AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

© 2025 The Author(s)
ISSN: 0022-0663

2025, Vol. 117, No. 6, 980–997
https://doi.org/10.1037/edu0000949

Combining Generative Tasks and Retrieval Tasks

Niklas Obergassel[1], Alexander Renkl[2], Tino Endres[2], Matthias Nückles[3], Shana K. Carpenter[4], and Julian Roelle[1]

[1] Faculty of Philosophy and Educational Research, Ruhr University Bochum
[2] Department of Psychology, University of Freiburg
[3] Department of Educational Science, University of Freiburg
[4] Department of Psychology, Oregon State University

Both tasks designed to elicit generative activities (i.e., generative tasks) and tasks designed to elicit retrieval activities (i.e., retrieval tasks) can substantially foster learning after an initial study phase in which learners encountered new content. More specifically, in line with the theoretical notion that generative tasks serve the function of fostering the construction of mental representations, generative tasks have been found to substantially foster comprehension, whereas retrieval tasks, in line with the theoretical notion that retrieval practice fosters the consolidation of mental representations in memory, have been found to foster retention. In view of these complementary functions, it is surprising that, to date, it has scarcely been investigated whether these tasks can be combined to good effects. Likewise, potential sequence effects, that is, whether engaging learners in generative tasks first (i.e., generative-first sequence) or in retrieval tasks first (i.e., retrieval-first sequence) would matter, have widely been ignored. The present study was designed to address these open issues. In a preregistered experiment with $N = 340$ university students, we varied whether learners engaged in generative tasks, retrieval tasks, both types of tasks (in two different sequences), or restudy tasks after an initial study phase. The combination of generative and retrieval tasks resulted in better retention than generative tasks as well as in better comprehension than retrieval tasks. No sequence effects were observed. We conclude that, regardless of the sequence, the combination of generative tasks and retrieval tasks is more effective than providing learners with either type of task alone.

Educational Impact and Implications Statement

Is it more effective to engage learners in a combination of generative tasks and retrieval tasks than in either type of task alone, and if so, does the sequence in which the two types of tasks are combined matter? In the present experiment with university students, combining generative tasks and retrieval tasks proved to add statistically identifiable value over using either type of task alone—irrespective of choice of sequence. While this finding emphasizes the high complementarity of generative tasks and retrieval tasks in promoting learning, it also gives rise to tentative objections against the claim that choice of sequence plays a prominent role when it comes to combining the two types of tasks to good effect.

This article was published Online First April 24, 2025.

Olusola O. Adesope served as action editor.

Niklas Obergassel https://orcid.org/0000-0002-2311-820X
Alexander Renkl https://orcid.org/0000-0001-7971-0619
Tino Endres https://orcid.org/0000-0003-3334-4064
Matthias Nückles https://orcid.org/0000-0001-9924-5806
Shana K. Carpenter https://orcid.org/0000-0003-0784-9026
Julian Roelle https://orcid.org/0000-0001-9581-9120

This study is part of the project “How to combine generative learning and retrieval practice in fostering the acquisition of declarative concepts,” funded by the Deutsche Forschungsgemeinschaft (German Research Foundation; Grant 462211397). We would like to thank the research assistants Lisa Camp, Chantal Cebula, Pernilla Cordes, Robin Fischer, Anton Lachmann, Maximilian Luppold, Tabea Steins, Leon van den Berg, and Melina Vogel for their help in conducting the experiment as well as Pernilla Cordes and Tabea Steins for coding the process data. Moreover, we would like to thank Jens Fleischer and Anne Scheunemann for their advice on the confirmatory factor analyses. We have no known conflicts of interest to disclose. This study’s design, hypotheses, and analysis plan were preregistered prior to data collection (see https://osf.io/wvq5n). All data as well as analyses are publicly available on the Open Science Framework and can be accessed via the following link: https://osf.io/u46c9.

Niklas Obergassel served as lead for data curation, formal analysis, software, visualization, and writing–original draft. Alexander Renkl and Matthias Nückles served as lead for conceptualization and funding acquisition and served in a supporting role for supervision. Tino Endres and Shana K. Carpenter served in a supporting role for conceptualization, funding acquisition, and supervision. Julian Roelle served as lead for conceptualization, funding acquisition, project administration, supervision, and writing–review and editing. Niklas Obergassel, Alexander Renkl, Tino Endres, Matthias Nückles, Shana K. Carpenter, and Julian Roelle contributed equally to methodology. Alexander Renkl, Tino Endres, Matthias Nückles, and Shana K. Carpenter contributed equally to writing–review and editing.

Correspondence concerning this article should be addressed to Niklas Obergassel, Faculty of Philosophy and Educational Research, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany. Email: niklas.obergassel@ruhr-uni-bochum.de

Keywords: generative learning, retrieval practice, sequence effects, comprehension, retention

Supplemental materials: https://doi.org/10.1037/edu0000949.supp

Making sense of previously provided information through generative processes and practicing retrieval of previously provided information from memory are two potent ways to enhance learning. From a theoretical view, these two activities promote learning by different functions. Generative (learning) activities mainly enhance the coherence of learners' mental representations and their integration with prior knowledge (i.e., construction function; Fiorella, 2023), whereas retrieval (practice) activities mainly contribute to the consolidation of learners' mental representations (i.e., consolidation function; Karpicke, 2017).

In view of these different functions, it is reasonable to assume that tasks designed to elicit generative activities (hereafter referred to as generative tasks) and tasks designed to elicit retrieval activities (hereafter referred to as retrieval tasks) differ in terms of their main benefits (see McDaniel, 2023; Roelle et al., 2023). Whereas generative tasks should mainly foster comprehension, retrieval tasks should primarily promote the degree to which learners can retain previously learned information over time. As in education usually both comprehension and retention are key, it might hence overall be more effective to engage learners in both types of tasks than engaging them in one type of task only. Surprisingly, however, this theoretical prediction has scarcely been tested (for a rare exception, see O'Day & Karpicke, 2021).

Likewise, potential sequence effects in combining the two types of tasks have widely been ignored. As engaging learners in generative tasks prior to retrieval tasks (i.e., generative-first sequence) should better align with the construction-before-consolidation logic proposed by theoretical accounts of knowledge and cognitive skill acquisition (e.g., Bjork et al., 2013; VanLehn, 1996), combining both types of tasks in a generative-first sequence might more effectively promote comprehension than combining them in a retrieval-first sequence that features the reverse order of tasks.

The present study was designed to address the outlined theoretical predictions. After learners had read an expository text that introduced them to new concepts, we varied whether learners were engaged (a) in generative tasks only (generative-only sequence), (b) in retrieval tasks only (retrieval-only sequence), (c) in generative tasks before retrieval tasks (generative-first sequence), (d) in retrieval tasks before generative tasks (retrieval-first sequence), or (e) in restudy tasks only (restudy-only sequence; control sequence). As main dependent variables, we assessed learners' retention and comprehension performance on an immediate or a 1-week-delayed posttest.

Generative Tasks

Tasks designed to engage learners in generative activities, such as organization and elaboration, after an initial study phase are beneficial. A remarkable body of research clearly demonstrates that generative tasks such as concept mapping (e.g., Schroeder et al., 2018; $g = 0.58$), example generation (e.g., Rawson & Dunlosky, 2016; $0.39 \leq$ pooled $d \leq 0.49$ compared to restudy), self-explaining (e.g., Bisra et al., 2018; $g = 0.55$), drawing (e.g., Fiorella & Zhang, 2018; $0.46 \leq d \leq 0.70$), or teaching (e.g., Kobayashi, 2019; $0.35 \leq g \leq 0.56$) can promote learning considerably. Specifically, in comparison to the common learning activity of restudy, generative tasks can promote both surface learning outcomes such as retention of factual information (e.g., Hoogerheide, Renkl, et al., 2019; $d = 0.55$) and deep-oriented learning outcomes such as comprehension or transfer (e.g., Hoogerheide, Visee, et al., 2019; $d = 0.71$), although the latter effects are typically (far) more substantial than the former (e.g., $0.46 \leq d \leq 1.15$ for retention vs. $0.76 \leq d \leq 1.34$ for comprehension in Fiorella & Kuhlmann, 2020) and thus often regarded as the main benefit of generative tasks.

Following generative learning theory (e.g., Fiorella, 2023; Fiorella & Mayer, 2016), the explanation for the outlined effects of generative tasks is that generative activities such as organization, elaboration, and inference generation serve a construction function. That is, the activities are believed to contribute to the construction of coherent and well-integrated mental representations for to-be-learned content. These structurally enhanced mental representations, in turn, foster comprehension (i.e., the main benefit of generative tasks) and, because they entail more retrieval routes than less coherent and less integrated mental representations (see Kintsch et al., 1990), retention to some extent as well (i.e., a side benefit of generative tasks).

At least when retention of factual information is tested after a delay of some days or more (hereafter referred to as delayed retention), however, the effects of generative tasks are relatively modest. That is, although previous research shows that generative tasks can foster delayed retention in comparison to restudy (e.g., Rawson & Dunlosky, 2016; $0.39 \leq$ pooled $d \leq 0.49$), empirical comparisons with retrieval tasks hint toward a potential suboptimality of generative tasks when it comes to fostering lasting knowledge (e.g., $-1.50 \leq d \leq -1.01$ compared to a retrieval task in Karpicke & Blunt, 2011; $-1.54 \leq d \leq -0.85$ compared to a retrieval task in O'Day & Karpicke, 2021).

Retrieval Tasks

Tasks designed to engage learners in practicing retrieval of provided information after an initial study phase are beneficial. A wealth of research clearly shows that retrieval tasks such as free recall (e.g., Karpicke & Blunt, 2011; $1.01 \leq d \leq 1.50$), cued recall (e.g., Rowland, 2014; $0.61 \leq g \leq 0.72$), or short answer quiz questions (e.g., Heitmann et al., 2021; $0.55 \leq d \leq 0.94$) can promote learning to a considerable extent. Specifically, the literature shows that in comparison to the common learning activity of restudy, retrieval tasks can promote both surface learning outcomes such as retention of factual information as well as deep-oriented learning outcomes such as comprehension or transfer, although the former effects are typically (far) more substantial than the latter (e.g., $1.25 \leq d \leq 2.13$ for retention vs. $0.54 \leq d \leq 1.03$ for transfer in Butler, 2010; for recent meta-analytical overviews, see e.g., Adesope et al., 2017; Pan & Rickard, 2018; Yang et al., 2021; see also Carpenter et al., 2022) and thus usually regarded as the main benefit of retrieval tasks. Importantly, however, these benefits mainly occur when learners' performance on retention or comprehension questions is measured after a delay (e.g., 1 day or 1 week). When retention or comprehension performance is assessed immediately after the learning phase (hereafter referred to as immediate retention and immediate comprehension), retrieval practice usually does not show beneficial effects (e.g., $-1.22 \leq d \leq -0.52$ for immediate retention in Roediger & Karpicke, 2006; $-0.88 \leq d \leq 0.07$ for immediate comprehension in Wong et al., 2023).

Following retrieval-based learning theories (e.g., Karpicke, 2017; Karpicke & Grimaldi, 2012), the theoretical explanation for the outlined effects of retrieval tasks is that practicing retrieval from memory serves a consolidation function. That is, by requiring learners to retrieve specific information from their memory, retrieval activities are believed to contribute to the consolidation of mental representations in memory. The consolidation can be explained, for example, by elaborative retrieval (i.e., semantically related knowledge items become associated with the retrieved knowledge and hence can serve as retrieval cues on future occasions; see Carpenter, 2009, 2011) or by episodic context updating (i.e., more different episodic context features become associated with the retrieved knowledge, making it easier to access on future occasions; see Karpicke et al., 2014). The consolidation of the mental representations slows down forgetting of learned information over time (e.g., Carpenter et al., 2008; Nickl & Bäuml, 2023), which is why it substantially fosters performance on delayed retention tasks that require memorization of previously encountered information (i.e., the main benefit of retrieval tasks; e.g., Rowland, 2014). Moreover, as memorization of previously encountered information can also be crucial for performing well on deep-oriented learning outcome measures such as comprehension and transfer, retrieval practice can foster performance on these measures to some degree as well (i.e., a side benefit of retrieval tasks; e.g., Pan & Rickard, 2018).

However, as deep-oriented comprehension and transfer tasks involve not only memorizing the corresponding materials but also understanding them (see Butler et al., 2017; Corral & Carpenter, 2020), the effects of retrieval tasks on deep-oriented outcome measures are relatively modest. That is, although previous research shows that retrieval tasks can foster comprehension in comparison to restudy (e.g., Agarwal & Roediger, 2011; $d = 0.54$), empirical comparisons with generative tasks hint toward a considerable suboptimality of retrieval tasks when it comes to fostering comprehension (e.g., $d = -0.54$ compared to organization and elaboration prompts in Roelle & Nückles, 2019, Experiment 2; $d = -0.76$ compared to elaboration prompts in Endres et al., 2024; $d = -0.88$ compared to a self-explanation task in Lachner et al., 2021, Experiment 2; $-0.89 \leq d \leq -0.88$ compared to a teaching task in Wong et al., 2023).

Combining Generative Tasks and Retrieval Tasks

In view of the outlined complementarity of generative tasks and retrieval tasks with regard to functions and benefits, there is reason to believe that in combination both types of tasks could benefit learning even further (see Hinze et al., 2013; McDaniel, 2023; Roelle et al., 2023). More specifically, since in a combination the benefits associated with generative tasks would likely compensate for the limitations associated with retrieval tasks (and vice versa), engaging learners in both types of tasks should likely produce more lasting learning (especially better delayed retention) than generative tasks alone and better comprehension than retrieval tasks alone.

Against this background, in recent years, researchers have started to investigate potential benefits of combining generative learning and retrieval practice. However, in most of the studies, such combination was realized within one task, that is, by implementing generative tasks in a closed-book format rather than in an open-book format. A closed-book format of learning tasks is characterized by the fact that previously introduced learning material is unavailable to the learners during task execution, whereas in an open-book format, learners have access to the previously introduced learning material during task execution (see Agarwal et al., 2008; Hiller et al., 2020; Roelle & Nückles, 2019). Thus, in contrast to an open-book format, a closed-book format of generative tasks requires learners to retrieve the idea units from the previously introduced learning material that are needed to perform the generative activities from memory (i.e., engage in retrieval activities) before they can execute the generative activities.

Overall, such within-task combination of generative learning and retrieval practice has yielded rather mixed findings (for a recent overview, see Roelle et al., 2023). That is, while some of the studies indicated that eliciting generative and retrieval activities through one task can be beneficial (e.g., combination benefit of $d = 0.43$ for comprehension compared to pure retrieval activities in Endres et al., 2024; combination benefit of $d = 0.42$ compared to pure retrieval activities in Hinze et al., 2013, Experiment 3; combination benefit of $0.71 \leq d \leq 0.86$ compared to pure generative activities in Rummer et al., 2019), there were lots of null results as well (see, e.g., Arnold et al., 2021; Roelle & Nückles, 2019, Experiment 2; Waldeyer et al., 2020). One explanation for the mixed effects of such a within-task combination on learning outcomes is that at least the generative part of tasks that combine the two types of learning activities within one task is usually substantially hampered by the retrieval hurdle that is associated with a closed-book format (e.g., closed-book disadvantage of $-0.80 \leq d \leq 0.53$ for essay quality in Arnold et al., 2021; closed-book disadvantage of $-1.47 \leq d \leq -1.31$ for explanation quality in Sibley et al., 2022). That is, because learners typically do not manage to recall or reconstruct all of the required information correctly (e.g., Roelle & Berthold, 2017; Waldeyer et al., 2020), a closed-book format usually hinders the execution of the generative activities, meaning that eliciting generative and retrieval activities through one task is likely suboptimal for exploiting the full potential of a combination of generative learning and retrieval practice.

In contrast to combining generative and retrieval activities within one task, the combination of the two activities through sequencing generative tasks and retrieval tasks should not entail the outlined suboptimalities and hence should be better suited to investigate the potential that generative learning and retrieval practice should entail for promoting learning outcomes. Surprisingly, however, the effects of a sequential combination of generative and retrieval tasks have scarcely been investigated so far. In a rare exception, O'Day and Karpicke (2021, Experiment 2) tested the combined effectiveness of established generative tasks (here: concept mapping tasks) and established retrieval tasks (here: free recall tasks) by pitting a sequential combination of the two types of tasks against either type of task alone. The authors found the combination of both types of tasks to result in better delayed retention ($d = 1.08$) and better delayed comprehension ($d = 0.57$) than the generative task only. Conversely, when compared with the retrieval task only, no such benefits of a sequential combination emerged.

However, these findings have to be interpreted cautiously, not only because of the probabilistic nature of statistical tests (i.e., Type II errors are possible), but also because there are reasonable doubts regarding the comparability of the implementation quality of the two types of tasks that were used in the respective study (see Roelle et al., 2023). Specifically, while providing learners with a retrieval task produced exceptional delayed retention and comprehension performance ($1.25 \leq d \leq 2.32$ compared to study-only), engaging them in the generative task was surprisingly ineffective (see O'Day & Karpicke, 2021). That is, even in comparison to a simple study-only task, which served as a manipulation check, providing the generative task resulted in no statistically identifiable benefits for comprehension ($p = .142$, $d = 0.38$), that is, in the outcome measure typically enhanced the most by generative tasks (see, e.g., Fiorella, 2023; Fiorella & Mayer, 2016). Considering that, at least when not attributable to statistical issues like Type II errors, such a failed manipulation check likely reflects methodological issues in the form of disparities in the implementation quality between the generative task and the retrieval task, the outlined findings might not be reliable, in particular with respect to the absent advantage of the combination over retrieval practice alone. Against this background, further research is warranted on the question of the combined effectiveness under the premise that established, effective, and well-implemented generative tasks and retrieval tasks are used.

Further research is also warranted with respect to the examination of sequence effects in combining generative tasks and retrieval tasks; that is, potential differences in combined effectiveness as a function of whether learners engage in generative tasks prior to retrieval tasks (i.e., generative-first sequence) or vice versa (i.e., retrieval-first sequence). From a theoretical perspective, sequence effects are highly plausible. Following the construction-before-consolidation logic proposed by theoretical accounts of knowledge and cognitive skill acquisition (e.g., Bjork et al., 2013; VanLehn, 1996), engaging learners in generative tasks prior to retrieval tasks should be more effective than the reverse order of tasks. More specifically, since learners who engage in generative tasks first should be able to first build coherent and well-integrated mental representations and then consolidate them, they should likely consolidate mental representations of relatively high quality. Hence, compared to learners who engage in retrieval tasks first, learners who engage in generative tasks first should likely benefit from superior consolidation of meaningful knowledge, allowing them to perform better when comprehension is tested after some delay. Unfortunately, however, conclusive research on the issue of sequence effects in combining generative tasks and retrieval tasks is scarce to date.

One study that has recently examined the role of the sequence in combining generative tasks and retrieval tasks is a study by Roelle and colleagues (Roelle, Froese, et al., 2022). Contradicting the theoretical prediction mentioned above, the authors found that a generative-first sequence showed no statistically identifiable advantage over a retrieval-first sequence in terms of delayed comprehension ($d = 0.17$). However, as learners' performance on the generative task that was provided during the follow-up learning phase (i.e., example generation) was quite low (i.e., performance of 45% and below), it is reasonable to assume that the implementation quality of the generative task was insufficient, which potentially prevented beneficial effects of a generative-first sequence to occur. In consideration of this potential methodological limitation (and the fact that findings derived from probabilistic statistical tests in single experiments should generally be treated with some caution), it is questionable whether the lack of difference between the generative-first sequence and the retrieval-first sequence with regard to delayed comprehension represents a reliable finding.

The Present Study

Building on the outlined theoretical considerations and empirical findings, the present study aimed at investigating (a) whether engaging learners in both generative tasks and retrieval tasks would be more effective than engaging learners in generative tasks or retrieval tasks alone (i.e., combined effectiveness) as well as (b) whether the sequence in which the two types of tasks are combined (generative-first sequence vs. retrieval-first sequence) would matter (i.e., sequence effects).

To address these aims, in the present study two different sequential combinations of generative tasks and retrieval tasks (generative-first sequence and retrieval-first sequence) were compared against each other as well as against sequences that consisted only of generative tasks (generative-only sequence) or only of retrieval tasks (retrieval-only sequence). As generative tasks and retrieval tasks, we used slightly modified variations of the respective tasks that had proven effective for fostering learning of materials similar to ours in comparison to common control tasks such as restudy (e.g., Rawson & Dunlosky, 2016). This was done to ensure an effective implementation of both types of tasks. A sequence consisting only of restudy tasks (restudy-only sequence) was implemented as a manipulation check to verify the implementation quality of the respective generative tasks and retrieval tasks.

We tested the following five preregistered hypotheses (see https://osf.io/wvq5n). First, as a manipulation check, we assumed that engaging learners in generative tasks or retrieval tasks only would result in specific benefits compared to engaging learners in restudy tasks only. In terms of generative tasks, we hypothesized these tasks to be more effective than restudy tasks for fostering both immediate and delayed comprehension (i.e., main benefit of generative tasks) and immediate and delayed retention (i.e., side benefit of generative tasks) (Hypothesis 1 [H1]). In terms of retrieval tasks, we hypothesized better results than restudy tasks with respect to promoting both delayed retention (i.e., main benefit of retrieval tasks) and delayed comprehension (i.e., side benefit of retrieval tasks) (Hypothesis 2 [H2]).

Second, we assumed that a combination of generative tasks and retrieval tasks would have specific benefits in comparison to providing generative tasks or retrieval tasks only. In comparison to generative tasks only, we expected the combination of generative and retrieval tasks to yield better delayed retention and comprehension (Hypothesis 3 [H3]). In comparison to retrieval tasks only, we expected the combination of generative and retrieval tasks to yield better immediate and delayed comprehension (Hypothesis 4 [H4]).

Third, we expected that in combining generative tasks and retrieval tasks, it would matter which type of task comes first. Specifically, based on the aforementioned rationale, we assumed that providing generative tasks prior to retrieval tasks (i.e., generative-first sequence) would result in better delayed comprehension than providing retrieval tasks prior to generative tasks (i.e., retrieval-first sequence) (Hypothesis 5 [H5]).

Method

Sample and Design

We determined the required sample size for our laboratory experiment by performing an a priori power analysis with G*Power (Faul et al., 2007). The power analysis was adapted to the type of statistical analyses stated in the preregistration (i.e., contrast analysis in a design with 10 groups and two covariates; see Preregistered Analyses Concerning our Hypotheses section). Since we aimed at examining effects of educational relevance (i.e., effects of at least medium size; $d = 0.40$ or $f = 0.20$; see Hattie, 2008), the input parameters for our power analysis were set to $f = 0.20$, $\alpha = .01$, and $1 - \beta = .80$. The α-level of .01 was due to the fact that not all preregistered contrasts (see Table 1 for an overview of all contrasts and contrast weights) were orthogonal to each other and hence required an adjusted α-level.

The power analysis indicated a required sample size of at least $N = 296$ participants. Hence, we recruited $N = 340$ students (254 female, two nonbinary, 84 male; $M_{\text{age}} = 23.33$, $SD_{\text{age}} = 5.35$) from two German universities as participants for our experiment.[1] Upon completion of the study, students received monetary compensation or course credit. The experiment was approved by the Ethics Committee of Ruhr University Bochum (EPE-2022-002). Due to technical difficulties (e.g., computer crash of the experiment; $n = 2$) and noncompliant participants (e.g., participants providing false information or showing unusually short learning times; $n = 32$), data of 34 participants had to be excluded. Thus, the final sample for all statistical analyses consisted of $N = 306$ participants.

The experiment followed a 5 × 2-factorial between-subjects design with the factors sequence of learning tasks and timing of posttest. The factor sequence of learning tasks related to the follow-up learning phase that was scheduled after all learners had read a text that covered four declarative concepts in an initial study phase. Dependent on the experimental condition, learners engaged either (a) in a generative task twice (generative-only sequence), (b) in a retrieval task twice (retrieval-only sequence), (c) in a restudy task twice (restudy-only sequence), (d) in a generative task first and a retrieval task second (generative-first sequence), or (e) in a retrieval task first and a generative task second (retrieval-first sequence). The factor timing of posttest had two levels: Learners took either (a) an immediate posttest (sequences with immediate posttest) or (b) a 1-week-delayed posttest (sequences with delayed posttest).

Materials

Expository Text

All learners read an adapted version of the textbook passage used by Zamary and Rawson (2018). The text covered four declarative concepts from the domain of social and cognitive psychology, namely (a) representativeness heuristic, (b) availability heuristic, (c) mere exposure effect, and (d) social facilitation. The text consisted of 346 words, and reading time was set to 8 min for all learners.

Learning Tasks

For the follow-up learning phase, three types of learning tasks were created: (a) a retrieval task, (b) a generative task, and (c) a restudy task. To ensure an effective implementation of the generative task and the retrieval task, we used generative and retrieval tasks that had already proven to effectively foster the acquisition of declarative concepts as compared to restudy tasks (e.g., Rawson & Dunlosky, 2016) and slightly modified them.

The retrieval task consisted of three cycles. During each cycle, learners were first presented with a cued recall task for each declarative concept (note that the order of these tasks was identical for all learners and across cycles), asking them to type in the correct definition for a given concept without having access to the expository text (i.e., closed-book format). Next, based on an established feedback procedure by Dunlosky and Rawson (2012), learners rated their answers by determining for each recalled definition whether the idea units that formed the correct definition of the respective concept (i.e., the correct solution to the cued recall task) were fully, partially, or not at all included in the recalled definition. Once learners had rated all idea units, the next cycle began after a break of 1 min (for an overview of the stages completed during each of the three cycles of the retrieval task, see Figure 1). Learners who engaged in the retrieval task twice (i.e., those assigned to a retrieval-only sequence) completed the outlined procedure 2 times (i.e., six cycles in total).

The generative task followed a similar procedure. First, learners were asked to generate their own example for each of the four declarative concepts while being re-presented with the expository text (note that the order in which the examples were generated was identical for all learners). The open-book format was chosen to reduce the requirement to engage in retrieval while performing this task (see, e.g., Roelle, Schweppe, et al., 2022). To provide learners with feedback, we used a procedure that was developed by Froese and Roelle (2023, 2024) for example-generation tasks. That is, after generating their own example for each concept, learners studied one correct example per concept that illustrated how the idea units that formed the respective concepts' definition could be correctly instantiated in an example. Using the correct example, learners then rated their self-generated examples by determining whether each idea unit included in the concept definition was fully, partially, or not at all illustrated in their example. Finally, all learners were asked to revise their examples using the text, the respective correct example, and the idea unit feedback as guides (for an overview of the stages of the generative task, see Figure 2). Learners who engaged in the generative task twice (i.e., those assigned to a generative-only sequence) were asked to come up with different examples during the second time.

As its main purpose was to serve as a control condition against which, in case of a successful implementation, the retrieval and generative tasks should show the established main benefits (i.e., better retention for the retrieval tasks and better comprehension for the generative tasks), the restudy task was closely oriented to previous studies in the generative and retrieval tasks literature (e.g., Butler, 2010; Roelle, Froese, et al., 2022; Sibley et al., 2022). Consequently, it simply required learners to restudy the learning material (i.e., the expository text) that was provided in the initial study phase and, other than

[1] Due to the COVID-19 situation and distance learning in most courses, students' willingness to return to campus to participate in the experiment in person was far lower than expected at the beginning of data collection. Thus, in order to increase our chances of recruiting an adequate number of participants, we decided to slightly deviate from the preregistered participation requirements by removing the age restriction (i.e., opening the experiment for university students younger than 18 years or older than 30 years).

Table 1

Overview of the Contrasts and Contrast Weights

Contrast	Experimental groups with contrast weight									
	Immediate posttest					Delayed posttest				
	Generative-only	Retrieval-only	Restudy-only	Generative-first	Retrieval-first	Generative-only	Retrieval-only	Restudy-only	Generative-first	Retrieval-first
Generative-only versus restudy-only (H1)	1	0	−1	0	0	1	0	−1	0	0
Retrieval-only versus restudy-only (H2)	0	0	0	0	0	0	1	−1	0	0
Generative-first and retrieval-first versus generative-only (H3)	0	0	0	0	0	−2	0	0	1	1
Generative-first and retrieval-first versus retrieval-only (H4)	0	−2	0	1	1	0	−2	0	1	1
Generative-first versus retrieval-first (H5)	0	0	0	0	0	0	0	0	1	−1

Note. H = hypothesis.

the retrieval and generative tasks, did not involve different stages or had explicit feedback components. Specifically, learners were instructed to reread the text carefully and completely and to take as much time as needed for properly understanding all aspects (with mandatory reading time being set to at least 8 min; see Figure 3).

Instruments and Measures

For all aggregated measures, we conducted confirmatory factor analyses (CFAs) with maximum likelihood estimation in IBM SPSS Amos (Version 29.0.0) to test the assumed measurement model. Model fit was evaluated based on guidelines by Gäde et al. (2020) and Weiber and Mühlhaus (2014): $\chi^2/df \leq 3$, Tucker-Lewis index (TLI) ≥ .900, comparative fit index (CFI) ≥ .900, root-mean-square error of approximation (RMSEA) ≤ .080, and standardized root-mean-square residual (SRMR) ≤ .100 for an acceptable model fit. In case of unsatisfactory model fit, we optimized the measurement model wherever possible. Specifically, we removed items with negative variance (i.e., Heywood cases) and, as recommended by Brown (2015), items with nonsalient standardized factor loadings (i.e., standardized factor loadings < .30; see, e.g., Kline, 2005). In case of measurement models with more than one factor, we also considered unintended cross-loadings of items during model optimization (see Brown, 2015). All measurement models can be found in Figures S1–S7 in the online supplemental materials.

Academic Self-Concept

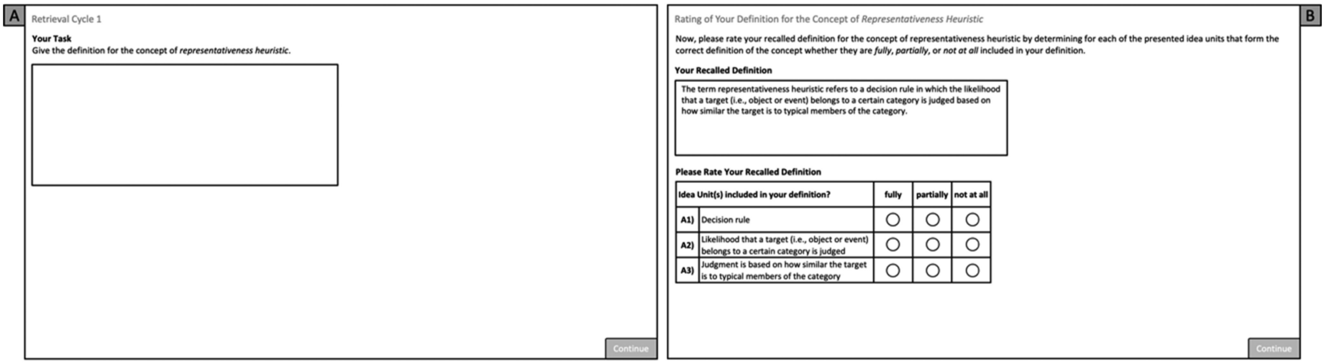
In view of the recent finding that learners' academic self-concept can affect the extent to which learners benefit from engaging in generative learning and retrieval practice (e.g., Roelle & Renkl, 2020), learners' academic self-concept was assessed as a control variable, using the subscale absolute academic self-concept from the Academic Self Concept Scales (Dickhäuser et al., 2002). The subscale comprised five items (e.g., "My study-related abilities are …"), answered on a Likert scale ranging from 1 to 7, with each extreme linked to an item-specific ending (e.g., 1 = … *low* and 7 = … *high*). For our analyses, we intended to build a mean score across all five items to obtain a score for learners' academic self-concept. However, the fit indices obtained by the CFA ($\chi^2/df = 8.367$, TLI = .883, CFI = .942, RMSEA = .155, SRMR = .047) only partly supported the assumed model (see Figure S1 in the online supplemental materials). Given that there was no room for optimization (i.e., all items showed positive variance and salient standardized factor loadings), we decided to deviate from the preregistration and not include academic self-concept in our analyses.

Pretest

As a second control variable, we assessed learners' prior knowledge regarding the four declarative concepts, using a pretest with eight questions (two questions for each concept; all questions were presented in the same order for all learners). Four of these questions were designed to measure factual knowledge by asking learners to type in the correct definition for a given concept (i.e., cued recall questions), whereas the other four questions were designed to assess comprehension by asking learners to generate an illustrative example for a given concept using a particular scenario (i.e., example generation questions). Two independent raters scored the answers of 20 learners

Figure 1

Screenshots Illustrating the Stages Completed by Learners During Each of the Three Cycles of the Retrieval Task in the Follow-Up Learning Phase



Retrieval Cycle 1

Your Task

Give the definition for the concept of *representativeness heuristic*.

Rating of Your Definition for the Concept of *Representativeness Heuristic*

Now, please rate your recalled definition for the concept of representativeness heuristic by determining for each of the presented idea units that form the correct definition of the concept whether they are *fully*, *partially*, or *not at all* included in your definition.

Your Recalled Definition

The term representativeness heuristic refers to a decision rule in which the likelihood that a target (i.e., object or event) belongs to a certain category is judged based on how similar the target is to typical members of the category.

Please Rate Your Recalled Definition

Idea Unit(s) included in your definition?		**fully**	**partially**	**not at all**
A1)	Decision rule	○	○	○
A2)	Likelihood that a target (i.e., object or event) belongs to a certain category is judged	○	○	○
A3)	Judgment is based on how similar the target is to typical members of the category	○	○	○

Note. (A) Cued recall stage. (B) Idea unit feedback stage. During each cycle, the two stages were completed for all four concepts.

on each question, using a scoring procedure that involved comparing the idea units included in the learners' answers with the idea units of the correct answers (scoring per idea unit: 1 point if completely correct, 0.5 points if partially correct, and 0 points if incorrect). Interrater reliability, as determined by the intraclass correlation coefficient (ICC) with measures of absolute agreement, was very good for all questions (all ICCs $>$ .85). Hence, the remaining answers were scored by only one rater. Through dividing the absolute score on each question by the theoretically attainable maximum score, a relative score (theoretical min: 0, theoretical max: 1) was built for each question.

For our analyses, we intended to build a pretest score by aggregating the relative scores of all eight questions. However, the CFA conducted on the specified measurement model (see Figure S2 in the online supplemental materials) indicated insufficient model fit ($\chi^2/df = 8.777$, TLI = .730, CFI = .807, RMSEA = .160, SRMR = .091). Hence, we optimized the measurement model by the above-mentioned means (i.e., four items were removed; see Figure S3 in the online supplemental materials) to obtain acceptable model fit ($\chi^2/df = 2.697$, TLI = .956, CFI = .985, RMSEA = .075, SRMR = .029).

Self-Reported High School Graduation Grade

Going beyond the preregistration, learners' self-reported high school graduation grade functioned as a third control variable. To indicate their high school graduation grade, learners were presented with a drop-down menu that offered values between 1.0 and 4.0. In accordance with the German grading system, lower values indicated a better graduation grade achieved by the learners (and vice versa).

Performance on the Retrieval Task(s)

To assess learners' retrieval performance during the follow-up learning phase, we examined learners' answers on the cued recall tasks for the idea units that made up the to-be-recalled concept definition (scoring per idea unit: 1 point if completely included, 0.5 points if partially included, and 0 points if not at all included). Interrater reliability, as determined by the ICC with measures of absolute agreement, was very good for all cued recall tasks (all ICCs $>$ .85). By dividing the absolute retrieval success scores on each cued recall task by the theoretically attainable maximum score, we obtained relative retrieval success scores (theoretical min: 0, theoretical max: 1) for each cued recall task in each cycle. For the further analyses, we averaged the relative retrieval success scores to compute an overall retrieval success score.[2] The specified measurement model (see Figure S4 in the online supplemental materials) showed acceptable model fit ($\chi^2/df = 2.103$, TLI = .958, CFI = .974, RMSEA = .076, SRMR = .043).

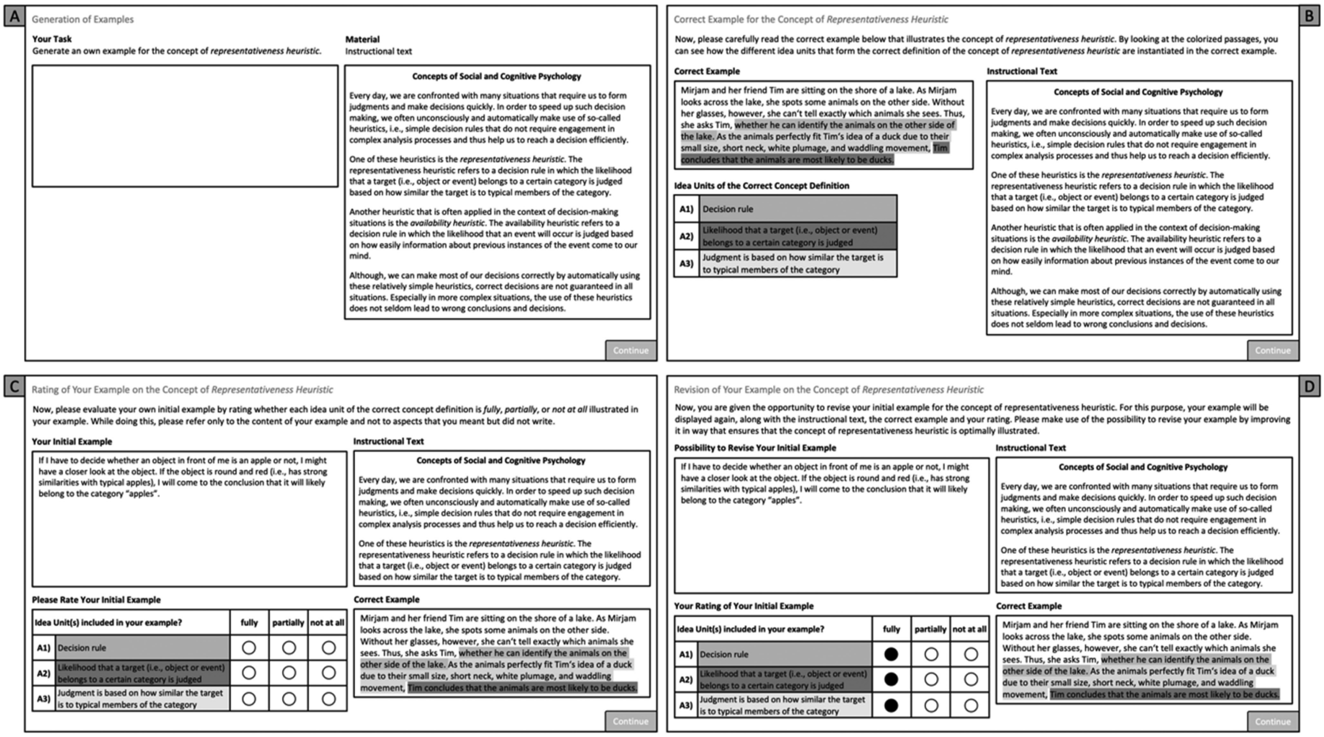
Performance on the Generative Task(s)

To assess learners' generative performance during the follow-up learning phase, we evaluated the quality of both the initial and the revised examples. Specifically, for each idea unit related to the concept to be exemplified (e.g., Idea Unit 3 of social facilitation: "Works only if performed task is easy or well-learned."), learners were awarded 1 point for a completely correct exemplification (e.g., "Mike is better at playing the guitar when his friends are around, because playing the guitar is a well-learned task for him"), 0.5 points for a partially correct exemplification (e.g., "While Mike is a skilled guitarist, he is really struggling with playing other instruments. However, whenever his friends are around, he plays the guitar and all other kind of instruments like never before"), and 0 points for a completely incorrect or missing exemplification (e.g., "Mike is really struggling with playing the piano. However, when his friends are around, he plays piano like never before"). This way, the extent to which each concept was correctly captured in the respective examples was determined. Interrater reliability as determined by the ICC with measures of absolute agreement was very good for all examples (all ICCs $>$ .85). Dividing the absolute quality scores by the theoretically attainable maximum score led to a relative quality score (theoretical min: 0, theoretical max: 1) for each initial and revised example. For the further analyses, we aggregated the relative

[2] Note that in the retrieval-only groups, retrieval success scores from two executions of the retrieval task were available. We expected the retrieval success to be higher in the second execution and hence the scores of the second execution to best capture the benefits of engaging in the retrieval task twice. Paired samples t tests comparing the mean scores from the two executions for each of the two retrieval-only groups backed this assumption, $t(33) = 8.17$, $p < .001$ (two-tailed), $d = 1.40$, 95% CI [0.92, 1.87], and $t(29) = 8.44$, $p < .001$ (two-tailed), $d = 1.54$, 95% CI [1.00, 2.07]. Thus, for the retrieval-only groups, we determined the overall score for retrieval success based on the scores from the second execution.

Figure 2

Screenshots Illustrating the Stages Completed by Learners During the Execution of the Generative Task in the Follow-Up Learning Phase



Generation of Examples

Your Task
Generate an own example for the concept of *representativeness heuristic*.

Material
Instructional text

Concepts of Social and Cognitive Psychology

Every day, we are confronted with many situations that require us to form judgments and make decisions quickly. In order to speed up such decision making, we often unconsciously and automatically make use of so-called heuristics, i.e., simple decision rules that do not require engagement in complex analysis processes and thus help us to reach a decision efficiently.

One of these heuristics is the *representativeness heuristic*. The representativeness heuristic refers to a decision rule in which the likelihood that a target (i.e., object or event) belongs to a certain category is judged based on how similar the target is to typical members of the category.

Another heuristic that is often applied in the context of decision-making situations is the *availability heuristic*. The availability heuristic refers to a decision rule in which the likelihood that an event will occur is judged based on how easily information about previous instances of the event come to our mind.

Although, we can make most of our decisions correctly by automatically using these relatively simple heuristics, correct decisions are not guaranteed in all situations. Especially in more complex situations, the use of these heuristics does not seldom lead to wrong conclusions and decisions.

Correct Example for the Concept of *Representativeness Heuristic*

Now, please carefully read the correct example below that illustrates the concept of *representativeness heuristic*. By looking at the colorized passages, you can see how the different idea units that form the correct definition of the concept of *representativeness heuristic* are instantiated in the correct example.

Correct Example

Mirjam and her friend Tim are sitting on the shore of a lake. As Mirjam looks across the lake, she spots some animals on the other side. Without her glasses, however, she can't tell exactly which animals she sees. Thus, she asks Tim, whether he can identify the animals on the other side of the lake. As the animals perfectly fit Tim's idea of a duck due to their small size, short neck, white plumage, and waddling movement, Tim concludes that the animals are most likely to be ducks.

Idea Units of the Correct Concept Definition

A1)	Decision rule
A2)	Likelihood that a target (i.e., object or event) belongs to a certain category is judged
A3)	Judgment is based on how similar the target is to typical members of the category

Instructional Text

Concepts of Social and Cognitive Psychology

Every day, we are confronted with many situations that require us to form judgments and make decisions quickly. In order to speed up such decision making, we often unconsciously and automatically make use of so-called heuristics, i.e., simple decision rules that do not require engagement in complex analysis processes and thus help us to reach a decision efficiently.

One of these heuristics is the *representativeness heuristic*. The representativeness heuristic refers to a decision rule in which the likelihood that a target (i.e., object or event) belongs to a certain category is judged based on how similar the target is to typical members of the category.

Another heuristic that is often applied in the context of decision-making situations is the *availability heuristic*. The availability heuristic refers to a decision rule in which the likelihood that an event will occur is judged based on how easily information about previous instances of the event come to our mind.

Although, we can make most of our decisions correctly by automatically using these relatively simple heuristics, correct decisions are not guaranteed in all situations. Especially in more complex situations, the use of these heuristics does not seldom lead to wrong conclusions and decisions.

Rating of Your Example on the Concept of *Representativeness Heuristic*

Now, please evaluate your own initial example by rating whether each idea unit of the correct concept definition is *fully*, *partially*, or *not at all* illustrated in your example. While doing this, please refer only to the content of your example and not to aspects that you meant but did not write.

Your Initial Example

If I have to decide whether an object in front of me is an apple or not, I might have a closer look at the object. If the object is round and red (i.e., has strong similarities with typical apples), I will come to the conclusion that it will likely belong to the category "apples".

Instructional Text

Concepts of Social and Cognitive Psychology

Every day, we are confronted with many situations that require us to form judgments and make decisions quickly. In order to speed up such decision making, we often unconsciously and automatically make use of so-called heuristics, i.e., simple decision rules that do not require engagement in complex analysis processes and thus help us to reach a decision efficiently.

One of these heuristics is the *representativeness heuristic*. The representativeness heuristic refers to a decision rule in which the likelihood that a target (i.e., object or event) belongs to a certain category is judged based on how similar the target is to typical members of the category.

Please Rate Your Initial Example

Idea Unit(s) included in your example?		fully	partially	not at all
A1)	Decision rule	○	○	○
A2)	Likelihood that a target (i.e., object or event) belongs to a certain category is judged	○	○	○
A3)	Judgment is based on how similar the target is to typical members of the category	○	○	○

Correct Example

Mirjam and her friend Tim are sitting on the shore of a lake. As Mirjam looks across the lake, she spots some animals on the other side. Without her glasses, however, she can't tell exactly which animals she sees. Thus, she asks Tim, whether he can identify the animals on the other side of the lake. As the animals perfectly fit Tim's idea of a duck due to their small size, short neck, white plumage, and waddling movement, Tim concludes that the animals are most likely to be ducks.

Revision of Your Example on the Concept of *Representativeness Heuristic*

Now, you are given the opportunity to revise your initial example for the concept of representativeness heuristic. For this purpose, your example will be displayed again, along with the instructional text, the correct example and your rating. Please make use of the possibility to revise your example by improving it in way that ensures that the concept of representativeness heuristic is optimally illustrated.

Possibility to Revise Your Initial Example

If I have to decide whether an object in front of me is an apple or not, I might have a closer look at the object. If the object is round and red (i.e., has strong similarities with typical apples), I will come to the conclusion that it will likely belong to the category "apples".

Instructional Text

Concepts of Social and Cognitive Psychology

Every day, we are confronted with many situations that require us to form judgments and make decisions quickly. In order to speed up such decision making, we often unconsciously and automatically make use of so-called heuristics, i.e., simple decision rules that do not require engagement in complex analysis processes and thus help us to reach a decision efficiently.

One of these heuristics is the *representativeness heuristic*. The representativeness heuristic refers to a decision rule in which the likelihood that a target (i.e., object or event) belongs to a certain category is judged based on how similar the target is to typical members of the category.

Your Rating of Your Initial Example

Idea Unit(s) included in your example?		fully	partially	not at all
A1)	Decision rule	●	○	○
A2)	Likelihood that a target (i.e., object or event) belongs to a certain category is judged	●	○	○
A3)	Judgment is based on how similar the target is to typical members of the category	●	○	○

Correct Example

Mirjam and her friend Tim are sitting on the shore of a lake. As Mirjam looks across the lake, she spots some animals on the other side. Without her glasses, however, she can't tell exactly which animals she sees. Thus, she asks Tim, whether he can identify the animals on the other side of the lake. As the animals perfectly fit Tim's idea of a duck due to their small size, short neck, white plumage, and waddling movement, Tim concludes that the animals are most likely to be ducks.

Note. (A) Example generation stage. (B) Correct example stage. (C) Idea unit feedback stage. (D) Example revision stage. All stages were completed for each of the four concepts.

quality scores to obtain an overall example quality score.[3] The specified measurement model (see Figure S5 in the online supplemental materials) showed acceptable model fit ($\chi^2/df = 2.024$, TLI = .955, CFI = .974, RMSEA = .075, SRMR = .038).

Time on Learning Tasks

In addition to example quality and retrieval success, we also measured time on learning tasks (i.e., the time in minutes that learners actively spent on the assigned learning tasks). We summed up the time scores from each of the two learning tasks that learners engaged in to obtain a score for learners' total time spent on learning tasks.

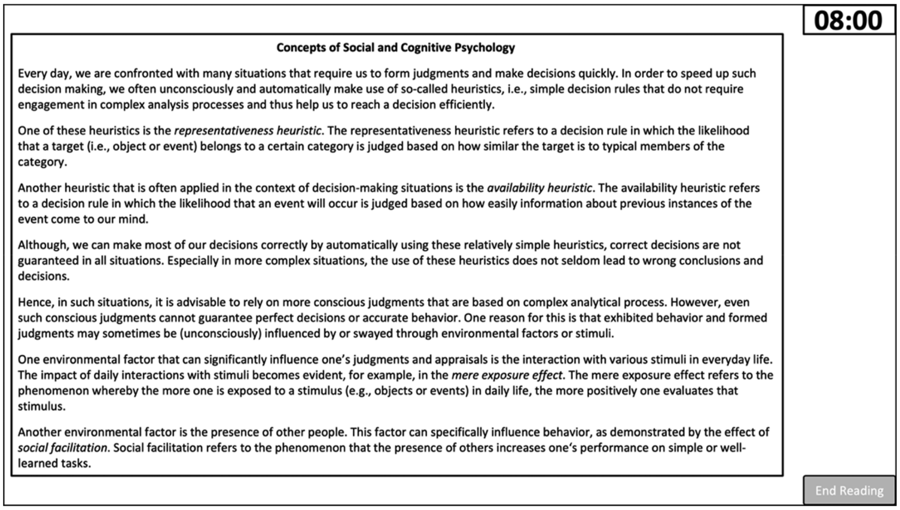
Posttest

We administered a posttest with 24 questions, either immediately after the follow-up learning phase (to assess immediate retention and comprehension) or with a delay of 1 week (to assess delayed retention and comprehension). The order of the questions was identical for all learners. Four questions (one per declarative concept) were designed to assess retention by asking learners to type in the correct definition for a given concept name (i.e., cued recall questions; identical to the cued recall questions in the pretest and to the cued recall tasks that formed the retrieval task in the follow-up learning phase). The remaining 20 questions were designed to measure comprehension by asking learners to either (a) generate an illustrative example for a given concept using a particular scenario (i.e., four example generation questions; although identical with the example generation questions in the pretest, these questions differed from the example generation tasks in the follow-up learning phase due to the fact that learners were required to generate examples that would not only exemplify the respective concept but also comply with a given scenario), (b) classify given examples or scenarios by correctly identifying the concept illustrated (i.e., 12 classification questions, adapted from Rawson et al., 2015; exclusive to the posttest), (c) identify similarities and differences between two given concepts at a time (i.e., two comparison questions; exclusive to the posttest), or (d) describe and explain how two concepts could be applied in a given scenario (i.e., two application questions; exclusive to the posttest). Two raters scored the answers of 20 participants (all ICCs > .85). Dividing the absolute scores on each question by

[3] Note that in the generative-only groups, example quality scores from two executions of the generative task were available. We expected the example quality scores to be higher in the second execution and hence the scores of the second execution to best capture the benefits of engaging in the generative task twice. Paired samples t tests conducted to compare the scores from the two executions for the two generative-only groups did not support this assumption, $t(29) = 1.80$, $p = .083$ (two-tailed), $d = 0.33$, 95% CI [−0.04, 0.69], and $t(26) = 0.91$, $p = .372$ (two-tailed), $d = 0.18$, 95% CI [−0.21, 0.55]. However, given that the scores from the second execution were at least descriptively higher, we nevertheless decided to base the overall score for example quality in the generative-only groups on the scores from the second execution and thereby maintained consistency in how the scores were determined for the retrieval and the generative task.

Figure 3

Screenshot Illustrating the Restudy Task in the Follow-Up Learning Phase



Concepts of Social and Cognitive Psychology

Every day, we are confronted with many situations that require us to form judgments and make decisions quickly. In order to speed up such decision making, we often unconsciously and automatically make use of so-called heuristics, i.e., simple decision rules that do not require engagement in complex analysis processes and thus help us to reach a decision efficiently.

One of these heuristics is the *representativeness heuristic*. The representativeness heuristic refers to a decision rule in which the likelihood that a target (i.e., object or event) belongs to a certain category is judged based on how similar the target is to typical members of the category.

Another heuristic that is often applied in the context of decision-making situations is the *availability heuristic*. The availability heuristic refers to a decision rule in which the likelihood that an event will occur is judged based on how easily information about previous instances of the event come to our mind.

Although, we can make most of our decisions correctly by automatically using these relatively simple heuristics, correct decisions are not guaranteed in all situations. Especially in more complex situations, the use of these heuristics does not seldom lead to wrong conclusions and decisions.

Hence, in such situations, it is advisable to rely on more conscious judgments that are based on complex analytical process. However, even such conscious judgments cannot guarantee perfect decisions or accurate behavior. One reason for this is that exhibited behavior and formed judgments may sometimes be (unconsciously) influenced by or swayed through environmental factors or stimuli.

One environmental factor that can significantly influence one's judgments and appraisals is the interaction with various stimuli in everyday life. The impact of daily interactions with stimuli becomes evident, for example, in the *mere exposure effect*. The mere exposure effect refers to the phenomenon whereby the more one is exposed to a stimulus (e.g., objects or events) in daily life, the more positively one evaluates that stimulus.

Another environmental factor is the presence of other people. This factor can specifically influence behavior, as demonstrated by the effect of *social facilitation*. Social facilitation refers to the phenomenon that the presence of others increases one's performance on simple or well-learned tasks.

Note. Besides being required to restudy the text for at least 8 min (as indicated by the timer), learners worked on the restudy task without time constraints.

the theoretically attainable maximum score resulted in a relative score for each of the 24 posttest questions.

By averaging the relative scores on the cued recall questions and the comprehension questions, we intended to build a retention and a comprehension score (theoretical min: 0, theoretical max: 1) for the further analyses. However, a CFA revealed that fit indices of the specified measurement model (see Figure S6 in the online supplemental materials) were not fully satisfactory ($\chi^2/df = 1.964$, TLI = .786, CFI = .806, RMSEA = .056, SRMR = .062). Hence, through the above-mentioned means we optimized the measurement model (i.e., nine comprehension items were removed; see Figure S7 in the online supplemental materials), which resulted in the retention score to be built as initially intended (i.e., based on the four cued recall questions) and the comprehension score to be based on 11 items in total ($\chi^2/df = 1.703$, TLI = .914, CFI = .927, RMSEA = .048, SRMR = .054).

Procedure

The study was conducted in the laboratory, where all learners worked individually in a digital learning environment, presented via the online platform Labvanced (Scicovery GmbH, 2023). First, academic self-concept and demographic data, including self-reported high school graduation grade, were assessed. Afterwards, all learners first took the pretest and then, during the initial study phase, read the expository text for 8 min. In the follow-up learning phase, learners worked on two tasks according to the assigned sequence of learning tasks (i.e., generative task twice, retrieval task twice, generative task prior to retrieval task, retrieval task prior to generative task, or restudy task twice). Learners completed the experiment by taking the posttest, either immediately or 1 week after the follow-up learning phase.[4]

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow JARS (Kazak, 2018). All data and analyses are publicly available on the Open Science Framework and can be accessed at https://osf.io/u46c9. Data were analyzed using JASP, Version 0.19.1 (JASP Team, 2024) and IBM SPSS Statistics (Version 29.0.2.0). This study's design, hypotheses, and analysis plan were preregistered prior to data collection (see https://osf.io/wvq5n).

Results

An α-level of .05 was used for all statistical tests. Tables 2 and 3 show the mean scores and standard deviations on all control and dependent variables of the study.

Preliminary Analyses

Prior to addressing our hypotheses, we tested for a priori differences between the experimental groups with regard to pretest score (note that academic self-concept was dropped from all analyses due to unsatisfactory psychometric validity) and examined correlations between this control variable and learners' posttest performance. Going beyond the preregistration, both types of preliminary analyses were also performed for learners' self-reported high school graduation grade as an additional control variable. To account for potential aptitude-treatment interactions (i.e., differential intervention effects), all correlations between the above variables were computed separately for each experimental group.

[4] Consistent with the preregistration, we also assessed cognitive load (using a questionnaire by Klepsch et al., 2017) as well as active and passive load (using two items by Klepsch & Seufert, 2021) at the end of (a) the initial study phase and (b) each of the two learning tasks during the follow-up learning phase. However, as these measures were not directly related to our hypotheses, we decided to refrain from reporting them in the present article. Nevertheless, to uphold open science standards, all data on cognitive load as well as active and passive load are part of the data set that is available on OSF under the following link: https://osf.io/u46c9.

Table 2

Means and (Standard Deviations) on All Control and Dependent Variables of the Study for Experimental Groups With Immediate Posttest

	Experimental groups				
Variable	Generative-only // immediate posttest ($n = 30$)	Retrieval-only // immediate posttest ($n = 34$)	Restudy-only // immediate posttest ($n = 32$)	Generative-first // immediate posttest ($n = 35$)	Retrieval-first // immediate posttest ($n = 32$)
Control variables					
Pretest score (0–1)	.00 (.02)	.00 (.02)	.00 (.02)	.01 (.03)	.00 (.00)
Self-reported high school graduation grade (1–4)[a]	1.97 (0.59)	1.99 (0.62)	2.08 (0.53)	1.99 (0.48)	2.08 (0.64)
Dependent variables					
Example quality (0–1)	.63 (.19)			.55 (.15)	.58 (.22)
Retrieval success (0–1)		.75 (.23)		.66 (.17)	.66 (.25)
Time on learning tasks (in min)	43.57 (14.05)	37.26 (10.09)	16.54 (0.55)	52.88 (18.51)	44.94 (13.19)
Retention score (0–1)	.72 (.17)	.78 (.19)	.62 (.26)	.73 (.21)	.78 (.24)
Comprehension score (0–1)	.58 (.14)	.51 (.14)	.50 (.17)	.59 (.10)	.56 (.13)

[a] Lower grades indicate better performance.

In terms of a priori group differences, neither a statistically detectable difference regarding pretest score, $F(9, 296) = 1.82$, $p = .064$, $\eta_p^2 = .05$, 95% confidence interval (CI) [0.00, 0.08], nor a statistically detectable difference regarding self-reported high school graduation grade, $F(9, 296) = 0.23$, $p = .991$, $\eta_p^2 = .01$, 95% CI [0.00, 0.00], emerged. Thus, the groups were comparable regarding these learning prerequisites.

The correlation analyses (see Table 4 for all results) showed no statistically detectable correlations between learners' pretest score and their retention or comprehension score in any of the 10 experimental groups (note that this is likely due to the fact that learners in most experimental groups had virtually no prior knowledge on the learning content). However, there were statistically detectable, negative correlations of self-reported high school graduation grade with learners' retention score and learners' comprehension score in some of the experimental groups (see Table 4; note that these correlations are negative due to the fact that in Germany, lower grades indicate better performance). Based on these results, and to reduce error variance, we decided to include learners' self-reported high school graduation grade as a covariate in all subsequent analyses on retention or comprehension scores that involved experimental groups for which a statistically detectable correlation between self-reported high school graduation grade and the respective score was found (note that the assumption of homogeneity of regression slopes was not violated in any of these analyses).

Preregistered Analyses Concerning our Hypotheses

Following the recommendations of the American Psychological Association guidelines regarding the use of statistical methods for addressing specific research questions in experimental designs with more than two conditions (see Wilkinson & APA Task Force on Statistical Inference, 1999), we tested all of our preregistered hypotheses by performing contrast analyses (see Table 1 for the specified contrasts and the related contrast weights). As the contrasts performed for retention (i.e., three contrasts) and comprehension (i.e., five contrasts) were not orthogonal, all obtained p-values were adjusted according to the Benjamini–Hochberg procedure (see Benjamini & Hochberg, 1995).

Generative-Only Versus Restudy-Only (Manipulation Check; H1)

In the sense of a manipulation check for the generative-only sequence, we predicted this sequence to result in better immediate

Table 3

Means and (Standard Deviations) on All Control and Dependent Variables of the Study for Experimental Groups With Delayed Posttest

	Experimental groups				
Variable	Generative-only // delayed posttest ($n = 27$)	Retrieval-only // delayed posttest ($n = 30$)	Restudy-only // delayed posttest ($n = 27$)	Generative-first // delayed posttest ($n = 33$)	Retrieval-first // delayed posttest ($n = 26$)
Control variables					
Pretest score (0–1)	.02 (.07)	.00 (.02)	.04 (.11)	.00 (.01)	.02 (.07)
Self-reported high school graduation grade (1–4)[a]	2.07 (0.60)	2.10 (0.34)	2.03 (0.57)	2.09 (0.56)	2.04 (0.51)
Dependent variables					
Example quality (0–1)	.54 (.18)			.53 (.17)	.57 (.19)
Retrieval success (0–1)		.76 (.22)		.68 (.19)	.65 (.23)
Time on learning tasks (in min)	50.60 (19.51)	44.76 (13.04)	16.52 (1.03)	50.91 (15.55)	40.85 (9.08)
Retention score (0–1)	.47 (.29)	.60 (.25)	.46 (.25)	.59 (.25)	.66 (.23)
Comprehension score (0–1)	.59 (.12)	.50 (.16)	.47 (.17)	.56 (.13)	.57 (.15)

[a] Lower grades indicate better performance.

Table 4

Results of the Preliminary Correlation Analyses Conducted for Each Experimental Group

	Correlated variables			
	Pretest score		Self-reported high school graduation grade	
Experimental group	With retention score	With comprehension score	With retention score	With comprehension score
Groups with immediate posttest				
Generative-only	$r(28) = .32, p = .085$	$r(28) = .13, p = .480$	$r(28) = -.31, p = .097$	$r(28) = -.30, p = .107$
Retrieval-only	$r(32) = -.21, p = .240$	$r(32) = .07, p = .716$	$r(32) = -.39, p = .024$	$r(32) = -.48, p = .004$
Restudy-only	$r(30) = .08, p = .673$	$r(30) = .09, p = .633$	$r(30) = -.15, p = .424$	$r(30) = -.51, p = .003$
Generative-first	$r(33) = -.11, p = .525$	$r(33) = -.05, p = .776$	$r(33) = -.39, p = .019$	$r(33) = -.30, p = .076$
Retrieval-first			$r(30) = -.53, p = .002$	$r(30) = -.22, p = .222$
Groups with delayed posttest				
Generative-only	$r(25) = .16, p = .418$	$r(25) = .11, p = .575$	$r(25) = -.47, p = .012$	$r(25) = -.21, p = .306$
Retrieval-only	$r(28) = -.02, p = .900$	$r(28) = -.12, p = .536$	$r(28) = -.16, p = .412$	$r(28) = -.05, p = .782$
Restudy-only	$r(25) = .08, p = .690$	$r(25) = .18, p = .382$	$r(25) = -.14, p = .476$	$r(25) = -.25, p = .208$
Generative-first	$r(31) = .09, p = .633$	$r(31) = -.04, p = .820$	$r(31) = .08, p = .660$	$r(31) = -.13, p = .488$
Retrieval-first	$r(24) = .23, p = .261$	$r(24) = .12, p = .568$	$r(24) = -.03, p = .887$	$r(24) = -.27, p = .184$

Note. Statistically undetectable correlations are grayed out. Due to zero variance in pretest scores in the retrieval-first group with immediate posttest, correlations between learners' pretest score and their retention or comprehension score could not be computed.

and delayed retention and comprehension than a restudy-only sequence (H1). Contrasting the respective groups (see first contrast in Table 1) on retention score, there was no statistically detectable difference between the groups, $t(295) = 1.19$, corrected $p = .118$ (one-tailed), $d = 0.44$, 95% CI [−0.29, 1.18]. However, with regard to comprehension, the outcome measure typically fostered the most by generative tasks, there was a statistically detectable difference that favored the generative-only groups, $t(295) = 4.07$, corrected $p < .001$ (one-tailed), $d = 1.52$, 95% CI [0.77, 2.26]. Hence, although, compared to engaging learners in restudy tasks only, engaging learners in generative tasks only was not sufficient to allow for a typical side benefit of generative tasks (i.e., better retention) to manifest, it did allow for the main benefit of these tasks (i.e., better comprehension) to materialize to a considerable extent. The implementation of the generative task can hence be considered successful.

Retrieval-Only Versus Restudy-Only (Manipulation Check; H2)

In the sense of a manipulation check for the retrieval-only sequence, we predicted this sequence to result in better delayed retention and comprehension than a restudy-only sequence (H2). Contrasting the respective groups (see second contrast in Table 1), a statistically detectable difference in terms of retention score emerged, favoring the retrieval-only group with delayed posttest over the restudy-only group with delayed posttest, $t(296) = 2.35$, corrected $p = .015$ (one-tailed), $d = 0.62$, 95% CI [0.10, 1.15]. In terms of comprehension score, however, there was no statistically detectable difference between the groups, $t(296) = 0.99$, corrected $p = .268$ (one-tailed), $d = 0.26$, 95% CI [−0.26, 0.79]. Thus, although compared to engaging learners in restudy tasks only, engaging them in retrieval tasks only was not sufficient to allow for better delayed comprehension on part of the learners (i.e., a typical side benefit of retrieval practice), it did allow for substantially better delayed retention (i.e., the main benefit of retrieval practice). Hence, the implementation of the retrieval task can be considered successful.

Combination Versus Generative-Only (H3)

In H3, we assumed that a combination of generative tasks and retrieval tasks would result in better delayed retention and comprehension than a generative-only sequence. Contrast analyses (the generative-first and retrieval-first groups were aggregated for these analyses; see third contrast in Table 1) revealed higher retention scores for the combined groups with delayed posttest than for the generative-only group with delayed posttest, $t(295) = 3.02$, corrected $p = .004$ (one-tailed), $d = 1.41$, 95% CI [0.48, 2.33]. However, there was no statistically detectable difference between the groups with regard to comprehension, $t(296) = -0.77$, corrected $p = .276$ (one-tailed), $d = -0.36$, 95% CI [−1.28, 0.56].

Combination Versus Retrieval-Only (H4)

In H4, we expected that a combination of generative tasks and retrieval tasks would result in better immediate and delayed comprehension than a retrieval-only sequence. Contrasting the respective groups (see fourth contrast in Table 1), we discovered a greater comprehension score in the combined groups than in the retrieval-only group, which proved to be statistically detectable, $t(295) = 3.21$, corrected $p = .002$ (one-tailed), $d = 1.98$, 95% CI [0.75, 3.20].

Generative-First Versus Retrieval-First (H5)

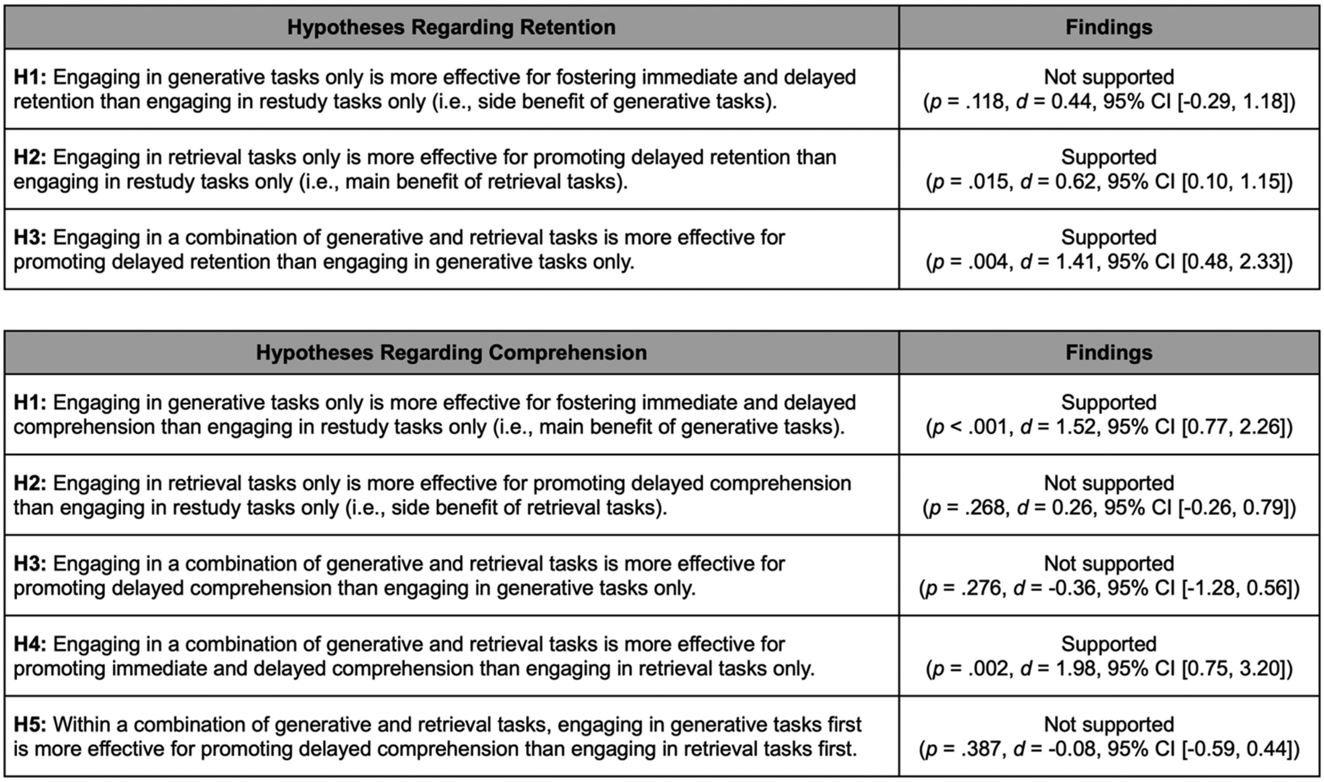
In H5, we predicted that a generative-first sequence would result in better delayed comprehension than a retrieval-first sequence. However, contrasting the respective groups (see fifth contrast in Table 1) regarding comprehension score, we did not find a statistically detectable difference, $t(296) = -0.29$, corrected $p = .387$ (one-tailed), $d = -0.08$, 95% CI [−0.59, 0.44].

Exploratory Analyses

Analyses on Learning Processes

To better understand the findings from the preregistered analyses (see Figure 4 for an overview of all hypotheses and the

Figure 4
Hypotheses and Corresponding Findings (Separate for Retention and Comprehension)



Hypotheses Regarding Retention	Findings
H1: Engaging in generative tasks only is more effective for fostering immediate and delayed retention than engaging in restudy tasks only (i.e., side benefit of generative tasks).	Not supported (p = .118, d = 0.44, 95% CI [-0.29, 1.18])
H2: Engaging in retrieval tasks only is more effective for promoting delayed retention than engaging in restudy tasks only (i.e., main benefit of retrieval tasks).	Supported (p = .015, d = 0.62, 95% CI [0.10, 1.15])
H3: Engaging in a combination of generative and retrieval tasks is more effective for promoting delayed retention than engaging in generative tasks only.	Supported (p = .004, d = 1.41, 95% CI [0.48, 2.33])

Hypotheses Regarding Comprehension	Findings
H1: Engaging in generative tasks only is more effective for fostering immediate and delayed comprehension than engaging in restudy tasks only (i.e., main benefit of generative tasks).	Supported (p < .001, d = 1.52, 95% CI [0.77, 2.26])
H2: Engaging in retrieval tasks only is more effective for promoting delayed comprehension than engaging in restudy tasks only (i.e., side benefit of retrieval tasks).	Not supported (p = .268, d = 0.26, 95% CI [-0.26, 0.79])
H3: Engaging in a combination of generative and retrieval tasks is more effective for promoting delayed comprehension than engaging in generative tasks only.	Not supported (p = .276, d = -0.36, 95% CI [-1.28, 0.56])
H4: Engaging in a combination of generative and retrieval tasks is more effective for promoting immediate and delayed comprehension than engaging in retrieval tasks only.	Supported (p = .002, d = 1.98, 95% CI [0.75, 3.20])
H5: Within a combination of generative and retrieval tasks, engaging in generative tasks first is more effective for promoting delayed comprehension than engaging in retrieval tasks first.	Not supported (p = .387, d = -0.08, 95% CI [-0.59, 0.44])

Note. H = hypothesis; CI = confidence interval.

corresponding findings), we conducted different types of exploratory (i.e., non-preregistered) analyses. First, we were interested in whether the groups that were contrasted concerning learning outcomes (i.e., retention and/or comprehension) differed with regard to effects on learning processes. We compared the different groups with respect to several measures related to learning processes (i.e., example quality, retrieval success, and/or time on learning tasks) wherever possible and reasonable, using the same contrasts as in the preregistered analyses. As contrasts performed for total time on learning tasks were not orthogonal, the respective *p*-values were adjusted according to the Benjamini–Hochberg procedure (see Benjamini & Hochberg, 1995). Second, we were interested in whether these potential differences would serve as mediators for the respective effects on retention and/or comprehension. Specifically, when at least one statistically detectable difference between the contrasted groups on any of the learning process measures was found, we conducted a mediation analysis with 95% percentile bootstrap CIs from 10,000 bootstrap samples using the SPSS macro PROCESS (Version 4.2; see Hayes, 2013). The path results of all mediation analyses that were conducted are presented in Figure 5.

Generative-Only Versus Restudy-Only. Contrasting the generative-only groups with the restudy-only groups regarding time on learning tasks, a statistically detectable difference emerged, $t(296) = 12.53$, corrected $p < .001$ (two-tailed), $d = 4.67$, 95% CI [3.84, 5.49]. In fact, time on learning tasks was statistically greater for learners in the generative-only groups than for learners in the restudy-only groups. However, this difference regarding time on learning tasks did not mediate any of the observed effects on learning outcomes, $a_1 \times b_1 = 0.024$ (-0.095, 0.123) for retention, and $a_2 \times b_2 = 0.014$ (-0.044, 0.061) for comprehension (see Panel A of Figure 5).

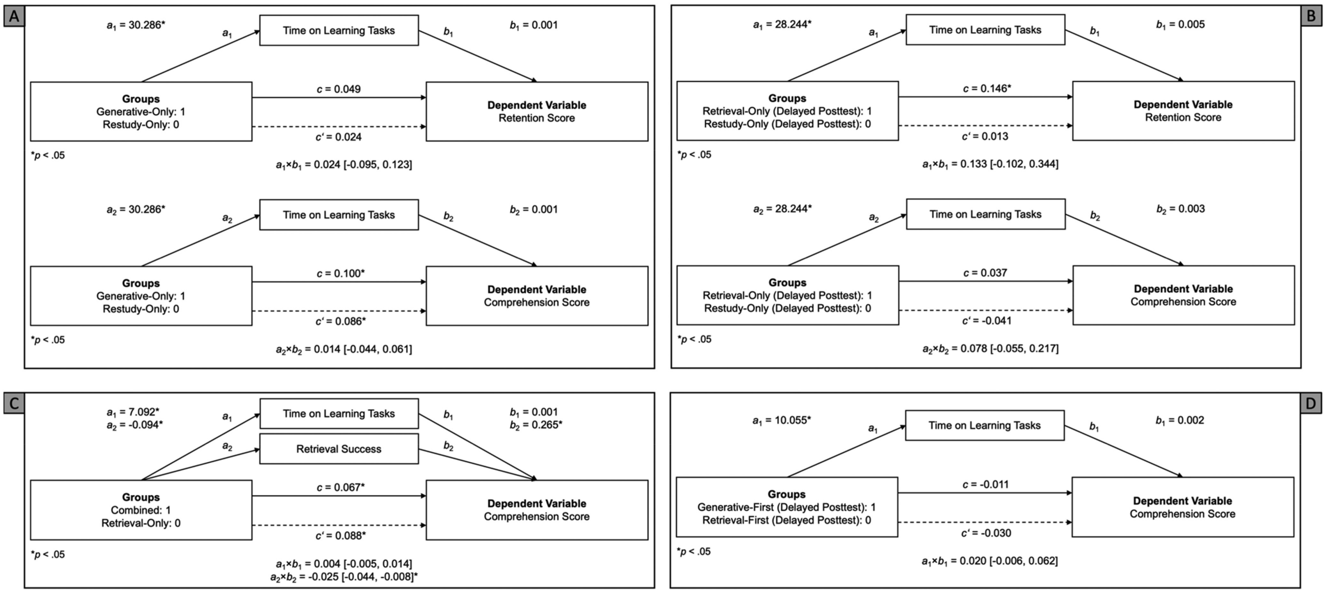
Retrieval-Only Versus Restudy-Only. Comparing the retrieval-only group with delayed posttest and the restudy-only group with delayed posttest regarding time on learning tasks, we found a statistically detectable difference, $t(296) = 8.13$, corrected $p < .001$ (two-tailed), $d = 2.16$, 95% CI [1.61, 2.71], indicating that the learners in the retrieval-only group spent more time on the learning tasks than their counterparts. However, no mediation effect on learning outcomes via time on learning tasks emerged, $a_1 \times b_1 = 0.133$ (-0.102, 0.344) for retention, and $a_2 \times b_2 = 0.078$ (-0.055, 0.217) for comprehension (see Panel B of Figure 5).

Combination Versus Generative-Only. Comparing the combined groups with delayed posttest with the generative-only group with delayed posttest, there was neither a statistically detectable difference for time on learning tasks, $t(296) = -1.55$, corrected $p = .123$ (two-tailed), $d = -0.72$, 95% CI [-1.64, 0.20], nor for example quality, $t(177) = 0.16$, $p = .872$ (two-tailed), $d = 0.08$, 95% CI [-0.84, 0.99].

Combination Versus Retrieval-Only. For the comparisons between the combined groups and the retrieval-only groups, contrast analyses revealed a statistically detectable difference with regard to time on learning tasks, $t(296) = 3.16$, corrected $p = .003$ (two-tailed), $d = 1.95$, 95% CI [0.73, 3.17], indicating that learners in the combined groups spent substantially more time on the assigned

Figure 5

Path Results of All Mediation Analyses Conducted



Note. (A) Generative-only groups versus restudy-only groups. (B) Retrieval-only group (delayed posttest) versus restudy-only group (delayed posttest). (C) Combined groups versus retrieval-only groups. (D) Generative-first group (delayed posttest) versus retrieval-first group (delayed posttest).
$^* p < .05$.

learning tasks. Moreover, there was a statistically detectable difference with regard to retrieval success, $t(184) = -2.90$, $p = .004$ (two-tailed), $d = -1.79$, 95% CI [−3.02, −0.56], indicating that the learners in the retrieval-only groups achieved a higher average retrieval success during the follow-up learning phase.

The mediation analysis (see Panel C of Figure 5) did not reveal a statistically detectable indirect effect on comprehension via time on learning tasks, $a_1 \times b_1 = 0.004$ (−0.005, 0.014). Hence, the superiority of the combined groups over the retrieval-only groups that was found for comprehension (H4) was not mediated by total time on learning tasks. However, there was a statistically detectable negative indirect effect on comprehension via retrieval success, $a_2 \times b_2 = -0.025$ (−0.044, −0.008). Accordingly, learners in the retrieval-only groups had an advantage of 0.025 units (i.e., 2.5%) with regard to comprehension score due to a substantially higher average retrieval success during the follow-up learning phase. However, the direct effect of group on comprehension was still statistically detectable and in favor of the learners in the combined groups. Hence, the results indicate that the learners in the combined groups outperformed their counterparts in the retrieval-only groups with regard to comprehension despite the disadvantage resulting from a lower average retrieval success.

Generative-First Versus Retrieval-First. Contrasting the generative-first group with delayed posttest and the retrieval-first group with delayed posttest, a statistically detectable difference regarding time on learning tasks emerged, $t(296) = 2.93$, corrected $p = .005$ (two-tailed), $d = 0.77$, 95% CI [0.25, 1.29]. The learners in the generative-first group spent substantially more time on the assigned learning tasks. With respect to example quality and retrieval success during the follow-up learning phase, however, no statistically detectable differences emerged, $t(177) = -0.90$, $p = .372$ (two-tailed), $d = -0.24$, 95% CI [−0.75, 0.28], and $t(184) = 0.47$, $p = .636$ (two-tailed), $d = 0.12$, 95% CI [−0.39, 0.64], respectively. The mediation analysis (see Panel D of Figure 5) did not reveal a statistically detectable indirect effect of group on comprehension score via time on learning tasks, $a_1 \times b_1 = 0.020$ (−0.006, 0.062). Thus, the observed null effects between the two groups with regard to delayed comprehension were not affected by the statistically detected difference in time on learning tasks.

Discussion

Benefits of Combining Generative Tasks and Retrieval Tasks

The primary goal of the present study was to examine the benefits that engaging learners in a combination of generative tasks and retrieval tasks would provide over either type of task alone. Based on the complementary potentials and limitations of generative and retrieval tasks for learning, we expected their combination to yield better delayed retention and comprehension than generative tasks alone (H3) as well as better immediate and delayed comprehension than retrieval tasks alone (H4).

Overall, combining generative tasks and retrieval tasks proved to be beneficial. That is, aside from a missing advantage of a combination over generative tasks only with regard to delayed comprehension that partially contradicted H3 (see further below), our hypotheses on the effectiveness of combining generative tasks and retrieval tasks were fully confirmed. In fact, combining both types of tasks did, regardless of potential differences in learning processes, result in a statistically detectable advantage over generative tasks only in terms of delayed retention (H3) as well as a statistically detectable advantage over retrieval tasks only in terms of immediate and delayed comprehension (H4).

At first glance, this pattern of results could simply be due to transfer appropriate processing (see Morris et al., 1977). That is, one could argue that the learners who engaged in both a generative and a retrieval task might have outperformed their generative-only counterparts on delayed retention and their retrieval-only counterparts on comprehension simply because the retrieval and generative activities in the follow-up learning phase were identical to the processes required on the retention and comprehension questions on the posttest. On closer inspection, however, transfer appropriate processing can only explain part of the results. Specifically, in terms of effects on retention, a transfer appropriate processing explanation is plausible. The retention measure on the posttest consisted of the same four cued recall questions from the retrieval task in the follow-up learning phase. By contrast, in terms of effects on comprehension, a transfer appropriate processing explanation is less plausible. Not only were the four example generation questions on the posttest different from the example generation tasks that formed the generative task in the follow-up learning phase (i.e., on the posttest the learners were required to exemplify the concepts using predefined scenarios, see the Posttest subsection in the Instruments and Measures section), but the comprehension measure, whose measurement model was optimized through CFAs in advance, involved 11 questions in total (i.e., seven questions that were substantially dissimilar to the example generation tasks in the follow-up learning phase).

An explanation that can account for the pattern of results concerning both retention and comprehension is that the retrieval task and the generative task served different functions for learning (see McDaniel, 2023; Roelle et al., 2023). That is, the retrieval task primarily contributed to consolidating (i.e., solidifying) the specific idea units of the definitions of the four declarative concepts in memory (i.e., consolidation function), which fostered performance on the retention questions on the posttest. The generative task, by contrast, primarily supported learners in constructing coherent and well-integrated mental representations for to-be-learned content, which fostered understanding of the idea units, and thus, performance on various types of comprehension questions on the posttest.

Jointly, these findings clearly indicate that the combination of generative and retrieval tasks can be fruitful. This conclusion contrasts with the study by O'Day and Karpicke (2021), who recently pitted a combination of generative and retrieval tasks against either type of task alone and concluded that combining the two types of tasks is not better than retrieval tasks alone. One reason for this discrepancy in results and conclusions could lie in the different implementation quality of the two types of tasks. Specifically, while in the study by O'Day and Karpicke (2021), the generative and retrieval tasks likely differed substantially in terms of implementation quality (i.e., while the conditions under which the two types of tasks were implemented proved to be quite favorable for the retrieval task, the opposite was true for the generative task), in the present study, both types of tasks were likely implemented well and in a typical fashion. That is, although testing the manipulation check hypotheses (H1 and H2) revealed that the benefits of the two types of tasks relative to restudy control tasks did not fully match our expectations (H1: no side benefit of the generative tasks for immediate and delayed retention; H2: no side benefit of the retrieval tasks for delayed comprehension), both types of tasks substantially promoted comprehension or delayed retention, respectively, thus fulfilling their main functions for learning. Hence, at least when generative tasks and retrieval tasks are implemented in a way that allows for the main functional benefits of both types of tasks to occur, engaging learners in both types of tasks seems to yield unique advantages over either type of task alone.

Given the novelty of this finding, future research should follow up on effective combinations of generative and retrieval tasks under conditions of high-quality implementation. While doing so, future research could also follow up on the abovementioned finding that, contrary to our expectations, the combination of generative and retrieval tasks did not produce better delayed comprehension than generative tasks alone (H3). Given the substantial benefits that retrieval tasks usually provide for lasting learning (e.g., Karpicke & Blunt, 2011) and the fact that these benefits should compensate the limitations of generative tasks in promoting delayed comprehension in a similar way than they did for the limitations of these tasks with respect to fostering delayed retention, this finding is surprising.

One explanation for this lack of combined advantage could lie in the fact that the posttest that was used to measure delayed retention and comprehension featured a delay of only 1 week. Specifically, given that coherent and well-integrated mental representations that facilitate comprehension usually entail more retrieval routes than rather superficial mental representations and should therefore be less prone to time-induced forgetting (see Kintsch et al., 1990), it is possible that delaying the comprehension questions by 1 week was not sufficient to induce forgetting of well-comprehended meaningful knowledge to a degree that would require additional consolidation as provided by retrieval tasks. Consequently, engaging learners in additional retrieval practice did not yield any benefits for their performance on the 1-week-delayed comprehension questions. Although this explanation might seem rather unlikely at first glance, given that a delay of 1 week is frequently used in research that addresses differential forgetting rates on retention and comprehension questions (e.g., Butler, 2010; Endres et al., 2017; Hinze & Rapp, 2014; Roelle & Nückles, 2019), our results concerning comprehension provide strong support for it. That is, while learners' performance on the retention questions that targeted more superficial mental representations was notably lower on a 1-week-delayed posttest (.46–.66) than on an immediate posttest (.62–.78), this was not the case for the comprehension questions that targeted more coherent and integrated mental representations (see Tables 2 and 3). In fact, learners' performance on these questions was nearly identical between an immediate posttest (.50–.59) and a 1-week-delayed posttest (.47–.59), a circumstance that likely reflected the fact that coherent and well-integrated mental representations are less prone to time-induced forgetting. Future studies should replicate the present study with a longer posttest delay (e.g., 2 weeks or more) to put the above explanation to the test.

Sequence Effects in Combining Generative Tasks and Retrieval Tasks

A secondary goal of the present study was to examine the role that the sequence in which generative tasks and retrieval tasks are combined (i.e., generative-first sequence vs. retrieval-first sequence) would play for combined effectiveness. Based on the rationale that, due to requiring learners to first construct and then consolidate coherent and well-integrated mental representations, a generative-first sequence should allow for consolidating mental representations of higher quality, we expected a generative-first sequence to produce better delayed comprehension than a retrieval-first sequence (H5). However, comparing the two sequences on delayed comprehension, no statistically detectable differences emerged. Hence, contrary to our predictions, it did not matter for delayed comprehension whether the two types of tasks were combined in a generative-first or in a retrieval-first sequence.

From an empirical perspective, this finding fits in well, given that it is perfectly in line with the results of Roelle and colleagues (Roelle, Froese, et al., 2022), who did not observe any sequence effects for delayed comprehension. Nevertheless, at least when viewed through the lens of knowledge and cognitive skill acquisition theories (e.g., Bjork et al., 2013; VanLehn, 1996), this finding is surprising. These theories clearly suggest a construction-before-consolidation logic, which should align best with a generative-first sequence.

One explanation for the absent advantage of a generative-first sequence regarding delayed comprehension could lie in the design of the retrieval task. Specifically, although our retrieval task resembled a variation of an established retrieval task and was designed in accordance with essential guidelines for effective retrieval practice (e.g., multiple cycles and corrective feedback), it might not have allowed learners assigned to a generative-first sequence to consolidate the generative-task-related improvements in their mental representations. More specifically, as the implemented retrieval task required learners to retrieve only concept definitions, instead of both definitions and generative products (i.e., self-generated examples), it may not have necessarily focused learners' consolidation efforts on the coherent and well-integrated mental representations that were formed during the preceding generative task. Although the decision to require learners to only retrieve concept definitions was inevitable, given that retrieving a generative product was only feasible in a generative-first sequence, it may nevertheless have compromised the benefits that a generative-first sequence should entail for delayed comprehension.

Limitations and Future Research

In addition to the outlined open issues concerning the underlying reasons for our pattern of findings, the present study has some further important limitations. First, it should be noted that the expository text that served as the learning basis in the present study was of relatively short length (i.e., 346 words). While such rather brief expository texts (i.e., texts of up to 500 words) are not uncommon in authentic learning material such as textbooks (see Rawson et al., 2015), texts of such short length might not particularly challenge learners to construct a situation model. That is, in the present study, learners were hardly required to form a mental model of the text that comprised text-derived propositions as well as prior knowledge-based interpretations (e.g., Kintsch & Rawson, 2005; see also Richter & Schnotz, 2018), which would usually be required by more extensive texts that are also commonly used in authentic college or university settings (e.g., undergraduate reading assignments of up to 12,000 words). Consequently, the expository text used in the present study might have been suboptimal for revealing the beneficial effects that engaging in knowledge construction (i.e., generative activities) should entail for comprehension. In order to test whether the present findings would generalize and, in terms of the effects of (additionally) engaging learners in generative tasks, potentially even show larger effect sizes, future studies should replicate the present study with a substantially longer expository text.

A second important limitation is that the retention measure in the posttest was based only on the cued recall questions from the retrieval task in the follow-up learning phase. That is, unlike the comprehension measure, the retention measure did not include any questions that were dissimilar to the follow-up learning tasks and therefore exclusive to the posttest. On this basis, although (a) our results regarding comprehension and (b) the well-established finding that engaging in follow-up retrieval tasks produces benefits for knowledge consolidation that even show up on posttest tasks that are dissimilar to the executed retrieval tasks (e.g., Adesope et al., 2017; Pan & Rickard, 2018; Rowland, 2014) speak against this explanation, it cannot be entirely ruled out that transfer appropriate processing advantages resulting from the follow-up retrieval tasks were accountable for the observed benefits on retention. Accordingly, future research that further explores the potential of combining generative and retrieval tasks should use posttests that include not only comprehension tasks that are dissimilar to the follow-up generative tasks (as in the present study), but also retention tasks that are dissimilar from the follow-up retrieval tasks, so that learning benefits can be more clearly related to the different functions of generative tasks and retrieval tasks.

Finally, it should be highlighted that in the present study, all learners took a pretest before they entered the learning phase and took the follow-up learning tasks. In view of research on pretesting effects (e.g., Pan & Carpenter, 2023), it is possible that the pretest not only captured learners' prior knowledge but also unintentionally guided learners' encoding efforts during the subsequent study phase toward those aspects of the learning materials that were part of the pretest. As the pretest included all four items that were subsequently used as retrieval tasks, it could be argued that the effects of the retrieval tasks that were found in the present study might depend on taking the pretest beforehand. Although research on retrieval practice tasks widely shows that retrieval practice effects occur even when no such pretesting is present (e.g., Adesope et al., 2017; Carpenter et al., 2018; Pan & Sana, 2021), future studies that delve into the benefits of combining retrieval and generative tasks and aim to test the generalizability of the present findings should control for potential pretesting effects by experimentally varying the provision of a pretest.

Conclusion

In summary, the present study has two main implications: First, engaging learners in a combination of generative tasks and retrieval tasks adds statistically identifiable value over engaging learners in either type of task alone. That is, although a combination of the two types of tasks was not superior in all aspects of learning, it largely compensated for the limitations typically associated with either type of task alone by yielding substantially better delayed retention than generative tasks, as well as substantially better comprehension than retrieval tasks. Hence, contrary to previous conclusions (see O'Day & Karpicke, 2021), a combination can yield benefits even compared to retrieval tasks alone.

Second, when it comes to combining generative tasks and retrieval tasks, sequence does not seem to play a prominent role. That is, although theoretical ideas (e.g., theories of knowledge and cognitive skill acquisition; Bjork et al., 2013; VanLehn, 1996) and empirical findings (see Roelle, Froese, et al., 2022) support the claim that choice of sequence (i.e., generative-first vs. retrieval-first) should, at least partly, make a difference in combined benefits, we found no evidence for this. However, given the paucity of research on the underlying mechanisms and potential moderators of sequence effects, it seems too early to draw robust conclusions on the role of sequence in combining generative and retrieval tasks.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

Agarwal, P. K., & Roediger, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*(8), 836–852. https://doi.org/10.1080/09658211.2011.613840

Arnold, K. M., Eliseev, E. D., Stone, A. R., McDaniel, M. A., & Marsh, E. J. (2021). Two routes to the same place: Learning from quick closed-book essays versus open-book essays. *Journal of Cognitive Psychology*, *33*(3), 229–246. https://doi.org/10.1080/20445911.2021.1903011

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

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

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied*, *23*(4), 433–446. https://doi.org/10.1037/xap0000142

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

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

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. https://doi.org/10.1038/s44159-022-00089-1

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448. https://doi.org/10.3758/MC.36.2.438

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied*, *24*(1), 34–42. https://doi.org/10.1037/xap0000145

Corral, D., & Carpenter, S. K. (2020). Facilitating transfer through incorrect examples and explanatory feedback. *The Quarterly Journal of Experimental Psychology*, *73*(9), 1340–1359. https://doi.org/10.1177/1747021820909454

Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes [The academic self concept scales: Construction and evaluation of a new instrument]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *23*(4), 393–405. https://doi.org/10.1024/0170-1789.23.4.393

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Endres, T., Carpenter, S., & Renkl, A. (2024). Constructive retrieval: Benefits for learning, motivation, and metacognitive monitoring. *Learning and Instruction*, *94*, Article 101974. https://doi.org/10.1016/j.learninstruc.2024.101974

Endres, T., Carpenter, S. K., Martin, A., & Renkl, A. (2017). Enhancing learning by retrieval: Enriching free recall with elaborative prompting. *Learning and Instruction*, *49*, 13–20. https://doi.org/10.1016/j.learninstruc.2016.11.010

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. (2023). Making sense of generative learning. *Educational Psychology Review*, *35*(2), Article 50. https://doi.org/10.1007/s10648-023-09769-7

Fiorella, L., & Kuhlmann, S. (2020). Creating drawings enhances learning by teaching. *Journal of Educational Psychology*, *112*(4), 811–822. https://doi.org/10.1037/edu0000392

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

Fiorella, L., & Zhang, Q. (2018). Drawing boundary conditions for learning by drawing. *Educational Psychology Review*, *30*(3), 1115–1137. https://doi.org/10.1007/s10648-018-9444-8

Froese, L., & Roelle, J. (2023). Expert example but not negative example standards help learners accurately evaluate the quality of self-generated examples. *Metacognition and Learning*, *18*(3), 923–944. https://doi.org/10.1007/s11409-023-09347-w

Froese, L., & Roelle, J. (2024). How to support self-assessment through standards in dissimilar-solution-tasks. *Learning and Instruction*, *94*, Article 101998. https://doi.org/10.1016/j.learninstruc.2024.101998

Gäde, J. C., Schermelleh-Engel, K., & Brandt, H. (2020). Konfirmatorische Faktorenanalyse (CFA) [Confirmatory factor analysis (CFA)]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 615–659). Springer. https://doi.org/10.1007/978-3-662-61532-4_24

Hattie, J. (2008). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Heitmann, S., Obergassel, N., Fries, S., Grund, A., Berthold, K., & Roelle, J. (2021). Adaptive practice quizzing in a university lecture: A pre-registered field experiment. *Journal of Applied Research in Memory and Cognition*, *10*(4), 603–620. https://doi.org/10.1037/h0101865

Hiller, S., Rumann, S., Berthold, K., & Roelle, J. (2020). Example-based learning: Should learners receive closed-book or open-book self-explanation prompts? *Instructional Science*, *48*(4), 623–649. https://doi.org/10.1007/s11251-020-09523-4

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice.

Applied Cognitive Psychology, *28*(4), 597–606. https://doi.org/10.1002/acp.3032

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151–164. https://doi.org/10.1016/j.jml.2013.03.002

Hoogerheide, V., Renkl, A., Fiorella, L., Paas, F., & van Gog, T. (2019). Enhancing example-based learning: Teaching on video increases arousal and improves problem-solving performance. *Journal of Educational Psychology*, *111*(1), 45–56. https://doi.org/10.1037/edu0000272

Hoogerheide, V., Visee, J., Lachner, A., & van Gog, T. (2019). Generating an instructional video as homework activity is both effective and enjoyable. *Learning and Instruction*, *64*, Article 101226. https://doi.org/10.1016/j.learninstruc.2019.101226

JASP Team. (2024). *JASP* (Version 0.19.1) [Computer software]. https://jasp-stats.org

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., pp. 487–514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*(3), 401–418. https://doi.org/10.1007/s10648-012-9202-2

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 237–284). Elsevier Academic Press.

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 211–226). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch12

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*(2), 133–159. https://doi.org/10.1016/0749-596X(90)90069-C

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*, Article 1997. https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2021). Making an effort versus experiencing load. *Frontiers in Education*, *6*, Article 645284. https://doi.org/10.3389/feduc.2021.645284

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford Press.

Kobayashi, K. (2019). Learning by preparing-to-teach and teaching: A meta-analysis. *Japanese Psychological Research*, *61*(3), 192–203. https://doi.org/10.1111/jpr.12221

Lachner, A., Jacob, L., & Hoogerheide, V. (2021). Learning by writing explanations: Is explaining to a fictitious student more effective than self-explaining? *Learning and Instruction*, *74*, Article 101438. https://doi.org/10.1016/j.learninstruc.2020.101438

McDaniel, M. A. (2023). Combining retrieval practice with elaborative encoding: Complementary or redundant? *Educational Psychology Review*, *35*(3), Article 75. https://doi.org/10.1007/s10648-023-09784-8

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nickl, A. T., & Bäuml, K.-H. T. (2023). Retrieval practice reduces relative forgetting over time. *Memory*, *31*(10), 1412–1424. https://doi.org/10.1080/09658211.2023.2270735

O'Day, G. M., & Karpicke, J. D. (2021). Comparing and combining retrieval practice and concept mapping. *Journal of Educational Psychology*, *113*(5), 986–997. https://doi.org/10.1037/edu0000486

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

Pan, S. C., & 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., & Sana, F. (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. *Journal of Experimental Psychology: Applied*, *27*(2), 237–257. https://doi.org/10.1037/xap0000345

Rawson, K. A., & Dunlosky, J. (2016). How effective is example generation for learning declarative concepts. *Educational Psychology Review*, *28*(3), 649–672. https://doi.org/10.1007/s10648-016-9377-z

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

Richter, T., & Schnotz, W. (2018). Textverstehen [Comprehension of texts]. In D. H. Rost, J. R. Sparfeldt, & S. R. Buch (Eds.), *Handwörterbuch Pädagogische Psychologie* (pp. 826–837). Beltz.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction*, *49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

Roelle, J., Endres, T., Abel, R., Obergassel, N., Nückles, M., & Renkl, A. (2023). Happy together? On the relationship between research on retrieval practice and generative learning using the case of follow-up learning tasks. *Educational Psychology Review*, *35*(4), Article 102. https://doi.org/10.1007/s10648-023-09810-9

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022). Sequence matters! Retrieval practice before generative learning is more effective than the reverse order. *Learning and Instruction*, *80*, Article 101634. https://doi.org/10.1016/j.learninstruc.2022.101634

Roelle, J., & Nückles, M. (2019). Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology*, *111*(8), 1341–1361. https://doi.org/10.1037/edu0000345

Roelle, J., & Renkl, A. (2020). Does an option to review instructional explanations enhance example-based learning? It depends on learners' academic self-concept. *Journal of Educational Psychology*, *112*(1), 131–147. https://doi.org/10.1037/edu0000365

Roelle, J., Schweppe, J., Endres, T., Lachner, A., von Aufschnaiter, C., Renkl, A., Eitel, A., Leutner, D., Rummer, R., Scheiter, K., & Vorholzer, A. (2022). Combining retrieval practice and generative learning in educational contexts: Promises and challenges. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *54*(4), 142–150. https://doi.org/10.1026/0049-8637/a000261

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

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: A field experiment. *Frontiers in Psychology*, *10*, Article 463. https://doi.org/10.3389/fpsyg.2019.00463

Schroeder, N. L., Nesbit, J. C., Anguiano, C. J., & Adesope, O. O. (2018). Studying and constructing concept maps: A meta-analysis. *Educational Psychology Review*, *30*(2), 431–455. https://doi.org/10.1007/s10648-017-9403-9

Scicovery GmbH. (2023). *Labvanced: Professional online experiments made easy*. https://www.labvanced.com/index.html

Sibley, L., Fiorella, L., & Lachner, A. (2022). It's better when I see it: Students benefit more from open-book than closed-book teaching. *Applied Cognitive Psychology*, *36*(6), 1347–1355. https://doi.org/10.1002/acp.4017

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, *47*(1), 513–539. https://doi.org/10.1146/annurev.psych.47.1.513

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001

Weiber, R., & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung—Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS (2. Aufl.)* [Structural equation modeling—A practical introduction to causal analysis using AMOS, SmartPLS and SPSS (2nd ed.)]. Springer. https://doi.org/10.1007/978-3-642-35012-2

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Wong, S. S. H., Lim, K. Y. L., & Lim, S. W. H. (2023). To ask better questions, teach: Learning-by-teaching enhances research question generation more than retrieval practice and concept-mapping. *Journal of Educational Psychology*, *115*(6), 798–812. https://doi.org/10.1037/edu0000802

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

Zamary, A., & Rawson, K. A. (2018). Are provided examples or faded examples more effective for declarative concept learning? *Educational Psychology Review*, *30*(3), 1167–1197. https://doi.org/10.1007/s10648-018-9433-y

Received May 10, 2024
Revision received December 3, 2024
Accepted February 7, 2025 ■