AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

ISSN: 0022-0663

2021, Vol. 113, No. 8, 1587–1608
https://doi.org/10.1037/edu0000499

Multimedia Effects During Retrieval Practice: Images That Reveal the Answer Reduce Vocabulary Learning

Gesa S. E. van den Broek, Tamara van Gog, Evelien Jansen, Mirja Pleijsant, and Liesbeth Kester
Department of Education, Utrecht University

Practicing retrieval of vocabulary items from memory (e.g., with flashcard software or practice tests) is an effective study strategy to remember vocabulary over time. Retrieval practice is often implemented in digital learning environments that increasingly include multimedia (i.e., combining textual and pictorial information). However, it is unknown how multimedia design affects the benefits of retrieval. Therefore, the present study tested the effect of adding images during retrieval practice on students' learning, affective-motivational outcomes, and judgments of learning. We experimentally manipulated the presence and timing of images during retrieval practice of foreign vocabulary in three classroom experiments with students in secondary education. Across experiments, students' vocabulary recall on a posttest (1 to 4 days after practice) was weaker after practice with images that helped them retrieve the answer, compared with practice without images (Experiments 2 and 3) and compared with practice with images that appeared after the retrieval attempt (Experiments 1 and 3). Images enhanced feelings of competence but not enjoyment of practice. The majority of students recognized the negative effects of images on their learning only when the images clearly revealed the answer (Experiment 1) but—incorrectly—considered images that provided partial hints about the answer to be helpful (Experiments 2 and 3). Moreover, students consistently overestimated how much they learned with images that helped them retrieve the answer. During retrieval practice of vocabulary words, informative images are thus potentially harmful and students have limited insight into these effects.

Educational Impact and Implications Statement

Practicing retrieval of vocabulary words from memory (e.g., with flashcard software or practice tests) is an effective strategy to remember the words over time. This study tested how adding images during such retrieval practice influences students' learning and motivation. In three classroom experiments, we found that retrieval practice is less effective when it includes images that provide hints about the answer, compared to no images. Students were unaware of this effect and overestimated how much they learned with images. Multimedia should thus be used cautiously in vocabulary learning software. To ensure that students can later recall vocabulary not only with the help of the images from practice but also without images, practice should not include images that provide hints about the to-be-retrieved answer. Images can, however, be presented as feedback that is shown *after* the learner has given a response.

Keywords: retrieval practice, testing effect, multimedia learning, vocabulary learning, judgments of learning

This article was published Online First October 14, 2021.

Gesa S. E. van den Broek https://orcid.org/0000-0001-7624-0957

Research plans and preliminary findings were presented at a roundtable session and a no-data-session at the EARLI sig2 meeting and the EARLI sig6/7 meeting, 2018.

Gesa S. E. van den Broek, Tamara van Gog, and Liesbeth Kester devised the main conceptual ideas. Evelien Jansen/ Mirja Pleijsant developed the materials for and conducted Experiments 1 and 2 together with Gesa S. E. van den Broek. Gesa S. E. van den Broek conducted the reported analyses. Gesa S. E. van den Broek, Tamara van Gog, and Liesbeth Kester wrote the article; all authors commented on and approved submission of the article.

The authors thank Anne Salimans, Lieke Overvelde, and Suzanne Straver for their help in conducting Experiment 3.

Correspondence concerning this article should be addressed to Gesa S. E. van den Broek, Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands. Email: g.s.e.vandenbroek@uu.nl

Retrieval practice, during which learners actively recall information from memory, enhances knowledge retention in comparison with restudy. For example, it is more effective to practice vocabulary words by repeatedly translating them from memory than by repeatedly reading the words (e.g., Goossens et al., 2014; van den Broek et al., 2014). These benefits of retrieval over restudying, also known as *testing effect,* have been documented in a large number of experiments (for overview studies, see Adesope et al., 2017; Karpicke, 2017; Roediger & Butler, 2011; Roediger & Karpicke, 2006a; Rowland, 2014). In experimental research, retrieval is typically prompted with verbal cues (e.g., a to-be-translated vocabulary word is shown on a computer screen, Carrier & Pashler, 1992; Karpicke & Roediger, 2008). However, educational materials and exercises often include images (Lindner et al., 2016). Multimedia

research has established conditions under which such images are beneficial during *studying* (e.g., reading) of materials (e.g., Butcher, 2014), but it is unclear what effect images have during retrieval practice, and it is likely that image effects are different during retrieval and restudy. For instance, images might act as hints or scaffolds that help learners retrieve information from memory. Whereas this would increase performance on assessment tests (e.g., Lindner, Eitel, et al., 2017; Lindner et al., 2016; Lindner, Lüdtke, et al., 2017; Schneider et al., 2020), it is unclear how images influence learning *from* taking a test (i.e., retrieval practice). Therefore, the present study investigated the effects of images during retrieval practice on learning and on students' affective-motivational outcomes in a series of three experiments, which will be introduced after discussing the relevant literature in more detail.

Potential Positive Effects of Images: Efficient Processing of Multimodal Input and Active Elaboration

Several lines of research suggest that adding images to learning materials can benefit learning, due to characteristics of the images themselves and due to the cognitive processing that images induce in combination with written text. Regarding characteristics of images, decades of memory research have documented a *picture superiority effect,* the finding that "[later] recall is generally higher for items presented as pictures than for items presented as words" (Paivio & Csapo, 1973, p. 176; for an overview of studies, see Ensor et al., 2019). Different mechanisms have been proposed to explain this effect, starting with Paivio's (1991, 2013) dual-coding theory that pictures are more likely than words to be encoded in both verbal and visual representations, and such dual coding leads to superior retention. In addition, images are said to offer a greater variety of unique visual features compared with written text. This may make images physically (e.g., Ensor et al., 2019) and conceptually (e.g., Hamilton & Geraci, 2006) more distinct than written text, and could enhance retention by stimulating elaborative processing during encoding.

Regarding the combination of text and images, according to the *multimedia principle* (Butcher, 2014), learning is enhanced by combining verbal information with pictorial information compared with verbal information alone. Such multimedia effects have been explained by the efficient use of separate processing channels for pictorial and verbal information in human working memory (e.g., according to the cognitive theory of multimedia learning, Mayer, 2014; and classic dual-coding theory, Clark & Paivio, 1991). Moreover, presenting combinations of pictures and text can lead learners to more actively encode the learning materials, increasing efforts to organize relevant information into coherent mental representations and integrate information with prior knowledge (e.g., inducing attempts to integrate verbal and pictorial representations and triggering self-explanations, Ainsworth & Loizou, 2003; Butcher, 2006). In this way, images may help learners to develop more accurate and integrated mental models (Butcher, 2006, 2014; Fiore et al., 2003).

The present study focuses on vocabulary learning, for which multiple studies have reported benefits of studying with (verbal explanations and) images over study without images (e.g., Akbulut, 2007; Andrä et al., 2020; Hald et al., 2016; Kim & Gilman, 2008; Shahrokni, 2009; Tonzar et al., 2009; Yeh & Wang, 2013; Yoshii, 2006; for an overview of older studies, see Sadoski, 2005), in line with the multimedia principle. There is also vocabulary research that found no significant multimedia effects (Al-Seghayer, 2001; Boers et al., 2009; Cohen & Johnson, 2011; Dubois & Vial, 2001) or negative effects of images (Acha, 2009). Such negative effects are likely when images do not match the provided verbal information (Rop et al., 2016), but could also occur if images are irrelevant or distracting to learners, acting as a *seductive detail* (e.g., Harp & Mayer, 1998; Kalyuga & Sweller, 2014). However, overall, prior research suggests that adding relevant images during encoding of new vocabulary has small but positive effects, especially for abstract words (Farley et al., 2012, 2014; Shen, 2003).

The research reviewed so far suggests that adding images during vocabulary practice is potentially beneficial. However, prior studies have focused on effects of images during *encoding*, that is, during (re)study and reading tasks. It is an open question how images influence learning from *retrieval* practice. It is possible that images could also trigger multimodal or elaborative processing during retrieval practice, and thereby enhance learning. Indeed, elaborative processing has been put forward as one possible explanation for benefits of retrieval practice (e.g., Carpenter, 2009, 2011; Carpenter & Delosh, 2006). However, at the same time, images may also change the dynamics of the retrieval process itself. This likely influences learning outcomes, but not necessarily in a positive way.

Potential Negative Effects of Images on Learning From Retrieval: Increased Retrieval Success, Reduced Mental Effort, and Risk of Context-Dependency

One consequence of adding images to retrieval practice is that the images might make retrieval practice easier and increase retrieval success, by providing hints about the to-be-retrieved answer. Although retrieval practice is known to only be beneficial for learning when learners gain access to the correct answer, learners can do so either by retrieving the answer successfully from memory or by studying feedback after failed retrieval attempts (Kornell et al., 2011). Retrieval success itself has only limited impact on learning when feedback is available (Kornell et al., 2015; Rowland, 2014). Therefore, it is questionable if enhanced retrieval success through the addition of images would be beneficial for learning. On the contrary, providing hints about the answer could have negative effects on learning if the hints reduce the depth of processing and the amount of effort needed to retrieve the correct answer.

When images reduce effortful retrieval processes by providing cues about the answer, this is likely to have negative effects on learning. Indeed, major cognitive accounts of retrieval practice have been characterized as "effortful retrieval theories" (Rowland, 2014, p. 1434) because they propose that the effective component of retrieval practice is the (effortful) mental search in memory to retrieve the answer. This effortful search is thought to change memory representations and facilitate later recall. The specific cognitive processes differ between accounts—some accounts propose that the search in memory leads to an elaboration of semantic networks, which leads to incorporation of related information that can later mediate recall (Carpenter, 2009, 2011; Carpenter & Delosh, 2006; Carpenter & Yeung, 2017) whereas other accounts emphasize the reinstatement of the retrieval context and an increasing reduction of the search set of possible answers (Lehman et al., 2014; Whiffen & Karpicke, 2017; see also Dikmans et al.,

2020; Rickard & Pan, 2018). However, all accounts build on empirical findings that more effortful (yet successful) retrieval is more beneficial than less effortful retrieval (Karpicke, 2017). For instance, successfully retrieving words after a longer delay is more beneficial than after a shorter delay (e.g., Pavlik & Anderson, 2008); and successful retrieval with the help of weak cues like "*s* _ _ _ _ _" is more beneficial than retrieval with strong cues that give away much of the answer, such as "*s t r e* _ _" (correct answer: *street*; Carpenter & Delosh, 2006; see also Finley et al., 2011). This negative effect of cues that make retrieval practice easier has been attributed to a reduction of beneficial effortful retrieval processes (Pyc & Rawson, 2009). Images could similarly function as cues that make it easier to retrieve an answer, and thereby reduce retrieval effort and thus learning.

Presenting images during retrieval practice might be particularly problematic when the images cause *context-dependency* (S. M. Smith & Handy, 2014, 2016). That is, images provide additional retrieval cues, and when learners depend on those cues during practice, they may fail at later recall when the images are not available anymore. In two earlier studies, learners who practiced retrieving the meaning of foreign words that were presented on an informative background image (e.g., an airport when the to-be retrieved word meaning was *pilot*), showed lower recall on a later test without background images, compared with learners who practiced the words with uninformative images that did not provide retrieval cues (S. M. Smith & Handy, 2016). Thus, informative images may act as "contextual crutches" (by providing cues) that facilitate retrieval practice but hamper later recall because learners become dependent on the cues in the images.

Similarly to context-dependency in vocabulary learning, images have also been found to impair young learners' reading fluency development. For instance, Torcasio and Sweller (2009) found that students who practiced reading an illustrated text of which the images provided information about the text content, showed lower proficiency on a later reading test without illustrations than students who read the text without images or with uninformative images. Similarly, a number of studies published in the 1960s and 1970s (reviewed in Schallert, 1980) showed that beginning readers were better at reading aloud single, written words when they had previously practiced reading with only the written words than when they had practiced reading the written words while also seeing a picture of the word meaning. Samuels (1970) referred to a *principle of least effort* to explain these findings, arguing that during practice, learners focus on whichever cue helps them respond with the least effort (in this case, the illustration rather than the written words) but their later performance deteriorates if that cue is absent. Overall, these results suggest that images that can be used as a substitute for actual practice of word retrieval, might impair learning.

Prior Research on the Effect of Images During Retrieval Practice on Learning

Although it is plausible that images positively influence learning when presented during encoding but negatively influence learning when interfering with effortful retrieval practice, there is very limited prior research available on the effects of images in retrieval practice. Carpenter and Olson (2012) compared performance *during* repeated retrieval in which learners recalled Swahili vocabulary (e.g., *kelb: dog*) when prompted with either images or translations in their native language English, immediately after having encoded the Swahili word in the same way, with images or translations. The image cues enhanced recall of Swahili words in the second and third round of retrieval practice. This finding suggests that images can have beneficial effects on retrieval success *during* practice. However, no information is available about learning outcomes *after* retrieval practice; so it is unclear how images influenced retention over time. Because benefits of retrieval practice tend to manifest over time, it is relevant to investigate effects of images on (delayed) learning outcomes (Kornell et al., 2011; Toppino & Cohen, 2009).

Another line of research has shown that representational images can improve students' performance on assessment tests (i.e., the multimedia effect in testing, Lindner et al., 2016; Martín-SanJosé et al., 2015), especially when the same images were also present in the encoding phase prior to the test (Lindner et al., 2021; see also Schneider et al., 2020) and when the images represent information that is needed to solve the test question (Lindner et al., 2016). For instance, in a recent classroom study in which students answered multiple choice questions about science facts, students were more accurate and showed better test taking behavior (i.e., reduced rapid guessing) when the test contained images that visualized the information in the question text (Lindner, Lüdtke, et al., 2017). Because assessment tests and retrieval practice involve similar cognitive processes (i.e., responding to questions that prompt the recall of knowledge from memory), it is likely that images have similar effects on the accuracy of responses during retrieval practice as during assessment tests. However, again, it is unclear what the effect is of such images on delayed learning outcomes. Only one study on multimedia effects in assessment tests included an immediate second test, a matching task, directly after a multiple choice test with or without images (Lindner et al., 2021). There was no significant effect of adding images during the multiple-choice test on subsequent matching performance, regardless of whether the encoding phase prior to multiple choice testing contained images or not. However, this was an immediate second test and benefits of retrieval practice typically show when the final test to measure learning outcomes takes place after a longer delay (Kornell et al., 2011; Toppino & Cohen, 2009). Therefore, the question remains if there are multimedia effects in retrieval practice, where adding images during retrieval practice would influence the performance on a delayed final test.

Effects of Images Beyond Cognitive Processing: Affective-Motivational Outcomes and Judgements of Learning

Images could change learning from retrieval practice not only by influencing cognitive processes like retrieval effort and elaboration but also by influencing learners' motivation to engage in practice. This is particularly relevant because retrieval practice is well suited to be done outside classroom instruction, as part of self-regulated vocabulary learning. For this to be optimally effective, learners should repeatedly practice retrieval and distribute practice sessions over time (e.g., Rawson et al., 2018). It is therefore important to take into account how the design of retrieval practice (e.g., the addition of images) influences affective-motivational outcomes such as students' enjoyment of practice, feelings of competence and experienced task value, as well as experienced effort and judgments of learning (JOLs), as these might influence

students' willingness to engage in (further) retrieval practice. At present, very little is known about motivational processes in retrieval practice (Kang & Pashler, 2014), besides the fact that learners often underestimate benefits of retrieval and prefer other study strategies (Bjork et al., 2013; Karpicke & Roediger, 2008). However, the *emotional design* of learning materials (Plass & Kaplan, 2016), for example, the choice of colors and shapes, can elevate learners' affective-motivational state and enhance learning (e.g., Mayer & Estrella, 2014; Um et al., 2012; for overviews, see Brom et al., 2018; Plass & Kaplan, 2016). Adding images to learning materials has been shown to induce more positive mood, higher alertness, higher satisfaction, and reduce the perceived difficulty of the learning materials (e.g., Lenzner et al., 2013; Sung & Mayer, 2012). Moreover, higher retrieval success during practice with images could increase pleasant feelings of competence and satisfaction with the correctness of the own responses, which motivate learners to engage in a task (Efklides, 2006). Overall, prior research suggests that adding images to retrieval practice might enhance affective-motivational outcomes.

Images could also change learners' judgments about their own learning, which in turn influence study behaviors (e.g., Finn & Tauber, 2015). In general, studies have shown that learners tend to underestimate the benefits of (repeated) effortful retrieval practice in comparison to other study strategies like restudying (e.g., Karpicke & Roediger, 2008; McCabe, 2011; Roediger & Karpicke, 2006b), and use retrieval tasks to measure their current state of knowledge rather than to practice materials for better long-term retention (e.g., Kornell & Son, 2009; for a review, see Bjork et al., 2013; Rivers, 2020). Learners who do not recognize the general benefits of effortful retrieval practice, are likely unaware that reducing retrieval effort (e.g., by providing hints that make retrieval exercises easier) could reduce their learning. This could result in an overestimation of learning with images in the present study, in particular because learners often hold multimedia heuristics and interpret fluency during practice as an indicator of successful learning. For one, learners hold beliefs about the effects of images on learning, also called *multimedia heuristics* (Serra & Dunlosky, 2010). When asked, many learners self-identify as "visual learners" who prefer materials with a visual component (Mayer & Massa, 2003 in Butcher, 2014) and learners predict their learning outcomes to be higher when instructional materials contain images in addition to written text, compared to text-only materials (e.g., Eitel, 2016; Jaeger & Wiley, 2014; Serra & Dunlosky, 2010). For instance, learners expect that vocabulary items that are paired with images of their meaning are learned better than the same vocabulary items paired with a written translation (Carpenter & Geller, 2020; Carpenter & Olson, 2012, Experiment 4). However, multimedia heuristics are not always accurate and might actually make learners *over*confident in their ability to recall learning materials that contain images (Carpenter & Geller, 2020; Carpenter & Olson, 2012). Retrieval practice with images might thus—incorrectly—appear more effective to learners than practice without images.

A second mechanism besides multimedia heuristics, which could further inflate JOLs about practice with images, is the experience of increased response fluency and reduced mental effort (de Bruin et al., 2020; Kirk-Johnson et al., 2019). When learners respond fluently, their JOLs tend to be higher and vice versa, as the time to retrieve a target increases, JOLs decrease (Benjamin et al., 1998; Koriat & Ma'ayan, 2005). Problematically, aspects of practice that increase fluency do not always also increase learning and may even hamper learning (for example, massed repetition in quick succession leads to fluent answers during practice, but produces inferior long-term learning compared with repetition that is distributed over time [Cepeda et al., 2008; see also the idea of desirable difficulties, Bjork, 1994; Yan et al., 2017]). When learners fail to realize that fluency during practice is driven by characteristics of the learning materials (e.g., massed repetition) rather than their own progress, this can lead to overestimation of learning outcomes (Benjamin et al., 1998). Thus, if images increase response fluency and reduce the experience of effort during retrieval practice without improving learning outcomes, this is a second mechanism besides multimedia heuristics, which might cause learners to overestimate themselves when practicing retrieval with images (compared with practice without images).

The Present Study

Prior research has established that retrieval practice is a beneficial study strategy for vocabulary learning (e.g., Barcroft, 2007; Goossens et al., 2014; Nakata, 2017), and that adding images to vocabulary exercises is potentially beneficial (Akbulut, 2007; Hald et al., 2016; Kim & Gilman, 2008; Shahrokni, 2009; Tonzar et al., 2009; Yeh & Wang, 2013; Yoshii, 2006; for an overview of older studies, see Sadoski, 2005). However, there is a lack of information about the effect of combining retrieval practice with images. This constitutes a gap in the literature because images are likely to affect interactive retrieval exercises in different ways than encoding or assessment tasks, which have been addressed in prior multimedia learning research (e.g., Butcher, 2006; Lindner, Lüdtke, et al., 2017). Therefore, the present study tested the effect of adding images during retrieval practice, taking into account students' learning outcomes, affective-motivational outcomes (i.e., enjoyment, perceived competence, task value, mental effort, experienced difficulty, and task preferences) and judgements of learning. In three classroom experiments, of which the first two were carried out in parallel, Dutch secondary education students practiced foreign vocabulary through retrieval practice with or without images (see Figure 1). Different vocabulary materials were used in each experiment because images might affect the retention of concrete words differently from more abstract idioms and words (Farley et al., 2012, 2014; Shen, 2003), and because images might affect retention differently depending on the degree to which retrieval effort is reduced (cf. retrieval-effort theories, Pyc & Rawson, 2009; Rowland, 2014).

First, we tested how adding images to retrieval practice influences learning outcomes (i.e., recall on a delayed posttest). On the one hand, positive effects of images that are found during encoding of vocabulary (i.e., efficient processing of multimodal input and enhanced elaboration) could also occur during retrieval exercises and thereby enhance learning. On the other hand, images could very well have negative effects if they interfere with the core retrieval process, reducing retrieval effort and creating context-dependency. The trade-off between these potential positive and negative effects has not yet been studied. We predict, however, that the overall effect on learning would also depend on *when* images are added to retrieval practice. Negative effects would be expected when images are visible during the retrieval attempt, when they may interfere with effortful retrieval by providing hints. In contrast, potential positive effects of images such as enhanced elaboration are most likely when

Figure 1
Retrieval Practice Conditions

Note. Experiment 1 compared the "images-during-and-after-retrieval" condition and the "images-after-retrieval" condition. Experiment 2 compared "images-during-and-after-retrieval" and "no images" conditions. Experiment 3 compared all three conditions. (The depicted item was used in Experiment 3.) See the online article for the color version of this figure.

learners encode presented images, such as during feedback processing after a retrieval attempt. To foreshadow: Experiment 1 therefore specifically focused on the effect of adding images during the retrieval attempt; Experiment 2 focused on the combined effect of adding images during the retrieval attempt and subsequent feedback processing; and Experiment 3 tested and contrasted both effects.

Regarding affective-motivational outcomes, we expected that images would increase students' feelings of competence and enjoyment and reduce perceived task difficulty, based on previous studies (e.g., Lenzner et al., 2013; Schneider et al., 2016). Third, we expected images to cause higher JOLs and (more) overestimation of the own performance due to learners' reliance on multimedia heuristics and increased response fluency during practice with images, in combination with a general unawareness of the benefits of effortful retrieval.

Experiment 1

Method

Participants

Participants were 80 students in Dutch secondary education (M_{age} = 13.5, *SD* = 1.04; 46 girls, 34 boys). The Netherlands has a tracked system for secondary education; the participating students were enrolled in the two highest tracks ("havo/vwo"), which prepare for higher education at a university of applied sciences or a research university. Six students missed Session 2, resulting in N = 80 for Session 1, and N = 74 for Session 2. One student indicated that she had prior knowledge of Spanish, but her performance was comparable with the other students and excluding her data did not change results, therefore her data was included. The other students reported no prior knowledge.

Materials

Stimuli. Participants studied 20 Spanish words with Dutch translation (e.g., *arbol* = tree, *reloj* = clock), distributed across two lists of ten words that were matched on imageability (determined with Van Loon-Vervoorn, 1985), frequency of the Dutch word (Corpus Hedendaags Nederlands, 2013), and word length in Spanish and Dutch. For each word, a color drawing was selected that depicted the word meaning (e.g., an image of a tree or clock).

Initial Encoding. During the initial encoding prior to retrieval practice, the Spanish words were presented one by one in a random order together with their Dutch translation and an image of the word meaning, for 4 s. This presentation was repeated once, in a different random order.

Retrieval Practice: Images-During-and-After-Retrieval or Images-After-Retrieval. During retrieval practice, the Spanish words were presented one by one, and students were asked to type in the Dutch translation (e.g., *arbol =?)*. Immediately after response submission, feedback was given: A green checkmark indicated a correct response or a red cross indicated an incorrect response. Then the Spanish word was shown with the correct translation and image for 4 s, independent of the correctness of the response. Providing feedback about the correct answer is common in retrieval practice research (cf., Adesope et al., 2017; Rowland, 2014), such feedback allows learners to restudy what they could not retrieve correctly and enhances learning (e.g., Pashler et al., 2005).

There were two different retrieval practice conditions, which differed in the timing of the display of the images (see Figure 1): In the *images-during-and-after-retrieval* condition, the images were shown in the response phase together with the Spanish prompt, so

that students could see them when retrieving the word meaning, and remained visible during the feedback phase. In the *images-after-retrieval* condition, the images were shown only after students had submitted a response, as part of the feedback phase. Each student practiced one list of words with images-during-and-after-retrieval, and one list of words with images-after-retrieval, with counterbalanced assignment.

Posttest Measures of Learning Outcomes. Three different recall tests were administered in Session 2. Students first translated from Spanish to Dutch (testing *recall of word meaning*), then from Dutch to Spanish (testing *recall of word form*), and then again from Dutch to Spanish while in addition to the Dutch word, the image from practice was visible (*recall of word form, prompted with images*). No feedback was given during the tests. To describe learning outcomes, the proportion of correct responses was calculated per test, per practice condition. For the recall of the foreign word form, we calculated two scores: a strict score (where only exact responses counted as correct, e.g., *reloj*) and a lenient score that counted responses with spelling errors as correct when they had an edit distance of two or lower (i.e., no more than two letters had to be added or removed to get to the perfect answer; *relo* or *reljo* instead of *reloj* were counted as correct; cf., van den Broek et al., 2018). For the recall of word meaning, responses with an edit distance below 2 were counted as correct.

Affective-Motivational Outcomes

Mental Effort. The experienced mental effort during practice was measured with two items: "How much effort did you invest in this task?" and "How difficult was the task?" Answers were given on a 7-point Likert scale from 1 = *very little effort* to 7 = *very much effort* and 1 = *very easy* to 7 = *very difficult* (cf. Schmeck et al., 2015). As a more objective indicator of effort during learning, we report response accuracy and reaction times for correct responses during practice.

Enjoyment, Feelings of Competence, Task Value, and Task Preference. Ten items from the Intrinsic Motivation Inventory (Center for Selfdetermination Theory, n.d.) were used to measure students' subjective experience of executing the task, in terms of enjoyment (four items of the intrinsic motivation scale, e.g., "I enjoyed this word learning task"), feelings of competence (four items, e.g., "I am satisfied with my performance during this word learning task"), and task value (two items, e.g., "I think that this task helps me learn new words").[1] The items were presented in an intermixed order; for each item, students responded on a 7-point Likert-scale from 1 = *not true* to 7 = *true*. For statistical analyses, the average response was calculated per subscale, resulting in values ranging from 1 to 7. The subscales all had high internal consistency (feelings of competence after practice with images-during-and-after-retrieval: Cronbach's alpha = .72, all other scales/measurement moments $\alpha > .83$). Finally, as a direct measure of task preference, students were asked at the end of the practice session which of the two retrieval tasks they would use to learn words on a different occasion.

Judgments of Learning. Students made two types of JOL after each block of retrieval practice. First, students were asked to indicate more globally how well they knew the words that they had practiced, on a scale from 1 = not well at all to 7 = very well. Henceforth, we will refer to this as "*general JOL*." Second, students were asked more specifically how many of the 10 practiced words they expected to be able to translate after 4 days (0–10; cf., Baars et al., 2013, 2014). This answer was divided by 10 to calculate the proportion of words that students expected to remember. For brevity, henceforth, we will refer to this measure as "*specific JOL*"—but it should not be confused with item-specific JOLs used in other studies.

Procedure

The two sessions of the experiment were conducted in a classroom at the students' school, where students worked individually at a computer. We used the Gorilla experiment Builder (www.gorilla.sc) to create and host our experiment (Anwyl-Irvine et al., 2020). A teacher and one researcher, who is a certified teacher, were present during the whole session. In the first session, students completed two practice blocks, which each consisted of initial encoding and three retrieval rounds. Per practice block, students first encoded one of the two lists of words. Then, participants engaged in three rounds of retrieval practice of that list either with images-during-and-after-retrieval or with images-after-retrieval. After retrieval practice, students completed ratings of affective-motivational outcomes and JOLs. Next, the same steps (encoding, retrieval, survey ratings) were repeated for the second list of words, in the other retrieval practice condition. In Session 2, four days after Session 1, the students completed the three posttests to measure learning outcomes. Session 1 took about 50 min; Session 2 took about 20 min.

Design and Data Analysis

We experimentally manipulated the within-subjects factor retrieval practice condition (images-during-and-after-retrieval, images-after-retrieval) by counterbalancing the assignment of the two matched word lists to the conditions and the order of conditions. Manipulation checks showed that the effect of the practice condition was comparable independent of assignment to stimulus lists and the order of conditions, therefore counterbalancing groups are combined in the following analyses. The main dependent variables were learning outcomes (the proportion of correct responses on the three recall posttests), affective-motivational outcomes (i.e., effort, difficulty, enjoyment, feelings of competence, task value), and JOLs. These were compared between the two retrieval practice conditions using t-tests for paired samples. In order to describe the accuracy of JOLs, the proportion of words that students expected to remember was compared with the actual proportion of remembered words, using t-tests for paired samples. Bayes factors (BF_{01}) are reported to quantify the evidence for the null hypothesis in case of nonsignificant differences (based on two-sided t-tests with a default Cauchy prior width of $r = .707$). The reported BF_{01} indicates how much more likely the observed data are under the null hypothesis than under the alternative hypothesis that there is a difference between the conditions (e.g., $BF_{01} = 5$ indicates that the data are five times more likely under the null hypothesis). We use verbal classifications to interpret evidence strength (cf. Jeffreys, 1961 in Wetzels & Wagenmakers, 2012). As it is increasingly recommended to use mixed-effects modeling in psycholinguistic research (Baayen et al., 2008), the main analyses were also replicated using mixed logit models with crossed random effects for items and participants (using the glmer function in the lme4 package, Bates et al., 2015; in R, Version

[1] The IMI subscales "experienced control," "effort," and "experienced pressure" were not used in the present study. The IMI provides reliable measures also when only subscales are selected (Center for Selfdetermination Theory, n.d.).

3.1.2; R Core Team, 2014). Unless stated otherwise in the text, all effects of t-tests or factorial analyses of the aggregated data were replicated in the mixed models.

In addition to the main analyses, we conducted two exploratory analyses. First, a repeated measures ANOVA tested whether the effect of the retrieval practice condition on learning outcomes was moderated by students' retrieval success. For this analysis, students were binned into categories of low/medium/high retrieval success, using *M* +/−1 *SD* as cut-off scores. Second, we conducted a mediation analysis to test if the effect of the practice condition on learning outcomes was mediated by mental effort experienced during practice, using the MEMORE macro for two-condition within-participant mediation analysis (Version 2.0, Montoya & Hayes, 2017) with 5,000 bootstrapped samples (percentile bootstrapping) in SPSS. The practice condition was entered as independent variable, mental effort as mediator, and learning outcomes (posttest score) as dependent variable.

The data files of this project can be accessed via https://osf.io/3kb2h/?view_only=b9b7d817af5f4a55b697caec2bcbd335

Results

What is the Effect of Images During the Response Phase of Retrieval Practice on Learning Outcomes?

On all three posttests in Session 2, students scored significantly lower on words practiced in the images-during-and-after-retrieval condition than on words practiced in the images-after-retrieval condition (recall of word meaning, $t_{SpNl}(73) = 8.29$, $p < .001$, $d = 1.01$, 95% CI_{diff} [.19, .31]; recall of word form, $t_{NlSp}(73) = 6.31$, $p < .001$, $d = .74$, 95% CI_{diff} [.12, .23]; recall of word form tested with images, $t_{NlSpImage}(73) = 6.27$, $p < .001$, $d = .72$, 95% CI_{diff} [.11, .21], average scores are presented in Table 1). When spelling errors were scored leniently for word form recall, effects remained significant and large ($d_{NlSp} = .88$; $d_{NlSpImage} = .87$). Descriptive statistics for all outcome measures are included in Figure 2 and Table 1.

What is the Effect of Images During the Response Phase of Retrieval Practice on Affective-Motivational Outcomes?

Mental Effort and Performance During Practice. Students rated both the experienced task difficulty and the amount of invested mental effort during practice significantly lower in the images-during-and-after-retrieval condition than in the images-after-retrieval condition, $t(79) = 10.62$, $p < .001$, $d = 1.63$ (large effect), 95% CI_{diff} [1.63, 2.38] and $t(79) = 9.09$, $p < .001$, $d = 1.27$, 95% CI_{diff} [1.34, 2.09]. Differences in practice performance were in line with these self-reports: Learners gave a significantly higher proportion of correct answers in the images-during-and-after-retrieval than in the images-after-retrieval condition, $t(79) = 11.36$, $p < .001$, $d = 1.59$, 95% CI_{diff} [.20; .28], and were also significantly faster to give correct responses in the images-during-and-after-retrieval condition; $t(79) = -8.20$, $p < .001$, $d = .75$, 95% CI_{diff} [−838, −511][2] (see Table 1 for descriptive statistics). All measures thus indicate that the images-during-and-after-retrieval condition was easier and less effortful than the images-after-retrieval condition.

Enjoyment, Feelings of Competence, Task Value, and Task Preference. Ratings of enjoyment did not differ significantly between the two conditions, $t(79) = 1.72$, $p = .089$, $d = .18$, 95% CI_{diff} [−.04, .56], $BF_{01} = 1.98$ (anecdotal evidence for null hypothesis). Students indicated significantly higher feelings of competence after practice in the images-during-and-after-retrieval condition than in the images-after-retrieval condition, $t(79) = 3.95$, $p < .001$, $d = .53$, 95% CI_{diff} [.38, 1.17]. In contrast, students rated the task value (i.e., the usefulness of the exercise) higher in the images-after-retrieval condition than in the images-during-and-after-retrieval condition, $t(79) = 5.83$, $p < .001$, $d = .81$, 95% CI_{diff} [.89, 1.82]. On the question which word learning task they would prefer when learning words in the future, 62 out of the 80 students chose the more effective images-after-retrieval condition, and only 18 chose the images-during-and-after-retrieval condition.

What is the Effect of Images During the Response Phase of Retrieval Practice on Students' Judgments of Learning (JOLs)?

Students' global judgments of how well they knew the practiced words were positive in both conditions ($M_{ImagesAfterRetrieval} = 4.61$ and $M_{ImagesDuringAndAfterRetrieval} = 4.34$, rated on a scale from 1 = *not good at all* to 7 = *very good*) and there was no significant difference between the conditions, $t(79) = 1.05$, $p = .30$, $d = .16$ (small effect), 95% CI_{diff} [−.25, .80], $BF_{01} = 4.80$ (moderate evidence for null hypothesis). Students' specific JOLs, that is, their predictions of how many items they expected to recall on a delayed test did also not differ significantly between conditions, $t(79) = 1.47$, $p = .144$, $d = .19$, 95% CI_{diff} [−.16, 1.09], $BF_{01} = 2.87$ (anecdotal evidence for null hypothesis). However, there was a difference in the accuracy of the specific JOLs between conditions (which is not surprising given that the actual results on the final test differed between conditions): On average, students *underestimated* their retention in the images-after-retrieval condition, $t(73) = -4.06$, $p < .001$, $d = .50$, 95% CI_{diff} [−.18, −.06] but *overestimated* their retention in the images-during-and-after-retrieval condition, $t(73) = 2.49$, $p = .015$, $d = .35$ (small effect), 95% CI_{diff} [.02, .16].[3] Descriptive statistics of the predicted scores and actual scores are included in Table 1.

Exploratory Analyses

Is the Effect of the Retrieval Practice Condition on Learning Outcomes Moderated by Retrieval Success During Practice? An additional exploratory repeated measures analysis was conducted with students grouped by retrieval success during practice with images-after-retrieval. This was done because benefits of retrieval practice sometimes depend on retrieval success during practice (Jang et al., 2012; Rowland & DeLosh, 2015; van den Broek et al., 2014), and the images enhanced retrieval success. This enhancement of retrieval success might, for instance, be especially beneficial for students who otherwise struggle with the task. However, results showed no such interaction between the practice condition

[2] Mixed modelling analyses of inverse-transformed reaction times (transformation selected based on correlation between observed and expected quantiles, cf. Baayen & Milin, 2015) also showed significantly faster reaction times with images-during-and-after-retrieval than images-after-retrieval.

[3] Analyses were repeated with the square of the difference between predicted and actual score (cf. calibration and absolute calibration scores in Baars et al., 2014). These analyses showed no significant difference between the conditions. On average, the square of the prediction error was 0.08 ($SD = 0.10$; $\sqrt{0.08} = 0.28$) for the no-image condition, and 0.11 ($SD = 0.17$; $\sqrt{0.11}=0.32$) for the images-during-and-after-retrieval condition.

Table 1
Average Learning Outcomes and Affective-Motivational Outcomes per Retrieval Practice Condition in Experiments 1 to 3

Outcome measures per experiment		Images-during-and-after-retrieval		Images-after-retrieval		No-images	
		M	*SD*	*M*	*SD*	*M*	*SD*
Experiment 1							
Learning outcomes (scale 0–1.0)	Recall of word meaning	0.39	0.26	0.64	0.23	na	na
	Recall of word form	0.27	0.21	0.44	0.25	na	na
	Recall of word form, lenient score	0.34	0.24	0.54	0.23	na	na
	Recall of word form, test with images	0.28	0.22	0.44	0.24	na	na
Affective-motivational outcomes							
Experienced mental effort (scale 1–7)	Effort	1.73	1.10	3.44	1.53	na	na
	Difficulty	1.51	0.86	3.51	1.50	na	na
Practice performance	Accuracy	0.95	0.07	0.71	0.19	na	na
	Reaction times (ms)	3,968	867	4,642	919	na	na
Intrinsic motivation (scale 1–7)	Enjoyment	3.51	1.46	3.77	1.46	na	na
	Competence	5.50	1.26	4.72	1.62	na	na
	Task value	4.04	1.94	5.39	1.34	na	na
	Task preference	22.5%		77.5%			
Judgements of learning	General JOL	4.34	1.73	4.61	1.68	na	na
	Predicted score	4.7	2.56	5.16	2.35	na	na
Experiment 2							
Learning outcomes (scale 0–1.0)	Transfer	0.30	0.21	na	na	0.34	0.22
	Recall of idioms	0.59	0.23	na	na	0.71	0.22
	Recall of idioms (test with images)	0.90	0.16	na	na	0.88	0.15
	Cloze task	0.93	0.10			0.91	0.12
Affective-motivational outcomes							
Experienced mental effort (scale 1–7)	Effort	2.89	1.36	na	na	2.98	1.40
	Difficulty	2.74	1.39	na	na	2.79	1.53
Practice performance	Accuracy	0.84	0.13	na	na	0.76	0.18
	Reaction times	7,306	1,370	na	na	7,910	1,398
Intrinsic motivation (scale 1–7)	Enjoyment	4.33	1.55	na	na	4.24	1.56
	Competence	5.04	1.30	na	na	4.73	1.49
	Task value	5.03	1.61	na	na	5.13	1.52
	Task preferences	64%		35%			
Judgements of learning	General JOL	5.47	1.38	na	na	5.22	1.42
	Predicted score	6.42	2.48	na	na	5.97	2.47
Experiment 3							
Learning outcomes	Recall of word meaning	0.40	0.23	0.64	0.25	0.59	0.25
	Recall of word meaning (test with images)	0.93	0.10	0.90	0.15	0.72	0.21
Practice performance	Accuracy	0.82	0.10	0.49	0.20	0.50	0.21
	Reaction times	4,652	935	5,179	998	5,176	897
	Task preference	55%		22%		6%	

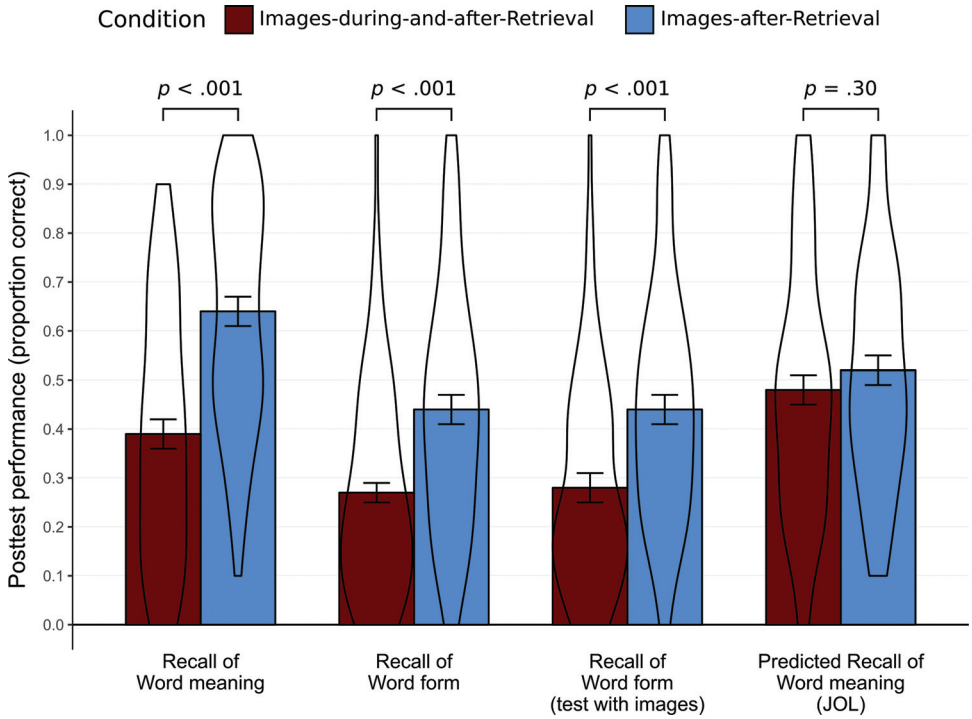
Note. JOL = judgments of learning.

and retrieval success: Students who had higher retrieval success during practice also showed higher recall on the final test, $F(2, 71) = 7.93$, $p = .001$, $\eta^2 = .18$. In addition, students showed lower recall for the words practiced with images-during-and-after-retrieval than with images-after-retrieval, $F(1, 71) = 24.53$, $p < .001$, $\eta^2 = .26$. However, this negative effect of adding images during the response phase of retrieval practice was independent of retrieval success (that is, there was no significant interaction effect, $F(2, 71) = 1.33$, $p = .27$, $\eta^2 = .04$).

To What Extent is the Effect of the Retrieval Practice Condition on Learning Outcomes Mediated by Mental Effort Experienced During Practice? One possible explanation for the effect of the practice condition on learning outcomes is that images-during-retrieval reduced retrieval effort compared to images-after retrieval, which in turn reduced learning (cf. the retrieval effort hypothesis, e.g., Pyc & Rawson, 2009). We therefore tested to what extent the reported mental effort—as a proxy of retrieval effort—mediated the relationship between the practice condition and learning outcomes. All paths in the model were significant, replicating the significant effects of practice condition on learning outcomes ($c = .25$, $p < .001$) and mental effort ($a = 1.68$, $p < .001$) reported in the previous section. The analysis further indicated a significant indirect effect of condition on learning outcomes via mental effort, where the 95% CI did not include zero ($ab = .07$, 95% CI [.01, .16]), as well as a significant direct effect ($c' = .18$, $p < .001$, 95% CI [.10, .27]). Mediation was thus partial; approximately 28% (.07/

Figure 2

Learning Outcomes on the Three Delayed Posttests and JOL Made After Retrieval Practice in Session 1, Split by Retrieval Practice Condition (in Experiment 1)



Note. Bar graphs display mean ± *SE*; the violin plots show a smoothed density curve to show the full distribution of recall scores. JOL = judgments of learning. See the online article for the color version of this figure.

.25) of the effect of the practice condition on learning outcomes can be explained by the mediating effect through mental effort.

Discussion

Students' learning outcomes on a delayed posttest—cued recall of word meaning, and recall of the word form with and without the images from practice—was significantly lower four days after practice with images-during-and-after-retrieval compared with practice with images-after-retrieval, where images appeared after the retrieval (attempt). This negative effect of adding images to the response phase of retrieval practice was independent of students' retrieval success during practice. That is, even students who showed low retrieval success—those who most markedly enhanced their retrieval success when images were added to the response phase—learned less with images-during-and-after-retrieval than with images-after-retrieval. A possible explanation for this finding is that while learning outcomes of retrieval practice sometimes increase with retrieval success during practice (Jang et al., 2012; Rowland & DeLosh, 2015; van den Broek et al., 2014), retrieval success has only limited impact when feedback is available like in the present study (Kornell et al., 2015; Rowland, 2014).

During practice with images-during-and-after-retrieval, students answered correctly more often and faster and experienced less mental effort compared with practice with images-after-retrieval. Exploratory analyses showed that this reduced mental effort partially mediated the effect of the practice condition on learning, suggesting that adding images during retrieval attempts had a negative effect on learning by reducing effortful processing. This finding is in line with retrieval effort theories that propose that making retrieval practice easier reduces learning (e.g., Carpenter, 2009; Carpenter & Delosh, 2006; Coppens et al., 2020; Pyc & Rawson, 2009).[4]

Regarding affective-motivational measures, images-during-and-after-retrieval furthermore increased feelings of competence compared to practice with images-after-retrieval, but did not influence students' enjoyment of practice. Images in the response phase of retrieval practice (accurately) reduced students' ratings of the task value and a majority of the students preferred the more effective images-after-retrieval condition over the images-during-and-after-retrieval condition. Yet, JOLs were less accurate in the images-during-and-after-retrieval condition: Students overestimated how much they learned with images-during-and-after-retrieval and underestimated how much they learned in the more effective practice condition with images-after-retrieval.

Overall, the results of Experiment 1 suggest that the reduction of retrieval effort and the possible creation of context-dependent

[4] Note, however, that the concept of retrieval effort as discussed in the literature (e.g., Rowland, 2014) is different from the broader concept of mental effort (Paas, 1992) that was measured in the present study. To allow stronger conclusions, future research should more directly measure retrieval effort to test the indirect effect of retrieval practice conditions on learning via retrieval effort.

memories through the introduction of images during the response phase of retrieval practice outweighed any potential benefits of elaboration or multimodal processing through extra exposure to images. In addition, images-during-and-after-retrieval did not enhance affective-motivational outcomes besides feelings of competence, and led learners to overestimate their retention.

Experiment 2

Experiment 2 tested how images influence learning of idioms. The basic paradigm was similar to Experiment 1 but we made a number of adjustments that increased the chance of finding benefits of images: Experiment 2 included more complex stimuli (idiomatic expressions) than the concrete nouns used in Experiment 1, images provided more subtle hints, and a transfer test was added. These characteristics of Experiment 2 increase the chance to find benefits of images because images are particularly beneficial for learning abstract words and idiomatic expressions (Farley et al., 2012; Szczepaniak & Lew, 2011), because less informative images might preserve beneficial retrieval effort during practice as they facilitate but do not obviate the retrieval process (cf. the retrieval effort hypothesis, Pyc & Rawson, 2009), and because multimedia effects may be stronger when outcome measures focus on understanding (i.e., transfer tests) rather than retention (Butcher, 2014). In addition, the comparison condition in Experiment 2 contained no images. This differed from Experiment 1, where the comparison condition included images in the feedback phase after the retrieval attempt (the images-after-retrieval condition, see Figure 1). Because of this comparison condition, Experiment 1 focused specifically on the effect of adding images in the response phase of retrieval practice. In contrast, the no-image comparison condition in Experiment 2 allowed us to evaluate the combined effect of adding images in both the response phase (where negative effects may occur due to reduced retrieval effort and increased context-dependency) and the feedback phase of retrieval practice (where the effect of images is predicted to be positive).

Method

Participants

Participants were 135 students from Dutch secondary education schools (M_{age} = 12.30, SD = 0.59, 51 girls, 84 boys). Data were collected in a similar population as in Experiment 1 (i.e., tracks of secondary education which prepare students for university education), during students' English-as-foreign-language classes.

Materials

Stimuli. Participants practiced 20 English idioms, such as "spill the beans" or "once in a blue moon." These idioms were paired with a Dutch explanation (e.g., "to spill the beans" = "een geheim verklappen" [English: "to reveal a secret"]) and a clipart image of the literal meaning of the idiom (e.g., a person spilling a bag of beans). Criteria for the selection of idioms were that there was no comparable idiom in Dutch, and that the literal and figurative meaning of the idiom were understandable for the students, as judged by one of the participating teachers. The idioms were distributed across two lists of 10 items that were matched on idiom/explanation length and that were used for counterbalancing. Manipulation checks showed that List 2 led to lower transfer and retention on the posttest compared with List 1, but to higher scores on the posttest with images if practiced with images. Moreover, performance in the first round of practice was significantly better for List 2 than List 1 if practiced with images (but not in the no-image condition). Possibly, the images of List 2 were more informative but led to lower learning outcomes on the posttests without images. However, the assignment of lists to conditions was counterbalanced and should therefore not have affected the overall within-subject comparison of the two practice conditions. Moreover, the collapsed effects reported hereafter were also tested (and replicated) with a mixed model that controlled for random item effects, suggesting that the reported effects are not due to item differences between conditions.

Initial Encoding. During encoding, the idioms were introduced one by one. Each encoding trial took 18 s: Idioms were first shown with a Dutch explanation for 4 s, then a context sentence was added that illustrated the meaning of the idiom (e.g., "My grandparents live in America. I see them only once in a blue moon"), and after another 10 s an image was added that depicted the literal meaning of the idiom. After 4 s, the next trial started. This encoding procedure was done twice per list of stimuli, in different random order. Encoding trials in Experiment 2 were longer in comparison to Experiment 1 due to the more complex nature of the stimuli (see Table 2 for a complete overview of differences in experimental procedures between experiments).

Retrieval Practice: Images-During-and-After-Retrieval or No-Images. During retrieval practice, students saw a Dutch (their first language) description and typed in the English idiom. In the images-during-and-after-retrieval condition, the Dutch description was shown together with the image (see Figure 1); in the no-image condition, no image was shown, neither during nor after retrieval. There was a time-out after 14 s where the program automatically proceeded to show feedback. After response submission or time-out, corrective feedback was shown as in Experiment 1. There were four rounds of retrieval practice per stimulus list.

Posttest Measures of Learning Outcomes. Four tests were administered in Session 2, two days after the practice session (the delay between practice and posttest differed from Experiment 1 because it was determined based on the participating school's schedule). First, students took a transfer test for which they read a context sentence and had to provide the practiced idiom that fit the context (e.g., "I have to go home and _____. I have an important exam next week." Correct answer: *hit the books*). Second, students took a retention test on which they read the Dutch explanations and typed in the English idiom. On the third test, the Dutch explanations were shown together with the image from the images-during-and-after-retrieval condition, and students again tried to type in the English idiom. Fourth, a cloze test was done in which students completed fragments of the idioms (e.g., "It is raining ___ and ___." Correct answer: *cats, dogs*). To describe learning outcomes, the proportion of correct responses was calculated per test, as in Experiment 1. This resulted in four measures: *Transfer, idiom recall, idiom recall on test with images*, and *idiom completion on cloze test*. Spelling errors were counted as correct answers when they had an edit distance of three or lower.

Table 2
Differences in Experimental Procedure Between Experiments

Methodological aspect	Experiment 1	Experiment 2	Experiment 3
Stimuli	20 concrete Spanish nouns paired with Dutch (L1) translation (e.g., "reloj = clock")	20 English idioms paired with Dutch (L1) definition (e.g., "spill the beans = reveal a secret")	24 abstract French words paired with Dutch (L1) translation (e.g., "hausse = increase")
Encoding	2 encoding rounds, each encoding trial took 4 s; encoding included images	2 encoding rounds, each encoding trial took 18 s; encoding included images and context example of idiom use	2 encoding rounds, each encoding trial took 6 s; encoding did not include images
Retrieval practice	4 rounds of retrieval practice, response: Dutch (L1) translation of Spanish items	4 rounds of retrieval practice, response: English idiom	6 rounds of retrieval practice, response: Dutch (L1) translation of French items
Experimental conditions	Images-during-and-after-retrieval, Images-after-retrieval	Images-during-and-after-retrieval, No-images	Images-during-and-after-retrieval, Images-after-retrieval, No-images
Delay until posttest	4 days after practice	2 days after practice	1 to 3 days (M = 2.1) after practice

Note. L1 = first language.

Ratings of Affective-Motivational Outcomes. The same measures and the same procedure was used as in Experiment 1, with students rating perceived mental effort, task difficulty, enjoyment, feelings of competence, task value, and metacognitive judgments after each practice block.

Results

What is the Effect of Images-During-and-After-Retrieval on Learning Outcomes?

Students' learning outcomes were significantly higher for words practiced without images than for words practiced with images-during-and-after-retrieval on both the transfer test, $t(134) = 2.36$, $p = .02$, $d = .21$, 95% CI_{diff} [.01, .08] and on the idiom recall test, $t(133) = 7.16$, $p < .001$, $d = .53$, 95% CI_{diff} [.09, .15].[5] Idiom recall on the test with images was overall high and not significantly different in the two conditions, $t(134) = -1.1$, $p = .28$, $d = .09$, 95% CI_{diff} [−.04, .01], $BF_{01} = 5.84$ (moderate evidence for H0). On the fourth test, in which participants completed idiom fragments (e.g., "Once in a ____ moon"), performance was higher for idioms practiced with images-during-and-after-retrieval than for idioms practiced without images, $t(134) = 2.24$, $p = .03$, $d = .21$, 95% CI_{diff} [.002, .04].[6] For descriptive statistics, see Figure 3, Table 1.

What is the Effect of Images-During-and-After-Retrieval on Learners' Affective-Motivational Outcomes?

Effect of Images-During-and-After-Retrieval on Mental Effort and Performance During Practice. Ratings of difficulty of practice and invested mental effort did not differ significantly between the two retrieval conditions, $t(134) = .81$, $p = .42$, $d = .06$, 95% CI_{diff} [−.13,.31], $BF_{01} = 7.56$, and $t(134) = .50$, $p = .62$, $d = .04$, 95% CI_{diff} [−.15, .26], $BF_{01} = 9.25$ (see Table 1 for descriptives). However, images influenced performance during practice: Learners gave significantly more correct answers during practice with images-during-and-after-retrieval ($M = .84$) than during practice without images ($M = .76$), $t(134) = 6.10$, $p < .001$, $d = .50$, 95% CI_{diff} [.05, .10]. Furthermore, learners were significantly faster to give correct responses in the condition with images-during-and-after-retrieval, $t(134) = 6.96$, $p < .001$, $d = .44$, 95% CI_{diff} [408, 732].

Effect of Images-During-and-After-Retrieval on Enjoyment, Feelings of Competence, Task Value, and Task Preference. Students gave similar enjoyment ratings in the two retrieval practice conditions, $t(134) = -1.06$, $p = .29$, $d = .06$, 95% CI_{diff} [−.27, .08], $BF_{01} = 6.01$, but indicated higher feelings of competence after practice with images-during-and-after-retrieval than after practice without images, $t(134) = 3.04$, $p = .003$, $d = .22$, 95% CI_{diff} [−.52, −.11]. Ratings for task value did not differ significantly between conditions, $t(134) = .92$, $p = .36$, $d = .06$, 95% CI_{diff} [−.11, .33], $BF_{01} = 6.92$. However, on the questionnaire at the end of Session 1, a majority of 86 students (63.7%) indicated that they preferred the images-during-and-after-retrieval condition, whereas only 49 (36.3%) preferred the no-images condition.

What is the Effect of Images-During-and-After-Retrieval on Students' Judgments of Learning (JOLs)?

Both global JOLs and specific JOLs (predicted recall out of 10 items) were significantly higher after practice with images-during-and-after-retrieval than after practice without images, global JOL: $t(134) = -2.29$, $p = .02$, $d = .18$, 95% CI_{diff} [−.47, −.03]; specific JOL: $t(134) = -2.50$, $p = .01$, $d = .18$, 95% CI_{diff} [−.81, −.09]. The accuracy of JOLs (calculated as: *predicted performance-actual test score*) differed between the two retrieval conditions, $t(133) = -6.68$, $p < .001$, $d = .58$, 95% CI_{diff} [−2.16;− 1.17][7]. After the no-image retrieval practice, students significantly underestimated their later test performance ($M = -1.09$, $SD = 2.80$), $t(133) = 4.49$, $p < .001$, $d = .47$, 95% CI_{diff} [−1.57, −.61]; after

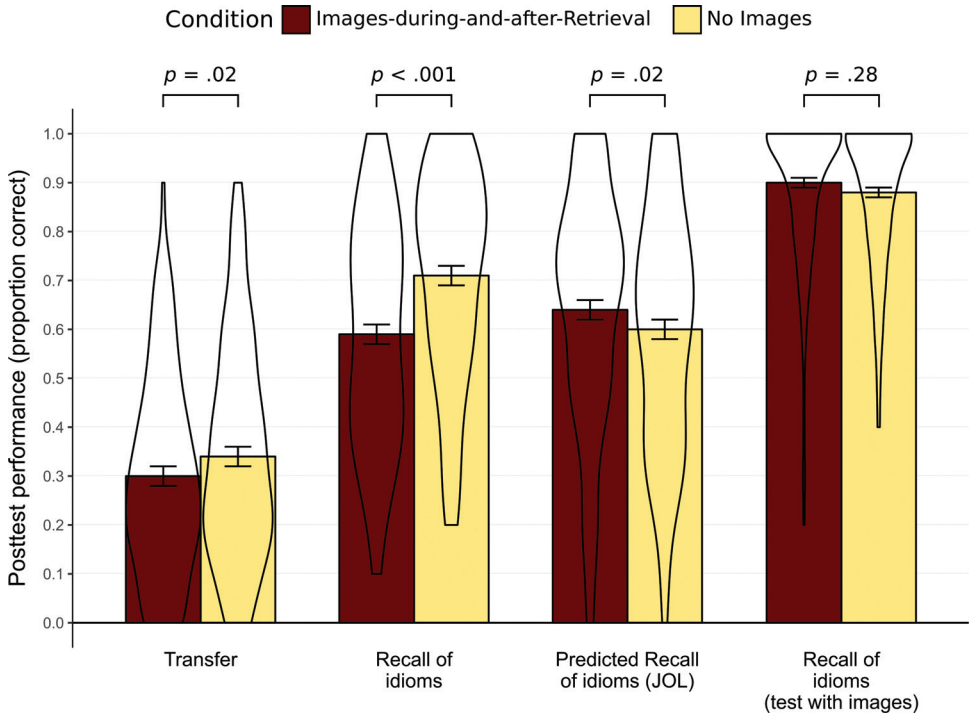
[5] $df = 133$ because the retention test data of one student were not correctly recorded due to a technical problem.

[6] We include the fourth test to provide a complete report of all findings. However, as we elaborate in the discussion, this test is problematic because it followed the recall test with images, which may have introduced a confound in favor of the images-during-and-after-retrieval condition. Therefore, differences between conditions on the fourth test should be interpreted cautiously.

[7] Analyses of the square of the difference between predicted and actual score showed no significant difference between the conditions. On average, the square of the prediction error was 9.02 ($SD = 12.33$, $\sqrt{9.02} = 3.0$) for the no-images condition, and 8.99 ($SD = 12.97$, $\sqrt{8.99} = 3.0$) for the images-during-and-after-retrieval condition.

Figure 3

Learning Outcomes on Three Delayed Posttests and JOL Made After Retrieval Practice in Session 1, Split by Retrieval Practice Condition (in Experiment 2)



Note. Bar graphs show the overall mean ± *SE*; the violin plots show the distribution of scores (for the rightmost bar, the width of the violin plot was determined separately from the other outcome measures, due to the pile-up of high scores). Test 4 is not included, as findings on this test are hard to interpret (see Footnote 5). JOL = judgments of learning. See the online article for the color version of this figure.

practice with images-during-and-after-retrieval, students *overestimated* their retention (M = .57, SD = 3.13), $t(133) = 2.25$, $p = .03$, d = .24, 95% CI_{diff} [.07, 1.08].

Exploratory Analyses

Is the Effect of the Retrieval Practice Condition on Learning Outcomes Moderated by Retrieval Success During Practice?

Students who had higher retrieval success during practice showed higher idiom recall on the posttest, $F(2, 131) = 40.27$, $p < .001$, $\eta^2 = .38$. In addition, idiom recall was higher for idioms practiced without images than with images-during-and-after-retrieval, $F(1, 131) = 27.45$, $p < .001$, $\eta^2 = .173$. This effect of the retrieval practice condition on later recall was independent of retrieval success during practice, $F_{int}(2, 131) = .17$, $p = .84$, $\eta^2 = .003$.

To What Extent is the Effect of the Retrieval Practice Condition on Learning Outcomes Mediated by Mental Effort Experienced During Practice? The mediation analysis indicated that mental effort did not mediate the relationship between the practice condition and learning outcomes in Experiment 2. As in the analyses reported before, the total effect of practice condition on learning outcomes was significant (c = .12, 95% CI [.09, .15], $p < .001$) but there was no effect of practice condition on mental effort (a = .07, 95% CI [−.14, .27], p = .52). The coefficient of the indirect effect of condition on learning outcomes via mental effort was close to 0, and the confidence interval included 0 (ab = .0006; 95% CI [.002, −.003]). The effect of the practice condition on learning outcomes can thus not be explained by a mediating effect through mental effort.

Discussion

In Experiment 2, Dutch students practiced the retrieval of English idioms with and without images. Two days after learning, both their transfer performance—students' ability to produce the appropriate idiom in response to a context description—and idiom recall were better after retrieval practice without images than after practice with images-during-and-after-retrieval. These negative effects of images on learning outcomes strengthen the conclusion from Experiment 1 that adding images to retrieval practice can reduce learning and furthermore show that a negative image effect is obtained also when images are added to both the response and feedback phase of retrieval practice.

Experiment 2 contained a total of four tests. Whereas the first two tests—the transfer test and the retention test in which idiom recall was prompted verbally with a definition—showed negative effects of images during retrieval practice, results were different on the third and fourth test. The third test, in which idiom recall was prompted both verbally and with the images from practice, showed comparable performance in the two retrieval practice conditions. Although findings need to be interpreted with caution because performance was very high on this test (Mdn = 0.9/1.0 for

the two conditions), this is an interesting finding that suggests that retrieval practice with images might specifically put learners at a disadvantage only when images are not available during later recall situations, but not when images are available during recall. A possible explanation for this is that the associations formed during retrieval practice with images incorporate images in such a way that later recall becomes dependent on the images (cf. the context-dependent memory idea, S. M. Smith & Handy, 2016). On the fourth test, the cloze test, participants completed idiom fragments. On this test, performance was also very high, and there was a benefit of the images-during-and-after-retrieval condition. However, the origin of this image effect is unclear because the cloze test was conducted directly after the retention test with images. It is possible that, although performance on the retention test with images was not significantly different between conditions, seeing the images on the third test was a stronger reminder for those idioms that students had practiced with images-during-and-after-retrieval than for idioms practiced without images. The third test may thus have helped students subsequently recognize and complete the idiom fragments on the fourth test. Therefore, results on Test 4 need to be interpreted with caution. Overall, Experiment 2 replicated negative effects of images-during-and-after-retrieval on delayed learning outcomes (both transfer and idiom recall), when learning was measured on a posttest without images.

Regarding affective-motivational outcome measures, images had no effect on the experienced difficulty and mental effort, nor on students' enjoyment ratings and assessment of task value. Unlike in Experiment 1, mental effort also did not mediate the effect of the practice condition on learning.[8] The lack of a main effect of the practice condition on mental effort suggested instead that producing the idioms felt similarly effortful to students with and without images that provided partial hints. However, students expressed higher feelings of competence in the images-during-and-after-retrieval condition, as in Experiment 1, and a majority of students (68% vs. 32%) preferred the condition with images-during-and-after-retrieval. Moreover, as in Experiment 1, JOLs were less accurate in the images-during-and-after-retrieval condition than in the no-image condition: Students predicted that they had learned more and overestimated themselves more after practice with images compared with practice without images.

Experiment 3

Experiment 3 was conducted to test an alternative explanation for the negative effect of images in Experiment 1 and 2, namely, that images might not specifically influence the retrieval process but might have a general negative effect on learning, for example, by distracting learners. Although multiple studies have shown positive effects of images in word learning (Akbulut, 2007; Kim & Gilman, 2008; Shahrokni, 2009; Tonzar et al., 2009; Yeh & Wang, 2013; Yoshii, 2006; for an overview of older studies, see Sadoski, 2005), the images in Experiments 1 and 2 represented similar semantic information as the translations and might therefore be considered partially redundant. Redundant or uninformative images can hamper learning (e.g., Harp & Mayer, 1998; Kalyuga & Sweller, 2014), and this could explain the negative effects of images in Experiments 1 and 2. Experiment 3 tested this alternative explanation by comparing three conditions: images-during-and-after-retrieval, images-after-retrieval, and no-images (see Figure 1). We reason that if negative effects of images are due to reduced retrieval effort and/or context-dependency, the images-during-and-after-retrieval condition should lead to lower learning outcomes than the other two conditions, whereas the images-after-retrieval condition should lead to comparable or higher learning outcomes than the no-images condition. However, if images have a general negative effect, the images-after-retrieval condition should also have lower learning outcomes than the no-image practice condition. An additional change in Experiment 3 concerned the encoding phase prior to retrieval practice. Experiment 1 and 2 included images during initial encoding; Experiment 3 did not include images during initial encoding. This allowed us to test whether the effect of images during retrieval practice was independent of the presence of images during prior encoding, which is informative because prior research showed that multimedia effects in assessment tests may differ depending on the presence of images in prior encoding (cf. Lindner et al., 2021; Schneider et al., 2020).

Method

Participants

Participants were 78 students (M_{age} = 13.8, *SD* = 0.71; 39 girls, 39 boys) from four Dutch secondary education classes (i.e., tracks of secondary education which prepare students for university education).

Materials

Stimuli. Participants practiced translating 24 French words into Dutch. The words were abstract nouns, with an imageability rating below 4 on a 7-point-scale (based on published Dutch norms, Van Loon-Vervoorn, 1985). The French words varied in length between four and 12 characters (*M* = 6.7, *SD* = 2.1), and the Dutch translations between six and 10 characters (*M* = 6.7, *SD* = 1.1). The words were checked by a teacher of one of the participating classes to exclude words that students likely already knew or might have trouble understanding. For each word, a color drawing was selected using Internet resources.

Initial Encoding. There were two encoding rounds in which the words were presented one by one, for 6 s, together with their Dutch translation (in a different random order in each encoding round). Encoding did not include images.

Retrieval Practice: Images-During-and-After-Retrieval, Images-After-Retrieval, or No-Images. Students practiced eight words each with *images-during-and-after-retrieval, images-after-retrieval* or *no-images* (see Figure 1). The words were randomly distributed across the three conditions for each participant and the whole set of 24 words was practiced six times, each time in a different random order. The practice condition was thus experimentally manipulated using interleaving. Additional retrieval practice rounds were added to practice in comparison to Experiments 1 and 2, in order to account for the larger number of stimuli (see Table 2).

Posttest Measures of Learning Outcomes. The test session took place 1 to 3 days (*M* = 2.1, *SD* = .6) after the first session.

[8] This null result should be interpreted with caution because mediation analyses require a relatively large sample size to detect indirect effects, even with recent approaches for within-subject designs (Montoya, 2020).

First, students were asked to translate the 24 French words to Dutch (*cued recall of word meaning*, comparable with Experiment 1). Next, the French words were presented together with the image from practice (*cued recall on test with images*). There was no test of word form recall. No feedback was provided on the test. Spelling errors were counted as correct answers when they had an edit distance of two or lower.

Ratings of Affective-Motivational Outcomes: Students' Task Preference. Due to the interleaved experimental design, there were no separate ratings of affective-motivational outcomes or metacognitive judgments per condition. However, at the end of practice, students answered the following question [translated from Dutch]: "You just practiced some words with images and some without images. Sometimes the images were immediately visible, and sometimes only after you submitted a response. Which way of practice did you like best? (1) Practice without images; (2) practice with images as hints; (3) Practice with images that appeared after you gave an answer; (4) No preference." After the multiple choice question, students were asked to explain their choice in an open answer.

Results

What is the Effect of Images During and After Retrieval on Learning Outcomes?

Cued Recall of Word Meaning. A repeated measures ANOVA showed a significant main effect of practice condition on recall on the first posttest, $F(1.84, 141.49) = 44.04$, $p < .001$, $\eta^2 = .364$, see Table 1 for descriptive statistics. Specifically, recall was significantly lower in the images-during-and-after-retrieval condition compared with the no-image condition, $d = .79$, $p < .001$, and compared with the images-after-retrieval condition, $d = .98$, $p < .001$. The images-after-retrieval condition led to numerically but not significantly higher recall than the no-image condition, $d = .18$, $p = .06$. See Table 1, Figure 4 for descriptives.

Cued Recall of Word Meaning on Test With Images. There was also a significant main effect of the practice condition on performance on the second test, which prompted recall with the images from practice in addition to the Dutch translation, $F(1.65, 126.82) = 69.36$, $p < .001$, $\eta^2 = .474$. On this test, the images-during-and-after-retrieval condition lead to *higher* recall than the images-after-retrieval condition, $p = .005$, $d = .32$, which in turn led to higher recall than the no-image condition, $p < .001$, $d = .93$, see Table 1, Figure 4.

What is the Effect of Images During and After Retrieval on Learners' Affective-Motivational Outcomes?

Performance During Practice. Experiment 3 included no surveys of motivation. The practice condition had a significant effect on practice performance (accuracy: $F(2, 154) = 197.1$, $p < .001$, $\eta^2 = .719$; reaction times for correct answers: $F(1.75, 134.81) = 39.48$, $p < .001$, $\eta^2 = .339$). Pairwise comparisons showed higher accuracy and faster response times in the images-during-and-after-retrieval condition compared to the other two conditions, all $p < .001$. In contrast, the no-image condition and the images-after-retrieval condition did not differ significantly from each other on accuracy ($BF_{01} = 6.98$) nor response times ($BF_{01} = 8.01$).

Task Preferences. Of the 78 students, 43 preferred the images-during-and-after-retrieval condition (55%), 17 preferred the images-after-retrieval condition (22%), five preferred to practice without image (6%), and 13 had no preference (17%). Frequently named explanations for a preference for images were that images or visual thinking improve retention (e.g., "You often remember the image better than the word"; mentioned by 32 students) and that images supported access to the translation during practice (e.g., "Then you had an idea what the answer could be"; mentioned by 21 students). Of the 22 students who preferred images-after-retrieval or no-images, the majority argued that they did not want to associate the response with the image instead of the French word ("[When] practicing with an image as hint, I remember the image instead of the word"; mentioned by 14 students) and five students mentioned that practice with images did not fit vocabulary tests ("Because on the test we also do not get hints"). Students typically referred to "linking" or "thinking about" words and images (e.g., "Because then you need to link the image also to the French word"); only two students mentioned recall processes (e.g., "Because then you [.] need to know the French word to give a correct response"). Of the 17 students who preferred images-after-retrieval, nine students gave multiple reasons (e.g., "Then you do not immediately know the word when you see the image but you do have the image as example that you can think of another time").

Discussion

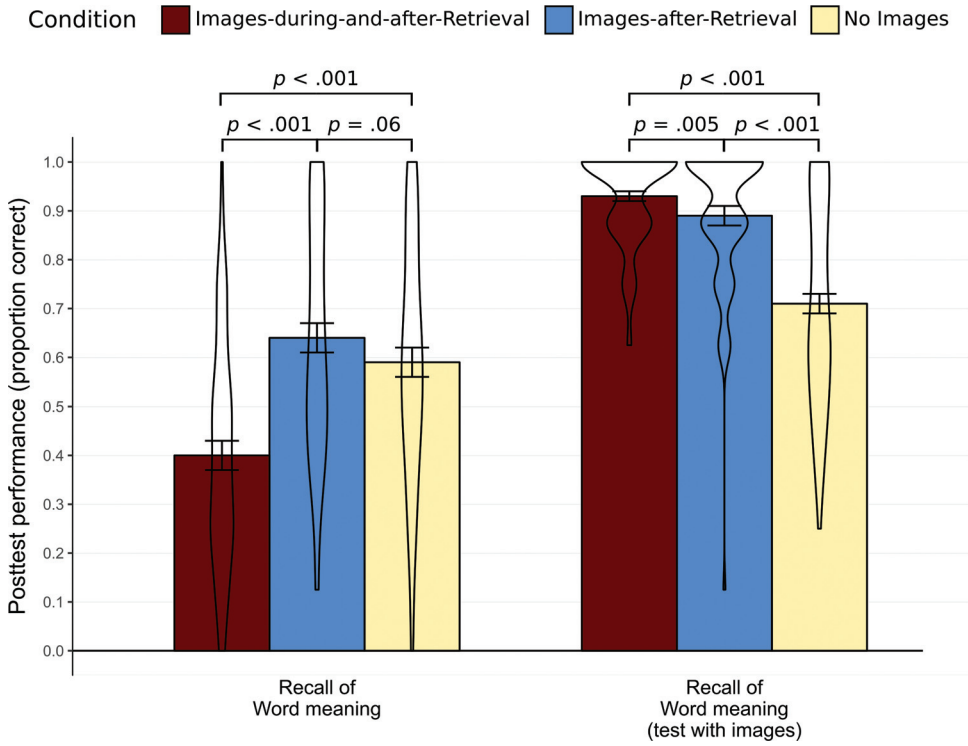
Experiment 3 showed that adding images to retrieval practice had a negative effect on later recall only when the images were visible while learners attempted to retrieve word knowledge from memory but not when images were visible *after* learners retrieved and submitted a response during practice. In fact, adding images *after* the retrieval (attempt) had a numerical positive effect on later recall. This pattern of results suggests that the present study indeed taps into image effects that are specific to retrieval practice, rather than a general negative effect of images on vocabulary learning. Although distractability effects may be to some extent material-specific, this suggests that it is unlikely that negative effects of images on learning outcomes in Experiment 1 and 2 were driven by general effects of distractability.

Experiment 3 also showed that effects of images in retrieval practice depended on whether the final test prompted recall with or without the images from practice: On the posttest that included the images, learners were significantly better when they had practiced with images (during or after retrieval of the answer) than when they had practiced without images. This suggests that recall on the posttest with images was driven by prior exposure to the images: Likely, students became better at producing the word meaning when they had previously seen the (abstract) word paired with a specific image. However, in line with the concept of context-dependency, students were unable to subsequently retrieve the word meaning when the images were not available anymore on the posttest (cf. S. M. Smith & Handy, 2014, 2016).

When asked to choose between the three practice conditions, the majority of the students (55%) preferred images-during-and-after-retrieval, possibly due to (incorrect) beliefs about benefits of visual thinking for retention and a preference for easier practice with images that provide hints. This suggests that, similar to Experiment 2, the majority of the students

Figure 4

Learning Outcomes on the Two Delayed Posttests in Experiment 3, Split by Practice Condition



Note. Bar graphs show the overall mean ± *SE*; the violin plots show the distribution of scores. See the online article for the color version of this figure.

lacked insight into benefits of effortful retrieval practice and/or did not recognize negative effects of images on learning from retrieval. However, there were individual differences between students. A minority of about 28% understood that the images were problematic, and some even included cognitive mechanisms in their explanations that were similar to discussions in the literature (e.g., the possibility that learners may become dependent on images if the images provide hints during retrieval practice; cf., S. M. Smith & Handy, 2016).

General Discussion

Retrieval practice is a well-established, effective learning strategy. By testing the effect of adding images (i.e., multimedia) to retrieval practice, the present study fits into the broader trend of recent research that is beginning to look into combinations of retrieval practice with other instructional principles and learning strategies (e.g., Kubik et al., 2020; Miyatsu & McDaniel, 2019). In three classroom experiments, we tested the effect of presenting images during retrieval practice on students' vocabulary learning. The main findings can be summarized as follows. First, we found a reversed multimedia effect in retrieval practice, where images that provided information about the to-be-retrieved answer during the response phase of retrieval practice reduced learning outcomes consistently across the three experiments. Second, regarding affective-motivational outcomes, images reduced the experienced mental effort and difficulty (Experiment 1) and increased feelings of competence (Experiments 1 and 2), but did not influence how much students reported to enjoy practice. When images provided very strong hints, students preferred the more effective condition without images in the response phase of retrieval practice (Experiment 1). However, when images provided partial hints (Experiment 2) or illustrated abstract words (Experiment 3), a majority preferred practice with images during retrieval, even though that condition was less effective. Third, on average, students did not adjust their judgements of learning to account for the negative effects of images, resulting in more overestimation of their performance in the images-during-and-after-retrieval condition than in the other conditions (Experiments 1 and 2).

Consistent Negative Effects of Images in the Response Phase on Learning Outcomes

During retrieval practice with images available in the response phase (*images-during-and-after-retrieval*), students provided more correct translations and responded faster. However, these images consistently reduced recall on the posttest. These negative image effects occurred across a range of vocabulary learning scenario's and were robust against changes in the experimental procedure (see Table 2): Across the three experiments, a negative effect of

images was found both when the encoding prior to retrieval practice included images (Experiments 1 and 2) and when prior encoding did not include images (Experiment 3); with different types of vocabulary learning materials (concrete nouns, idioms, abstract nouns); when students translated into their native language during practice (Experiments 1 and 3) and when they translated into the to-be-learned foreign language (Experiment 2). Moreover, the effect was found on the recall of word meaning (Experiments 1 and 3), recall of the foreign word form (Experiments 1 and 2), and on the ability to produce idioms in an appropriate context (Experiment 2, transfer test).[9] Exploratory analyses furthermore showed that the negative image effects were independent of students' retrieval success during practice: Students with comparably low, average or high performance during practice all showed negative effects of images-during-and-after-retrieval compared with no-images and images-after-retrieval practice.

Images as Retrieval Crutches

The only factor that moderated the effects of images during practice was the presence of images on the final test. In Experiment 2, the images-during-and-after-retrieval condition led to lower recall than the no-images condition on the transfer and retention tests that contained no images but the two conditions performed similarly on the retention test that contained the images from practice. In Experiment 3, the images-during-and-after-retrieval condition led to lower performance than the other two conditions on the test that contained no images, but to highest performance on the test that contained the images from practice. The finding that the presence of images on the *test* moderated the effects of images during retrieval *practice* on test performance in Experiment 3, is in line with the argument that images might act as cues or crutches during retrieval (cf. the idea of context-dependency, S. M. Smith & Handy, 2014, 2016). It appears that seeing images during the response phase of retrieval practice led to associations in memory that specifically supported later recall prompted with images, but not without images. Possibly, this is because during practice, students formed associations between the image and the target response rather than the foreign word or idiom meaning and the target response. In other words, students became better at recognizing and responding to the images but did not benefit from this practice during later recall without images.

A number of findings further support the interpretation that images reduced learning outcomes because they functioned as crutches during retrieval practice. First, Experiment 3 showed negative effects of images on the retention test that prompted recall without images only for the images-during-and-after-retrieval condition and not for the images-after-retrieval condition (compared with the no-image condition). This suggests that recall was not influenced by mere exposure to images during practice. Rather, recall was specifically driven by the effect of images in the response phase, that is, during the retrieval (attempt). Second, our results suggest that images which provided stronger cues during practice resulted in weaker learning outcomes. For one, the negative effects of images were larger in Experiment 1, in which the images provided stronger hints about the response, compared with Experiments 2 and 3, in which images were less guiding. Moreover, the manipulation check in Experiment 2 suggested that one list of idioms had less informative images and this list of idioms was remembered *better* after retrieval practice with images-during-and-after-retrieval than the other list of idioms.

Retrieval Effort

An additional, compatible interpretation of the negative effects of adding images to the response phase of retrieval practice is that images reduced retrieval effort, a central concept in cognitive accounts of retrieval practice (e.g., Bjork, 1994; Carpenter, 2011; Pyc & Rawson, 2009; overviews in Karpicke, 2017; Rowland, 2014). Students responded faster and more accurately in the images-during-and-after-retrieval conditions in all three experiments compared to practice conditions in which the retrieval attempt was done without images. In Experiment 1 (though not in Experiment 2), students moreover reported lower mental effort in the images-during-and-after-retrieval condition than in the other practice condition and this reduced effort partially mediated the effect of the practice condition on learning outcomes. Moreover, students in Experiment 3 commented that images-during-and-after-retrieval made it easier to find the correct answer. Overall, these results are in line with earlier claims that more effortful retrieval practice (here: without images) is more beneficial than easier retrieval practice (here: with images available in the response phase), comparable with findings of experiments that reduced effortful retrieval with orthographic cues (e.g., Carpenter & Delosh, 2006), contextual cues (van den Broek et al., 2018), or massed repetition (Karpicke & Roediger, 2007).

Multimodal and Elaborative Processing

As described in the Introduction, it is possible that positive effects of images on learning, as described in memory studies on the picture superiority effect (Ensor et al., 2019; Paivio & Csapo, 1973) and in multimedia research (e.g., Butcher, 2014), could occur not only during encoding tasks but also during retrieval practice. However, if any positive effects of images occurred in the present study, they were completely cancelled out by stronger negative effects of images interfering with retrieval processes during the response phase. Adding images in the feedback phase of retrieval practice did not have negative effects (tested in Experiment 3) but also did not increase learning significantly. Thus, the present study does not provide evidence for positive effects of images in vocabulary learning through retrieval practice, although adding images during feedback processing after the retrieval attempt appears unproblematic and numerical differences suggest that small, positive effects might exist.

[9] Note that we used multiple, consecutive tests in each experiment. We cannot rule out that repeated testing influenced our findings but there was no relation between the size or direction of image effects and the order of tests across experiments. Moreover, we took measures to reduce effects of repeated testing: tests were ordered by the amount of cues provided (e.g., the test that prompted recall with definition came before the test that prompted recall with the definition and an image), and no feedback was given on any of the tests. Moreover, if there was an effect of repeated testing, it is unlikely that this influenced the differences between experimental conditions because testing was done in the same way for all conditions: all items from all conditions were presented on a test before proceeding to the next test.

Effects of Images on Affective-Motivational Outcomes

Besides learning outcomes, this study focused on affective-motivational outcomes that might influence students' willingness to engage in (further) retrieval practice. To summarize the three aspects of motivation that we measured in Experiments 1 and 2, images consistently increased feelings of competence but did not influence enjoyment of retrieval practice, and images reduced the perceived task value in Experiment 1 but increased perceived task value in Experiment 2.

Increased feelings of competence after practice with images-during-and-after-retrieval likely reflected students' experience of being able to give more accurate and faster responses compared to the other practice conditions. This would also be in line with lower perceived mental effort and difficulty reported in Experiment 1 (though not in Experiment 2). However, increased feelings of competence did not go along with higher enjoyment. One possible explanation for this is that while having the images at hand may have led to a feeling of being able to master the specific retrieval task (which the competence ratings focused on), practice may not have given the students the satisfaction or pride of being competent at a *challenging* task because students may have recognized that the images made the task easier (Abuhamdeh & Csikszentmihalyi, 2012). This might also explain why unlike earlier studies (e.g., Lenzner et al., 2013; Sung & Mayer, 2012), we did not find that images generally increased learners' subjective experience of practice.

Students also rated the task value, or usefulness of practice for word learning. In Experiment 1, students accurately rated the images-after-retrieval condition as more valuable than the images-during-and-after-retrieval condition. Moreover, a majority of them preferred the images-after-retrieval condition. This assessment was in line with the actual learning outcomes. In Experiment 2, ratings did not differ between the two practice conditions and a majority of the students preferred the condition with images. Similarly, Experiment 3 (in which we only measured task preferences) showed that most students preferred the images-during-and-after-retrieval condition. Thus, whereas ratings of task value and task preferences suggest that students recognized potential negative effects of the images in Experiment 1—in which images provided very strong hints to the answer—students did not recognize the negative effects in Experiments 2 and 3, in which the images provided weaker cues. Students' open answers in Experiment 3 suggested that this was due to students (incorrectly) assuming that visual processing generally enhances retention, and a preference for practice during which retrieval success is high.

Effects of Images on Judgments of Learning

Students predicted how well they learned from the different practice conditions in Experiment 1 and 2. We had expected these JOLs to be less accurate after practice with images because learners tend to underestimate the benefits of effortful retrieval (e.g., Karpicke & Roediger, 2008; McCabe, 2011; Roediger & Karpicke, 2006b), might hold (inaccurate) multimedia heuristics that learning with images is more effective than learning without images (Carpenter & Olson, 2012; Serra & Dunlosky, 2010), and because learners might infer learning success from higher fluency and accuracy when practicing with images-during-and-after-retrieval. Indeed, JOLs were less accurate with images-during-and-after-retrieval than in the other practice conditions: In both experiments that measured JOLs, students overestimated how much they learned with images-during-and-after-retrieval, whereas they underestimated their learning outcomes in the conditions in which retrieval (attempts) were done without images.

Our findings provide some preliminary pointers about which factors—multimedia heuristics and/or increased fluency and accuracy—drove students JOLs. In Experiment 1, in which both conditions contained images (images-during-and-after-retrieval or images-after-retrieval), the absolute JOLs did not differ between conditions, even though images-during-and-after-retrieval led to higher accuracy and fluency during practice. In Experiment 2, in which the comparison condition contained no images, JOLs were higher in the images-during-and-after-retrieval condition. This suggests that multimedia heuristics may have played a larger role when students made JOLs, compared with perceived fluency and accuracy (fluency and accuracy were increased in both experiments in the images-during-and-after-retrieval condition but did not result in higher JOLs in Experiment 1). Further research would be needed to make stronger claims about the inputs that students used to make their JOLs in retrieval practice with multimedia, as students in Experiment 1 were also generally more negative about the images-during-and-after-retrieval condition, and may have corrected their JOLs to account for differences in task value.

Limitations and Suggestions for Further Research

A number of characteristics of this study need to be taken into account when generalizing conclusions to other learning situations. First, one potential limitation is that the paradigms and stimuli differed across the three experiments (see Table 2). However, this can also be regarded as a strength, as we conceptually replicated the negative effects of images in all three experiments. This strengthens the overall conclusion that image effects are robust and occur across a range of vocabulary tasks. Nevertheless, systematic task manipulations are recommended to establish possible boundary conditions of the effect in future research. In this section, we make a number of recommendations for this purpose. To begin with, all three experiments in the present study employed repeated retrieval practice in which each vocabulary item was practiced multiple times with the same image. It is possible that more variation—for example, a different image per retrieval trial, or a combination of initial practice with images and later practice without images—might reduce learners' dependency on the images and lead to better learning outcomes. Some authors argue that when learners practice retrieval of target information in different contexts, they eventually remember only retrieval cues that occur across contexts (i.e., decontextualization; e.g., Lehman et al., 2014). Therefore, using varying images during retrieval practice might be more beneficial than practice with constant images. On the other hand, a variety of images from which the answer can be derived might further draw learners' attention to the interpretation of the image instead of the association between the foreign word form and its meaning, and might for this reason increase the problem that images become a necessary crutch for recall.

The effect of images might also depend on the type of materials that we used. All three experiments used clipart that contained clues about the target response. It is an open question if adding images during retrieval practice also has a negative effect on learning if the

images do not provide cues about the response but, for example, merely represent the retrieval prompt. The context-dependency interpretation of our findings suggests that images would not have a negative effect if they do not reveal the answer, but this remains to be tested. In addition, it is possible that the type of image might influence learning. However, based on the cognitive processes that we assume to be involved in retrieval, it is unlikely that other types of illustrations would have a different effect, unless, for instance, the images also become more or less informative.

Another interesting avenue for future research is to test the effect of images in the feedback phase of retrieval practice. The focus of the present study was on effects of images which are available during the response phase of retrieval practice (and stay available in the feedback phase). These images-during-and-after-retrieval were compared with a condition which included images only in the feedback phase (Experiments 1 and 3) or a condition which included no images at all (Experiments 2 and 3, see Figure 1). However, the inclusion of a no-images comparison condition made it possible to also study the effects of images-after-retrieval in Experiment 3. This contrast was not significant (though numerically, there was a positive effect of the images which was not statistically significant with $p = .06$). Still, images which do not interfere with the core retrieval process because they are presented during feedback encoding *after* learners make a response, might trigger beneficial elaboration (as would be predicted based on multimedia effects during encoding, e.g., Butcher, 2014). Indeed, studies suggest that (failed) retrieval attempts improve subsequent encoding of feedback (cf. the literature on indirect effects of testing, e.g., Arnold & McDermott, 2013; and studies on test-potentiated learning, e.g., Vestergren & Nyberg, 2014). It is possible that these indirect effects of retrieval practice enhance processing of images in the feedback phase. This is an interesting topic for future research also because presenting images-after-retrieval might be a solution to not interfere with retrieval processes, yet make practice more appealing to learners who prefer materials with visuals.

A final characteristic of the materials that is relevant for the interpretation of our findings is that we focused on relatively simple materials with one clear, correct answer: word-meaning pairs and idioms paired with a definition. Effects of images might be more complex when students need to retrieve a larger amount of information, such that images can act as more subtle scaffolds and there is a larger range of retrieval success during practice (e.g., when students can answer questions partially correct and reach a higher level with the help of images). Under these circumstances, images might also have larger affective-motivational effects. Some prior research suggests that images can improve performance on (multiple choice) tests of complex science materials, both in terms of test-taking behavior (e.g., reduced rapid guessing) and in terms of higher test performance (Lindner, Lüdtke, et al., 2017). It is unclear, however, how such images influence later recall. Scaffolding retrieval practice, for example, by asking separate questions per paragraph of learning materials rather than one open question, does not consistently enhance learning outcomes (e.g., M. A. Smith et al., 2016).

Conclusion and Practical Implications

Learning increasingly involves digital resources that make it possible to add visuals to practice. Language learning applications like duolingo, busuu, and Memrise rank among the top downloaded educational apps for mobile phones (Sensortower, 2020), with several million active users according to company reports (e.g., busuu, 2020; Duolingo, 2020; Memrise, 2020). These applications frequently combine retrieval practice of vocabulary with images. However, the effect of images in retrieval practice is not well understood because multimedia research has focused widely on *encoding* tasks. The present study showed that design principles for an active study strategy like retrieval practice may differ from those for encoding tasks: We found consistent reversed multimedia effects for retrieval practice, where images reduced learning outcomes. In addition, seeing images during retrieval did not make practice more enjoyable, and led students to overestimate their learning outcomes. A practical implication of these findings is that if images are added to retrieval practice, this should be done without interrupting the core retrieval process. For example, images might be shown as feedback *after* the retrieval (attempt) but should not provide hints during the retrieval attempt, to avoid that learners infer the answer from the image instead of retrieving the answer from memory. Moreover, there is a clear need for guidance to enable students to use retrieval practice effectively because a majority of students prefers suboptimal practice conditions when given the choice between retrieval practice with and without images.

References

Abuhamdeh, S., & Csikszentmihalyi, M. (2012). The importance of challenge for the enjoyment of intrinsically motivated, goal-directed activities. *Personality and Social Psychology Bulletin*, *38*(3), 317–330. https://doi.org/10.1177/0146167211427147

Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology*, *40*(1), 23–31. https://doi.org/10.1111/j.1467-8535.2007.00800.x

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

Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, *27*(4), 669–681. https://doi.org/10.1207/s15516709cog2704_5

Akbulut, Y. (2007). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science*, *35*(6), 499–517. https://doi.org/10.1007/s11251-007-9016-7

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, *5*(1), 202–232.

Andrä, C., Mathias, B., Schwager, A., Macedonia, M., & von Kriegstein, K. (2020). Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. *Educational Psychology Review*, *32*(3), 815–850. https://doi.org/10.1007/s10648-020-09527-z

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's

monitoring accuracy. *Applied Cognitive Psychology*, *28*(3), 382–391. https://doi.org/10.1002/acp.3008

Baars, M., Visser, S., van Gog, T., de Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology*, *38*(4), 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2015). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Bjork, R. A. (1994). Memory and meta-memory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Boers, F., Píriz, A. M. P., Stengers, H., & Eyckmans, J. (2009). Does pictorial elucidation foster recollection of idioms? *Language Teaching Research*, *13*(4), 367–382. https://doi.org/10.1177/1362168809341505

Brom, C., Stárková, T., & D'Mello, S. K. (2018). How effective is emotional design? A meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning. *Educational Research Review*, *25*(1), 100–119. https://doi.org/10.1016/j.edurev.2018.09.004

busuu. (2020). *Busuu blog (company website)*. https://blog.busuu.com/

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, *98*(1), 182–197. https://doi.org/10.1037/0022-0663.98.1.182

Butcher, K. R. (2014). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 174–205). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.010

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., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., & Geller, J. (2020). Is a picture really worth a thousand words? Evaluating contributions of fluency and analytic processing in metacognitive judgements for pictures in foreign language vocabulary learning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *73*(2), 211–224. https://doi.org/10.1177/1747021819879416

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 92–101. https://doi.org/10.1037/a0024828

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*(1), 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/BF03202713

Center for Selfdetermination Theory. (n.d.). *Intrinsic Motivation Inventory (IMI)*. Retrieved from https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Cohen, M. T., & Johnson, H. L. (2011). Improving the acquisition of novel vocabulary through the use of imagery interventions. *Early Childhood Education Journal*, *38*(5), 357–366. https://doi.org/10.1007/s10643-010-0408-y

Coppens, L., de Jonge, M., van Gog, T., & Kester, L. (2020). The effect of practice test modality on perceived mental effort and delayed final test performance. *Journal of Cognitive Psychology*, *32*(8), 764–770. https://doi.org/10.1080/20445911.2020.1822366

Corpus Hedendaags Nederlands. (2013). *[Corpus Contemprary Dutch]*. Dutch Language Institute. http://hdl.handle.net/10032/tm-a2-m3

de Bruin, A. B. H., Roelle, J., Carpenter, S. K., & Baars, M. (2020). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educational Psychology Review*, *32*(4), 903–915. https://doi.org/10.1007/s10648-020-09576-4

Dikmans, M. E., van den Broek, G. S. E., & Klatter-Folmer, J. (2020). Effects of repeated retrieval on keyword mediator use: Shifting to direct retrieval predicts better learning outcomes. *Memory*, *28*(7), 908–917. https://doi.org/10.1080/09658211.2020.1797094

Dubois, M., & Vial, I. (2001). Multimedia design: The effects of relating multimodal information. *Journal of Computer Assisted Learning*, *16*(2), 157–165. https://doi.org/10.1046/j.1365-2729.2000.00127.x

Duolingo. (2020). *Language courses (company website)*. https://en.duolingo.com/courses/all

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, *1*(1), 3–14. https://doi.org/10.1016/j.edurev.2005.11.001

Eitel, A. (2016). How repeated studying and testing affects multimedia learning: Evidence for adaptation to task demands. *Learning and Instruction*, *41*(1), 70–84. https://doi.org/10.1016/j.learninstruc.2015.10.003

Ensor, T. M., Surprenant, A. M., & Neath, I. (2019). Increasing word distinctiveness eliminates the picture superiority effect in recognition: Evidence for the physical-distinctiveness account. *Memory & Cognition*, *47*(1), 182–193. https://doi.org/10.3758/s13421-018-0858-9

Farley, A., Pahom, O., & Ramonda, K. (2014). Is a picture worth a thousand words? Using images to create a concreteness effect for abstract words: Evidence from beginning l2 learners of Spanish. *Hispania*, *97*(4), 634–650. https://doi.org/10.1353/hpn.2014.0106

Farley, A., Ramonda, K., & Liu, X. (2012). The concreteness effect and the bilingual lexicon: The impact of visual stimuli attachment on meaning recall of abstract L2 words. *Language Teaching Research*, *16*(4), 449–466. https://doi.org/10.1177/1362168812436910

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7

Fiore, S. M., Cuevas, H. M., & Oser, R. L. (2003). A picture is worth a thousand connections: The facilitative effects of diagrams on mental model development and task performance. *Computers in Human Behavior*, *19*(2), 185–199. https://doi.org/10.1016/S0747-5632(02)00054-7

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*(1), 135–142. https://doi.org/10.1002/acp.2956

Hald, L. A., de Nooijer, J., van Gog, T., & Bekkering, H. (2016). Optimizing word learning via links to perceptual and motoric experience. *Educational Psychology Review*, *28*(3), 495–522. https://doi.org/10.1007/s10648-015-9334-2

Hamilton, M., & Geraci, L. (2006). The picture superiority effect in conceptual implicit memory: A conceptual distinctiveness hypothesis. *The American Journal of Psychology*, *119*(1), 1–20. https://doi.org/10.2307/20445315

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, *90*(3), 414–434. https://doi.org/10.1037/0022-0663.90.3.414

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction*, *34*, 58–73. https://doi.org/10.1016/j.learninstruc.2014.08.002

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quarterly Journal of Experimental Psychology*, *65*(5), 962–975. https://doi.org/10.1080/17470218.2011.638079

Kalyuga, S., & Sweller, J. (2014). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 247–262). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.013

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory & Cognition*, *3*(3), 183–188. https://doi.org/10.1016/j.jarmac.2014.05.006

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory* (Vol. 2, pp. 487–514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Roediger, H. L. I. I. I. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. I. I. I. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kim, D., & Gilman, D. A. (2008). Effects of text, audio, and graphic aids in multimedia instruction for vocabulary learning. *Journal of Educational Technology & Society*, *11*(3), 114–126.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*(1), 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

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

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

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting action into testing: Enacted retrieval benefits long-term retention more than covert retrieval. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *73*(12), 2093–2105. https://doi.org/10.1177/1747021820945560

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science*, *41*(5), 811–831. https://doi.org/10.1007/s11251-012-9256-z

Lindner, M. A., Eitel, A., Barenthien, J., & Köller, O. (2021). An integrative study on learning and testing with multimedia: Effects on students' performance and metacognition. *Learning and Instruction*. *71*(1), 101100. https://doi.org/10.1016/j.learninstruc.2018.01.002

Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and Instruction*, *47*(1), 91–102. https://doi.org/10.1016/j.learninstruc.2016.10.007

Lindner, M. A., Ihme, J. M., Saß, S., & Köller, O. (2016). How representational pictures enhance students' performance and test-taking pleasure in low-stakes assessment. *European Journal of Psychological Assessment*, *34*(6), 376–385. https://doi.org/10.1027/1015-5759/a000351

Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology*, *51*(1), 482–492. https://doi.org/10.1016/j.cedpsych.2017.09.009

Martín-SanJosé, J.-F., Lizandra, M. C. J., Vivó, R., & Abad, F. (2015). The effects of images on multiple-choice questions in computer-based formative assessment. *Digital Education Review*, *28*(1), 123–144. https://doi.org/10.1344/der.2015.28.123-144

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press https://doi.org/10.1017/CBO9781139547369

Mayer, R. E., & Estrella, G. (2014). Benefits of emotional design in multimedia instruction. *Learning and Instruction*, *33*(1), 12–18. https://doi.org/10.1016/j.learninstruc.2014.02.004

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

Memrise. (2020). *Company website*. https://www.memrise.com/about/

Miyatsu, T., & McDaniel, M. A. (2019). Adding the keyword mnemonic to retrieval practice: A potent combination for foreign language vocabulary learning? *Memory & Cognition*, *47*(7), 1328–1343. https://doi.org/10.3758/s13421-019-00936-2

Montoya, A. K. (2020). The power of design: Impact of experimental design outweighs impact of inferential methods on statistical power to detect indirect effects. *PsyArXiv*. https://doi.org/10.31234/osf.io/gqryz

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6–27. https://doi.org/10.1037/met0000086

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary

learning. *Studies in Second Language Acquisition*, *39*(4), 653–679. https://doi.org/10.1017/S0272263116000280

Paas, G. W. C. (1992). Training strategies for attaining transfer of problem solving skills in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press. https://doi.org/10.4324/9781315798868

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, *5*(2), 176–206. https://doi.org/10.1016/0010-0285(73)90032-7

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

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

Plass, J. L., & Kaplan, U. (2016). Emotional design in digital media for learning. In S. Y. Tettegah & M. Gartmeier (Eds.), *Emotions, technology, design, and learning* (pp. 131–161). Academic Press. https://doi.org/10.1016/B978-0-12-801856-9.00007-4

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied*, *24*(1), 57–71. https://doi.org/10.1037/xap0000146

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847–869. https://doi.org/10.3758/s13423-017-1298-4

Rivers, M. L. (2020). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*. Advance online publication. https://doi.org/10.1007/s10648-020-09578-2

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

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

Roediger, H. L. I. I. I., & Karpicke, J. D. (2006b). 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

Rop, G., van Wermeskerken, M., de Nooijer, J. A., Verkoeijen, P. P. J. L., & van Gog, T. (2016). Task experience as a boundary condition for the negative effects of irrelevant information on learning. *Educational Psychology Review*, *30*(1), 1–25. https://doi.org/10.1007/s10648-016-9388-9

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

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

Sadoski, M. (2005). A dual coding view of vocabulary learning. *Reading & Writing Quarterly*, *21*(3), 221–238. https://doi.org/10.1080/10573560590949359

Samuels, S. J. (1970). Effects of pictures on learning to read, comprehension and attitudes. *Review of Educational Research*, *40*(3), 397–407. https://doi.org/10.3102/00346543040003397

Schallert, D. L. (1980). The role of illustrations in reading comprehension. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 503–524). Routledge. https://doi.org/10.4324/9781315107493-27

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114. https://doi.org/10.1007/s11251-014-9328-3

Schneider, S., Nebel, S., Beege, M., & Rey, G. D. (2020). The retrieval-enhancing effects of decorative pictures as memory cues in multimedia learning videos and subsequent performance tests. *Journal of Educational Psychology*. *112*(6), 1111–1127, https://doi.org/10.1037/edu0000432

Schneider, S., Nebel, S., & Rey, G. D. (2016). Decorative pictures and emotional design in multimedia learning. *Learning and Instruction*, *44*(1), 65–73. https://doi.org/10.1016/j.learninstruc.2016.03.002

Sensortower. (2020). *Top Charts: IPhone—U.S. - Education*. https://sensortower.com/ios/rankings/top/iphone/us/education

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, *18*(7), 698–711. https://doi.org/10.1080/09658211.2010.506441

Shahrokni, S. A. (2009). Second language incidental vocabulary learning: The effect of online textual, pictorial, and textual pictorial glosses. *The Electronic Journal for English as a Second Language*, *13*(3), 1–17.

Shen, H.-J. (2003). The role of explicit instruction in ESL/EFL reading. *Foreign Language Annals*, *36*(3), 424–433. https://doi.org/10.1111/j.1944-9720.2003.tb02124.x

Smith, M. A., Blunt, J. R., Whiffen, J. W., & Karpicke, J. D. (2016). Does providing prompts during retrieval practice improve learning? *Applied Cognitive Psychology*, *30*(4), 544–553. https://doi.org/10.1002/acp.3227

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1582–1593. https://doi.org/10.1037/xlm0000019

Smith, S. M., & Handy, J. D. (2016). The crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory*, *24*(8), 1134–1141. https://doi.org/10.1080/09658211.2015.1071852

Sung, E., & Mayer, R. E. (2012). When graphics improve liking but not learning from online lessons. *Computers in Human Behavior*, *28*(5), 1618–1625. https://doi.org/10.1016/j.chb.2012.03.026

Szczepaniak, R., & Lew, R. (2011). The role of imagery in dictionaries of idioms. *Applied Linguistics*, *32*(3), 323–347. https://doi.org/10.1093/applin/amr001

Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning*, *59*(3), 623–646. https://doi.org/10.1111/j.1467-9922.2009.00519.x

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

Torcasio, S., & Sweller, J. (2009). The use of illustrations when learning to read: A cognitive load theory approach. *Applied Cognitive Psychology*, *24*(5), 659–672. https://doi.org/10.1002/acp.1577

Um, E., Plass, J. L., Hayward, E. O., & Homer, B. D. (2012). Emotional design in multimedia learning. *Journal of Educational Psychology*, *104*(2), 485–498. https://doi.org/10.1037/a0026609

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. https://doi.org/10.1080/09658211.2013.831455

van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, *68*(2), 546–585. https://doi.org/10.1111/lang.12285

Van Loon-Vervoorn, W. A. (1985). *Voorstelbaarheidswaarden van Nederlandse woorden: 4600 substantieven, 1000 verbalen 500 adjectieven* [Imageability scores of Dutch words: 4,600 nouns, 1,000 verbs, 500 adjectives]. Swets & Zeitlinger.

Vestergren, P., & Nyberg, L. (2014). Testing alters brain activity during subsequent restudy: Evidence for test-potentiated encoding. *Trends in Neuroscience and Education*, *3*(2), 69–80. https://doi.org/10.1016/j.tine.2013.11.001

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Yan, V. X., Clark, C. M., & Bjork, R. A. (2017). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. M. Lodge, & J. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 61–78) Routledge/Taylor & Francis Group.

Yeh, Y., & Wang, C.-W. (2013). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal*, *21*(1), 131–144. https://doi.org/10.1558/cj.v21i1.131-144

Yoshii, M. (2006). L1 and l2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, *10*(3), 85–101.

Received April 8, 2020
Revision received January 31, 2021
Accepted April 9, 2021 ■