appendix B

Examples of High- and Low-Quality Definitions Created in the User-Generated Condition of Experiment 3A and Featured in the Premade Conditions of Experiment 4A or 4B

Text passage	Key term	High-quality definition (used in premade condition of Experiment 4A)	Low-quality definition (used in premade condition of Experiment 4B)
Expressionist art	Impressionism	A style of art where an artist aims to objectively and accurately record visual reality using transient light and color.	Using colors and avoiding subjectivism
	Neue Sachlichkeit	An artistic style in which artists made artwork in a more realistic style in order to reflect the resignation and cynicism that surfaced following World War I in Germany.	Art that went against World War I
	New art	Art movement that sought to reflect the stimuli of the modern city through flat patterns and bold forms.	Using patterns
Ancient Rome	Insula	Practical housing that was provided to a large portion of Rome's population. Its residents were mainly labor class.	Where the laboring class lived
	Paterfamilias	The head of the household, usually oldest male of the family; wife, slaves, and descendants were his dependents, and he was in charge of properties.	Head of Roman family
	Sacramentum	A Roman soldier's oath of complete loyalty to the emperor, usually done in a sacred area and including a religious connotation.	Oath of loyalty

Received April 21, 2022
 Revision received August 18, 2022
 Accepted October 14, 2022 ■