Journal of Experimental Psychology: Applied

Effects of Elaborate Feedback During Practice Tests: Costs and Benefits of Retrieval Prompts

Gesa S. E. van den Broek, Eliane Segers, Hedderik van Rijn, Atsuko Takashima, and Ludo Verhoeven

Online First Publication, April 18, 2019. http://dx.doi.org/10.1037/xap0000212

American Psychological Association

Journal of Experimental Psychology: Applied

1076-898X/19/$12.00

http://dx.doi.org/10.1037/xap0000212

Effects of Elaborate Feedback During Practice Tests: Costs and Benefits of Retrieval Prompts

Gesa S. E. van den Broek
Radboud University and Utrecht University

Eliane Segers
Radboud University

Hedderik van Rijn
University of Groningen

Atsuko Takashima
Radboud University and Max Planck Institute for Psycholinguistics

Ludo Verhoeven
Radboud University

This study explores the effect of feedback with hints on students' recall of words. In three classroom experiments, high school students individually practiced vocabulary words through computerized retrieval practice with either standard show-answer feedback (display of answer) or hints feedback after incorrect responses. Hints feedback gave students a second chance to find the correct response using orthographic (Experiment 1), mnemonic (Experiment 2), or cross-language hints (Experiment 3). During practice, hints led to a shift of practice time from further repetitions to longer feedback processing but did not reduce (repeated) errors. There was no effect of feedback on later recall except when the hints from practice were also available on the test, indicating limited transfer of practice with hints to later recall without hints (in Experiments 1 and 2). Overall, hints feedback was not preferable over show-answer feedback. The common notion that hints are beneficial may not hold when the total practice time is limited.

Public Significance Statement
We compared how well high school students learned vocabulary words from translation exercises (retrieval practice) with different types of feedback. When the total study time was fixed, high school students learned equally well with hints feedback that helped the students recall the correct answer as with standard feedback that directly showed the correct answer. Hints feedback may not prepare students for later recall without hints.

Keywords: feedback, instructional scaffolding, vocabulary learning, retrieval practice, hints

Imagine two high school students, Ann and Bob, practicing Latin vocabulary. Ann asks Bob to translate the word *vestis*. Bob cannot recall the translation, so Ann says, "Think of the word vest!" Suddenly Bob remembers: "Oh, right! Vestis is clothing!" The two students in this fictional example practice the retrieval of words from memory—a practice strategy that a plethora of research has shown to enhance long-term retention (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006; Rowland, 2014). Retrieval practice is particularly effective with feedback that allows learners to correct errors and reexposes them to information that they cannot recall (Finley, Benjamin, Hays, Bjork, & Kornell, 2011; Kornell, Bjork, & Garcia, 2011). Different

Gesa S. E. van den Broek, Behavioural Science Institute, Radboud University, and Department of Education, Utrecht University; Eliane Segers, Behavioural Science Institute, Radboud University; Hedderik van Rijn, Department of Experimental Psychology, University of Groningen; Atsuko Takashima, Behavioural Science Institute, Radboud University, and Neurobiology of language, Max Planck Institute for Psycholinguistics; Ludo Verhoeven, Behavioural Science Institute, Radboud University.

We used an algorithm for spaced repetition in our experiments that is included in similar form in the educational software SlimStampen, for which Hedderik van Rijn is copyright owner. This affiliation has not influenced in any way the design of the experiments nor the report of results.

The article is based on Gesa S. E. van den Broek's doctoral thesis, completed at Radboud University. The research was supported by a grant from the National Initiative Brain & Cognition, Netherlands Organization for Scientific Research [Grant 056-33-014]. The authors thank Paul K. Gerke for technical support. The authors have declared that no competing interests exist.

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. E-mail: g.s.e.vandenbroek@uu.nl

feedback formats exist for this purpose: show-answer feedback (in the literature also known as "knowledge of correct response"—feedback [KCR]; Shute, 2008) presents the correct answer for restudy; more elaborated feedback presents additional explanations or requires the learner to make a new response. Overview studies show that elaborated feedback can lead to better learning outcomes than show-answer feedback (Kluger & DeNisi, 1996; Shute, 2008). However, feedback formats have varied widely across studies and the effect of specific elaborations needs further investigation (Attali & van der Kleij, 2017; Van der Kleij, Feskens, & Eggen, 2015). The present study focuses on one specific element of elaborated feedback, namely hints that create an extra opportunity for students to retrieve the correct answer from memory.

Empirical research regarding the creation of (new) retrieval opportunities as part of feedback is scarce. This is surprising given substantial evidence from prior research that practicing memory retrieval is more beneficial than restudying information (for recent reviews, see Karpicke, 2017; Rowland, 2014). Moreover, providing learners with scaffolds, assistance to perform tasks that they cannot complete on their own, has a long tradition in education (Wood, Bruner, & Ross, 1976). Different research fields thus suggest that it should be more beneficial to give learners hints to retrieve an answer from memory than to just show learners the correct answer for restudy. Yet, only few feedback studies have included hints that help learners respond again (for reviews of feedback research, see Narciss & Huth, 2004; Shute, 2008). As a case in point, the most recent meta-analysis on feedback interventions in computerized learning tasks (Van der Kleij et al., 2015) included only five studies with hints feedback that prompted a new response. Most of these studies focused on complex skills like mathematical operations or reading comprehension (Murphy, 2007, 2010; Narciss & Huth, 2004, in Van der Kleij et al., 2015). Only one dated study included hints that required learners to try again to retrieve factual information from memory (Hall, Adams, & Tardibuono, 1968). This study showed that learners' retention of geographical facts after retrieval practice was not enhanced by hints feedback: Learners who saw orthographic hints and tried again to type in the correct answer after an error (hints feedback) showed similar retention as learners who copied the correct response (copy-answer-feedback).

Three more recent studies reported the effect of hints specifically on the retention of items that learners initially could not retrieve correctly (Finn & Metcalfe, 2010; Kornell, Klein, & Rawson, 2015; Kornell & Vaughn, 2016). These studies drew different conclusions: On the one hand, Finn and Metcalfe (2010) found that participants who could not answer general knowledge questions remembered the correct answer better when they constructed the answer from an increasing number of letters (hints) than when they copied the correct response. On the other hand, Kornell and Vaughn (2016) and Kornell et al. (2015) concluded from a series of experiments that the retention of word-pairs was roughly equivalent when learners received copy-answer-feedback or completed word fragments (hints) after a failed retrieval attempt.

In sum, there is only a limited number of studies on hints feedback, and available studies have produced mixed findings. The application of these findings to instructional design is further hampered by the design of the studies: First, all experiments except Hall et al. (1968) included only a single presentation per item. This is relevant because repeated retrieval practice leads to better learning outcomes than a single retrieval (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011), and repetition could change feedback effects. Hints could, for example, influence subsequent repetitions of an item if they prevent repeated errors. Second, the lack of benefits of hints (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016) could be due in part to the type of hints that were used. Kornell et al. (2015) and Kornell and Vaughn (2016) used orthographic hints that almost always led to a correct response (e.g., *wine—vine__r*). Retrieval is, however, more beneficial, when it requires an effortful mental search for the correct answer (Pyc & Rawson, 2009). Moreover, all previous studies that we found used orthographic hints. A well-known principle in learning research holds that semantic processing leads to better retention than processing of physical features such as orthography (Craik & Lockhart, 1972; Craik & Tulving, 1975). Following this depth-of-processing principle, hints that lead to semantic processing might enhance retention more than the previously used orthographic hints. In sum, prior research leaves open whether hints feedback is beneficial for learning when administered during repeated retrieval and when designed to stimulate effortful retrieval and semantic processing.

The Present Study

The central research question of this study was whether retrieval practice with hints feedback is more efficient for recall several days after practice than retrieval practice with show-answer feedback. We investigated this question in three experiments, using three different types of hints. Students practiced vocabulary words in a foreign language by repeatedly translating the words from memory, a common retrieval activity for high school students. Scheduling of the repetitions was adaptive to learner performance and controlled with a learning system that modeled the memory strength per word based on the history of practice (for an extensive discussion of the system, see Sense, Behrens, Meijer, & van Rijn, 2016). Such adaptive scheduling produces better performance than repetitions in a random order or nonadaptive spacing strategies (e.g., Pavlik & Anderson, 2008) and has been applied successfully in foreign language courses (Lindsey, Shroyer, Pashler, & Mozer, 2014).

The main ways in which the present study goes beyond earlier studies on hints feedback are as follows. First, we manipulated feedback during repeated spaced retrieval. Second, we used different hints than previous studies: In Experiment 1, students received orthographic hints that required effortful retrieval (see Carpenter & Delosh, 2006). In Experiments 2 and 3, the hints focused on semantic processing (Craik & Lockhart, 1972; Craik & Tulving, 1975). Third, the study was conducted in a classroom setting and practice time was controlled, with students practicing for 15 min in each condition. This was done to investigate which form of practice—retrieval with standard show-answer feedback or retrieval with hints feedback—most efficiently uses a limited amount of study time. Previous studies used a fixed amount of trials, but such a design potentially favors hints feedback because hints feedback increases processing times after errors compared to standard feedback (Hall et al., 1968; Hays, Kornell, & Bjork, 2010).

This article is a complete report of all experiments which we have done to compare the effects of hints feedback and show-answer feedback during retrieval practice. We report all measures, conditions, data exclusions, and how we determined our sample size.

Experiment 1

In Experiment 1, we compared learning outcomes after retrieval practice with show-answer feedback and retrieval practice with orthographic hints feedback. The hints feedback was designed to trigger effortful retrieval (by providing only a small fragment of the translation of a new word form) and practice involved repeated presentations of each item. Therefore, we predicted positive effects of hints feedback on later recall in spite of the mixed findings in previous studies (Finn & Metcalfe, 2010; Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016).

In addition to later recall, we investigated how hints feedback changed the use of the available practice time, focusing on the number of trials and errors. We also report the number of practiced words, which depended on the number, speed and accuracy of learners' responses, and is a good overall measure of the rate of acquisition during practice. We expected that, because there is a trade-off between spending time on hints-feedback and spending time on additional repetitions (Hays et al., 2010), hints feedback would reduce the number of completed trials during practice. This could reduce the number of practiced words. However, hints feedback could also reduce the number of errors during practice and increase the chance that students learned from errors, which would increase the number of practiced words.

Method

Participants. A total of 108 students from three Dutch high schools took part in the experiment. The data of 85 students (64.71% female, M_{age} = 14.21 years, SD_{age} = 0.77) were analyzed. Of the 23 discarded data sets, 21 students had incomplete data because they were absent during one of the two sessions or experienced technical problems that led to incomplete or repeated practice blocks; one student did not provide consent to use his data, and one student did not follow instructions. All students were in Grade 8 or 9 of Dutch "havo/vwo" classes (higher secondary and preuniversity classes; these are the highest tracks of the Dutch educational system, see OECD, 2016). Necessary sample sizes were estimated using a power analysis with G-power (Version 3.1, Faul, Erdfelder, Lang, & Buchner, 2007). Assuming an effect size of 0.3 (small), 70 students would be needed to achieve a power of 0.80. Higher actual sample sizes (*n*s for Exp. 1/2/3 = 85/90/88) are due to recruitment of students in classes, and because we anticipated for data loss due to absence or technical issues.

Stimuli. Seventy-two English words were selected from vocabulary lists from the last chapters of two schoolbooks of Grade 9, which had not yet been covered in any of the participating classes. In each feedback condition, students practiced up to 36 of these words (see the Retrieval Practice section).

Design and experimental control. The experiment had a within-subject design with feedback condition (hints feedback or show-answer feedback) as independent variable. There was a practice block of 15 min for each of the feedback conditions; the order of the two practice blocks was counterbalanced across participants. Words were randomly assigned to the two retrieval practice blocks for each student.

Retrieval practice. Practice consisted of one initial study trial per word, followed by several retrieval trials. During the initial study trial, the English word was presented together with the Dutch translation and students retyped the translation. During the subsequent retrieval trials, only the English word was shown and students had to recall and type in the translation (see Figure 1), with a time-out after 60 s.

The number of words that students practiced out of the total of 36 words per condition depended on their performance during practice: The faster and the more accurately students responded, the more words were added to practice, with a maximum of 36 per condition. An adaptive learning system was used to determine the order in which items were presented for each student, using a mathematical model to continuously estimate the accessibility in memory of each practiced word (a proxy for memory strength) based on the number, timing, accuracy, and speed of previous retrievals during the study session (for a detailed description, see Sense et al., 2016). Briefly summarized, the purpose of the learning system is to maximize spacing of repetitions of each word while ensuring a high rate of retrieval success (Sense et al., 2016). This is achieved through stepwise addition of words to practice, and adaptive, mostly increasing spacing between repetitions. Such an approach leads to higher learning outcomes than common flashcard techniques and nonadaptive spacing models (Pavlik & Anderson, 2008; van Rijn, van Maanen, & van Woudenberg, 2009).

For the present study, the learning system (Sense et al., 2016) was used both to determine when to add more words to practice and when to repeat the words in practice. Normally, the system aims to repeat items in such a way that users answer around 70% of retrieval trials correctly during practice. For the present experiments, the model parameters were adjusted to increase the delay between repetitions of words. This made the retrieval more difficult and elicited a higher number of errors, and thus more feedback moments in the limited practice time. With the changed settings, students answered on average 60% of the retrieval trials correctly in Experiment 1. The practiced words appeared on average about six times during the 15-min block (see Table 1 for descriptive statistics per condition). Variations in the number of presentations per word are due to limited total study time; words that were introduced later during practice were repeated less frequently than words introduced earlier. The delay between repetitions of the same word increased over the course of practice but summarized across all trials, words were on average repeated 83 s (median 79.0 s) after the previous presentation.

Orthographic hints feedback and show-answer feedback. Feedback was given on all retrieval trials. In case of a correct answer, the word *correct* was displayed for 600 ms. In case of an empty or incorrect response, corrective feedback was shown that differed between the two experimental conditions. In the show-answer feedback condition, the word and its translation were presented for four seconds with the instruction to "try to remember". In the hints feedback condition, the first and the last grapheme of the response were shown as orthographic hints, with an instruction to try again (see Figure 1). The student could then submit another response, which was followed by the word *correct*

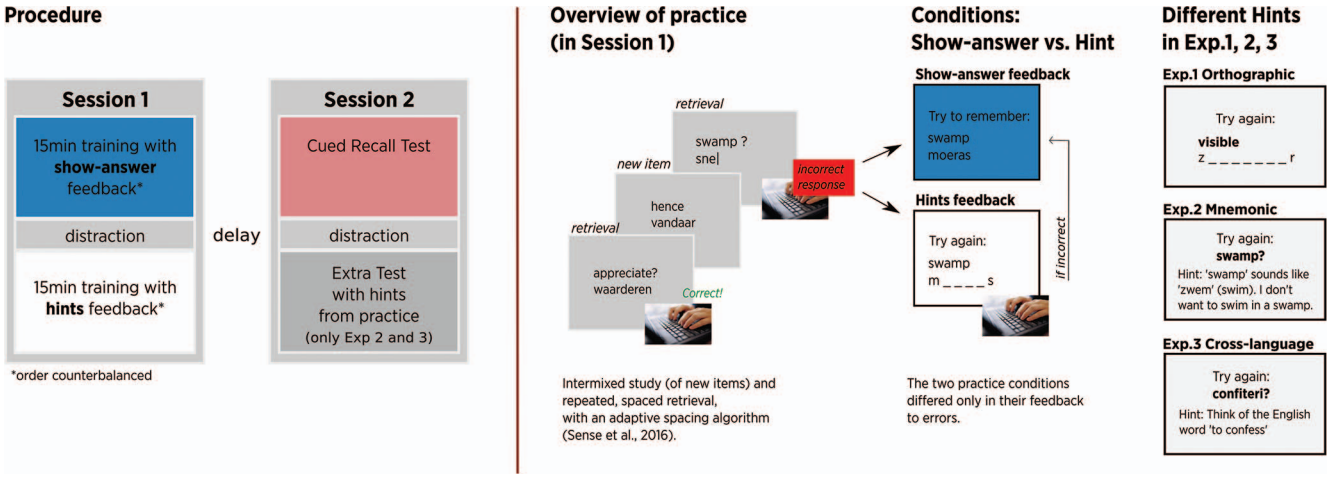

Figure 1. Overview of experimental procedure with feedback conditions. Left: Overview of the two sessions of the experiment with retrieval practice with experimental manipulation of the two different feedback conditions in Session 1, and performance measures with different recall tests in Session 2 several days later. Middle: The vocabulary practice consisted of intermixed studying of items with translations (when a new word was added to practice) and repeated, spaced retrieval. Shown are three English vocabulary items used in Experiment 1. In Experiments 1 and 2, students translated English vocabulary items into Dutch; in Experiment 3, students translated Latin items into German. The two feedback conditions differed only when students made an error: The show-answer feedback immediately revealed the correct answer; the hints feedback gave students a prompt to find the answer. Right: Examples of orthographic, mnemonic, and cross-language hints feedback. The instructions were in Dutch (Experiments 1 and 2) or German (Experiment 3); they were translated for the figure. See the online article for the color version of this figure.

in case of a correct response or by the show-answer feedback in case of an error. There was a time-out after 60 s.

Final recall test. Learning outcomes were measured 7 days after practice with a translation test in which the 72 English words were presented and students typed in the Dutch translation. In case of an incorrect response on the first attempt, the same orthographic hints were shown as during practice in the orthographic-hint condition and students submitted a second response (these hints on the test are called *prompts* in the following, to make a distinction with the *hints* given during practice). At the beginning of the test, students were instructed to try to translate the words as much as possible on the first attempt.

Students were tested on all 72 experimental words in a randomized order. The test thus included not only the words that students had practiced, but also any experimental words that had not been added to practice due to limited study time. These were on average 36.6 (SD = 13.6) unpracticed words per student. We calculated the proportion of the unpracticed words that were translated correctly on the test, as a proxy for the students' prior knowledge. Low test performance for unpracticed experimental words (M = 0.16, SD = 0.13)[1] suggested that students had little prior knowledge of the stimuli in Experiment 1. In addition, the test included 28 easy control words that were selected from the vocabulary lists of the beginners' edition of the students' schoolbook series. The control words were included in the test to ensure that there were at least some trials that the students could answer easily, and to control whether students filled out the test conscientiously. In Experiment 1, the students all recalled at least 67.9% of the easy control words (M = 91.5, SD = 8.6), suggesting that they complied with the given instructions.

Measures of learning outcomes. Responses on the translation test were categorized as either correct or incorrect, with obvious spelling errors (e.g., *sorroww* instead of *sorrow*) counted as correct. The main measure to describe learning outcomes was the number of words translated correctly on the test (short: *overall recall*). We chose the number of words rather than the proportion of practiced words as outcome measure in order compare the actual learning outcomes of the fixed amount of practice in each feedback condition, and to avoid confounding through list-length effects, as learners tend to remember a larger proportion of items from smaller study sets (Gillund & Shiffrin, 1984).

Overall recall on the test was calculated as the number of words that students translated correctly directly on the first attempt (when presented with just the vocabulary word) plus the number of words that students failed to recall on first attempt but then recalled correctly on a second attempt with orthographic prompts. Because successful recall on the first attempt without prompts might indicate stronger memory than recall on a second attempt with prompts, we also report for which proportion of their correct answers, students needed prompts on the test. This is the second outcome measure, short: *need for prompts.* Both overall recall and need for prompts were calculated separately for words practiced with hints feedback and for words practiced with show-answer feedback.

[1] We report the proportion instead of the number of unpracticed words that were translated in order to summarize data across students with different numbers of unpracticed words (range = 4 to 57).

Table 1
Descriptives of the Practice Phase With Show-Answer Feedback or Hints Feedback

Variable	Experiment 1 (n = 85)		Experiment 2 (n = 90)		Experiment 3 (n = 74)	
	Show-answer condition	Orthographic hints-condition	Show-answer condition	Mnemonic hints-condition	Show-answer condition	Cross-language hints-condition
Number of words practiced in 15 minutes	$M = 19.3^{***}$ $SD = 8.1$ $Mdn = 15.0$	$M = 16.5$ $SD = 7.2$ $Mdn = 19.0$	$M = 18.3$ $SD = 7.7$ $Mdn = 16.0$	$M = 20.2^{***}$ $SD = 7.4$ $Mdn = 19.5$	$M = 21.4^{*}$ $SD = 7.9$ $Mdn = 20.0$	$M = 19.5$ $SD = 7.2$ $Mdn = 19.0$
Total number of trials per practice block	$M = 111.0^{***}$ $SD = 27.1$ $Mdn = 112.0$	$M = 97.9$ $SD = 28.0$ $Mdn = 92.0$	$M = 105.6$ $SD = 26.4$ $Mdn = 103.5$	$M = 102.0$ $SD = 26.6$ $Mdn = 99.5$	$M = 109.4^{***}$ $SD = 23.1$ $Mdn = 108.0$	$M = 96.3$ $SD = 22.8$ $Mdn = 94.0$
Number of presentations per word (incl. 1 study trial)	$M = 6.3$ $SD = 1.8$ $Mdn = 5.9$	$M = 6.4$ $SD = 1.7$ $Mdn = 6.1$	$M = 6.3^{***}$ $SD = 1.7$ $Mdn = 5.9$	$M = 5.4$ $SD = 1.2$ $Mdn = 5.0$	$M = 5.6$ $SD = 1.6$ $Mdn = 5.2$	$M = 5.3$ $SD = 1.3$ $Mdn = 5.0$
Average time until next presentation of word (in seconds)	$M = 82.33$ $SD = 15.6$ $Mdn = 79.0$	$M = 83.5$ $SD = 18.6$ $Mdn = 81.9$	$M = 82.0$ $SD = 16.5$ $Mdn = 78.5$	$M = 88.2^{**}$ $SD = 16.3$ $Mdn = 86.7$	$M = 88.74$ $SD = 18.00$ $Mdn = 86.9$	$M = 91.82$ $SD = 22.22$ $Mdn = 84.99$
Number of incorrect responses per word	$M = 2.3$ $SD = 1.7$ $Mdn = 1.8$	$M = 2.5$ $SD = 1.5$ $Mdn = 2.2$	$M = 2.2^{***}$ $SD = 1.5$ $Mdn = 1.9$	$M = 1.4$ $SD = 1.0$ $Mdn = 1.2$	$M = 1.4$ $SD = 1.1$ $Mdn = 1.0$	$M = 1.4$ $SD = 1.1$ $Mdn = 1.1$
Chance for correct response at next presentation of a word, after an error	$M = .48$ $SD = .19$ $Mdn = .47$	$M = .45$ $SD = .21$ $Mdn = .45$	$M = .50$ $SD = .19$ $Mdn = .46$	$M = .64^{***}$ $SD = .22$ $Mdn = .62$	$M = .65^{***}$ $SD = .20$ $Mdn = .64$	$M = .59$ $SD = .22$ $Mdn = .58$
Proportion of correct responses	$M = .62^{*}$ $SD = .17$ $Mdn = .65$	$M = .58$ $SD = .17$ $Mdn = .57$	$M = .62$ $SD = .16$ $Mdn = .62$	$M = .71^{***}$ $SD = .15$ $Mdn = .71$	$M = .73$ $SD = .13$ $Mdn = .75$	$M = .71$ $SD = .16$ $Mdn = .73$
Proportion of correct responses to hints-feedback during practice	n/a	$M = .45$ $SD = .21$ $Mdn = .45$	n/a	$M = .53$ $SD = .18$ $Mdn = .54$	n/a	$M = .69$ $SD = .21$ $Mdn = .75$

Note. The table provides descriptive statistics of the practice phase. All measures were aggregated at participant level and then at group level. The number of presentations, errors, and time until next presentation per word was first aggregated at word level, then at participant level. The asterisks indicate the p-value obtained when comparing the scores in the two feedback conditions with paired t-tests, with $^{*}p < .05$, $^{**}p < .01$, or $^{***}p < .008$. Because of the number of statistical tests, only p values below .008 were considered significant.

Questionnaire measures. Students filled in pen and paper questionnaires in their native language at different moments during both sessions. These questionnaires were mainly used as distractors to introduce short breaks between the computer tasks and to obtain basic demographic information. We also obtained measures regarding students' language skills and preferred vocabulary learning strategies, which are not reported here because they are not directly related to the research questions.

Procedure. The experiment consisted of two sessions (see Figure 1), which were conducted in a classroom setting during the students' regular English lessons. The students worked individually at their computers. During the whole experiment, one or two researchers and the students' English teacher were present to ensure a quiet working atmosphere. Session 1 took 50 min. The session started with a brief group instruction, in which students were informed that they would practice vocabulary words with a computer program that adjusted practice to each student's learning rate. They then filled in a short pen and paper questionnaire and afterward started the first practice block of 15 min by opening a link in the web browser. After the first practice block, students filled in another short pen and paper questionnaire. Then they underwent a second practice block of 15 min. After the second practice block, a third questionnaire was administered and students were told that the researchers would come back for a second practice session seven days later. In Session 2, the students first completed a sustained attention test (from Smilek, Carriere, & Cheyne, 2010), which took 5 min and is not reported here. Afterward, they took the recall test and completed a final questionnaire. The remainder of the second session was spent with debriefing.

Statistical analyses. We tested the effect of the within-subject factor feedback condition (hints feedback or show-answer feedback) on six dependent variables with two-sided *t* tests for paired samples. The dependent variables describing learning outcomes on the final test were overall recall (the total number of words that were recalled correctly in each condition) and the need for prompts during the recall test (the proportion of correctly recalled words which students translated only on a second attempt after receiving prompts). The dependent variables describing the practice phase were the number of trials, the number of practiced words, the average number of errors made per word during practice, and the chance that students learned from errors. This last measure was calculated as the probability that a word was translated correctly on the next trial after it had been translated incorrectly, which was aggregated per learner across all (incorrect) practice trials.

Exact *p* values are reported; to control for the number of statistical tests (six *t* tests in Experiment 1), an adjusted alpha value of $0.05/6 = 0.008$ was used to determine significance; tests with *p* between 0.008 and 0.05 are reported as numerical difference. In addition to classic *t* tests, two-sided Bayesian paired *t* tests (with a default Cauchy prior width of $r = .707$) were used to quantify the evidence for or against the null hypothesis, using the JASP software (Version 0.8.0.0, JASP Team, 2016). To increase readability, we always report the Bayes factor for the alternative hypothesis (BF_{10}). Values of BF_{10} smaller than 1 indicate evidence in favor of the null hypothesis; a BF_{10} larger than 1 indicates evidence in favor of the alternative hypothesis. A BF_{10} of 10 indicates, for example, that the observed data are 10 times more likely under the alternative hypothesis that there is a difference between the conditions, than under the null hypothesis that no difference exists. A BF_{10} of 0.2 indicates that the data are $0.2^{-1} = 5$ times more likely under the null hypothesis than under the alternative hypothesis. We used a verbal classification scheme as proposed by Jeffreys (1961 in Wetzels & Wagenmakers, 2012) to interpret the evidence as "anecdotal" ($1 < BF < 3$), "moderate" ($3 < BF < 10$), "strong" ($10 < BF < 30$), or "very strong" ($BF > 30$). In case of a BF between 0 and 1, the inverse of the BF was calculated before applying this classification scheme.

Results

Descriptive statistics about the practice phase have been included in Table 1; descriptive statistics about recall performance on the final test have been included in Table 2.

Feedback effects on later recall. On the test 7 days after practice, overall recall was not significantly different for words from the practice block with show-answer feedback and for words from the practice block with hints feedback, $t(84) = 1.19$, $p = .24$, $d = 0.08$. A Bayes factor BF_{10} of 0.24 ($BF_{01} = 4.24$) indicated moderate evidence for the null hypothesis. Overall recall on the test was calculated as the sum of the number of words that students recalled correctly directly on the first attempt and the number of words that students failed to recall on first attempt, but then recalled correctly on a second attempt with orthographic prompts. Further analyses of these sub scores revealed that for words from the hints condition, students needed a second attempt with prompts for a larger proportion of the correctly recalled words ($M = 0.39$) than for words from the show-answer condition ($M = 0.25$), $t(80) = -2.90$, $p = .005$, $d = 0.40$.[2] Bayesian *t* tests indicated that the evidence for this effect of feedback on the need for prompts on the test was moderate ($BF_{10} = 5.84$).

Feedback effects during the practice phase. The number of trials that students went through in the 15 min of practice time and the number of words they practiced, were significantly higher in the show-answer condition than in the hints condition, $t(84) = 6.25$, $p < .001$, $d = 0.48$, and $t(84) = 4.22$, $p < .001$, $d = 0.36$. Bayes factors (BF_{10}) of 722435 and 304 indicated very strong evidence for these differences between conditions; see Table 1 for descriptive statistics. The number of errors that students made per word during practice was not significantly different between the two conditions, $t(84) = -1.39$, $p = .17$, $d = 0.15$, nor was the chance that students corrected their errors during practice (i.e., the chance that, after an error, students correctly responded on the next presentation of the same word), $t(84) = 1.84$, $p = .07$, $d = 0.18$. Bayes factors BF_{10} of 0.30 and 0.60 ($BF_{01} = 3.33$ and $BF_{01} = 1.69$) indicated moderate and anecdotal evidence for the null hypotheses.

Discussion

Overall recall performance on the final test was not significantly different after retrieval practice with orthographic hints feedback and retrieval practice with show-answer feedback. Students' test taking behavior indicated, however, that students more often used

[2] Degrees of freedom are 80 instead of 84 because four students recalled zero words from at least one of the two practice blocks, which made it impossible to calculate the proportion of words that were translated with prompts.

Table 2
Learning Outcomes on the Test in Session 2 for Words Practiced With Show-Answer Feedback and Hints Feedback

Outcome	Experiment 1		Experiment 2		Experiment 3	
	Show-answer condition	Orthographic hints condition	Show-answer condition	Mnemonic hints condition	Show-answer condition	Cross-language hints condition
Overall recall						
Total number of words recalled on first attempt or on second attempt with prompts	$M = 7.17$ $SD = 6.74$ $Mdn = 5.0$	$M = 6.67$ $SD = 6.34$ $Mdn = 4.0$	$M = 7.23$ $SD = 6.51$ $Mdn = 5.0$	$M = 7.56$ $SD = 5.92$ $Mdn = 6.0$	$M = 12.82$ $SD = 6.77$ $Mdn = 11.5$	$M = 11.20$ $SD = 6.57$ $Mdn = 11.0$
Need for prompt						
Proportion of overall recall that was translated on second attempt with prompts	$M = .25$ $SD = .30$ $Mdn = .14$	$M = .39^{***}$ $SD = .32$ $Mdn = .33$	$M = .20$ $SD = .22$ $Mdn = .15$	$M = .25^{*}$ $SD = .22$ $Mdn = .22$	$M = .32$ $SD = .19$ $Mdn = .29$	$M = .36$ $SD = .22$ $Mdn = .33$
Number of words recalled on test with hints from practice	n/a	n/a	$M = 7.53$ $SD = 6.67$ $Mdn = 5.0$	$M = 9.89^{***}$ $SD = 6.62$ $Mdn. = 9.0$	$M = 17.26$ $SD = 7.82$ $Mdn = 15.0$	$M = 15.63$ $SD = 7.27$ $Mdn = 15.0$

Note. The table provides descriptive statistics of the learning outcomes on the final test after practice. Overall recall denotes the number of words that were recalled either directly on the first attempt, when students saw a vocabulary word and attempted to type in the translation in their native language, or on the second attempt, after students saw the first and the last grapheme of the correct answer as orthographic prompt. Need for prompt (second row) is the proportion of correct recalls that were given only on the second attempt. In addition, Experiment 2 and 3 contained a separate test on which the hints from practice were presented together with the vocabulary words (results in third row). The asterisks indicate the *p*-value obtained when comparing the scores in the two conditions with paired *t*-tests, marking significantly higher scores at * $p < .05$, ** $p < .01$, or *** $p < .008$. Because of the number of statistical tests, only *p* values below .008 were considered significant.

recall prompts on the test for the words from the hints condition. Measures from the practice phase showed a higher number of trials and a higher number of practiced words in the show-answer condition than in the hints condition, whereas errors during practice were not influenced by the feedback condition.

The lack of feedback effects on the retention of vocabulary items is in line with the majority of previous memory studies on hints feedback, which found no differences between orthographic hints and copy-answer feedback (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016). For some of the earlier studies, the lack of effects could have been due to the use of strong hints (Kornell et al., 2015; Kornell & Vaughn, 2016). However, even with orthographic hints that required effortful processing, Experiment 1 did not show benefits of hints feedback compared to show-answer feedback. This result contradicts benefits of orthographic hints feedback reported by Finn and Metcalfe (2010). There are several possible explanations. One is that Finn and Metcalfe used incremental hints that increasingly revealed the correct answer until participants submitted a correct response, whereas the present study included only one fixed hint. This is unlikely to be the only explanation, however, because comparable incremental hints were not effective in Hall et al. (1968). A second explanation is that Finn and Metcalfe presented feedback when participants could not answer general knowledge questions based on their prior knowledge. Participants thus received feedback on information that they may have never learned before, making the hints a means to encode rather than recall the correct answer. This could explain differences with the present study because the same hints can have different effects when administered as part of encoding or as feedback after a failed retrieval (Kornell et al., 2015). Finally, Finn and Metcalfe (2010) did not control for time on task. The total study time was therefore likely longer in the hints feedback condition than in the show-answer condition, whereas it was equal in the two conditions in the present study. Follow-up research could establish how efficient incremental hints are under limited study time.

The finding that students relied more on recall prompts on the final test to translate the words from the hints condition than the show-answer condition was unexpected. One possible explanation is that the hints condition led to overall weaker memories than the show-answer feedback, making it necessary for students to rely more on prompts on the test (see also Halamish & Bjork, 2011, for a discussion how performance on tests of varying difficulty can indicate memory strength). An alternative explanation is that students may have been unable to transfer what they practiced with hints to the test without hints. During practice, students likely focused on the hints to figure out which of the practiced translations fit, rather than on the association of the foreign vocabulary word and its translation. However, such an association is needed for later recall of the word (e.g., Deconinck, Boers, & Eyckmans, 2017). The limited benefits of the hints feedback could thus be an example of context-dependent memory where later recall (here: recall of the translation) becomes dependent on cues that were available during practice (here: orthographic hints; Smith & Handy, 2014, 2016).

The fact that students went through fewer repetitions and practiced fewer words in the hints condition than in the show-answer condition, is in line with earlier experiments in which spending time on elaborate feedback took time away from further repetitions

(Hays et al., 2010). Surprisingly, although students spent extra time processing feedback after errors in the hints condition, they were not more likely to learn from feedback. The chance for a correct response on the next repetition after an error was similar in the two conditions, and so was the average number of errors per word. This suggests that responding to the hints feedback had few benefits, even during practice.

Experiment 2

Orthographic hints like those used in Experiment 1 can be easily automatically generated for computer-assisted learning and have been used as retrieval cues in previous memory studies (e.g., Finley et al., 2011). However, as we argued above, orthographic hints may focus learners' attention too much on the spelling of the response (finding the translation that fits the presented hints) rather than on the association between the vocabulary word and its translation. A more efficient way to strengthen this association might be to trigger semantic processing with so-called *keyword mediators* (Atkinson, 1975). Keyword mediators are an effective technique to encode the link between a vocabulary word and its meaning (e.g., Beaton, Gruneberg, Hyde, Shufflebottom, & Sykes, 2005; see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013, for a critical review). The technique involves two steps. First, the learner chooses a keyword, which is a known word that sounds similar to the new vocabulary. Next, the learner makes a meaningful association between the keyword and the translation, usually by forming a mental image. For example, to remember the meaning of the word *sorrow*, a learner could choose the keyword "Zorro" and create a mental image of Zorro feeling sorrow. Learners can generate keywords themselves but benefit in a comparable way from keywords generated by others (Shapiro & Waters, 2005). Therefore, we included experimenter-generated keywords in the hints feedback in Experiment 2.

Experiment 2 had a similar setup as Experiment 1 but the students now received feedback with mnemonic hints instead of orthographic hints during practice and took an additional recall test with mnemonic hints at the end of Session 2. This extra test with mnemonic hints allowed us to investigate whether the hints feedback during practice led to selective benefits on a test with the same hints from practice, as in Experiment 1. Alternatively, mnemonic hints that enforce the link between vocabulary words and translations, could also enhance recall on a test without hints.

Method

The research design, materials and procedure in Experiment 2 were identical to those in Experiment 1, with a few exceptions outlined hereafter. The most important difference between the experiments was that we used mnemonic hints instead of orthographic hints in Experiment 2.

Participants. A total of 120 students from two Dutch high schools took part in the experiment. The data of 90 students (56.7% female, $M_{age} = 13.96$ years, $SD_{age} = 0.76$) were analyzed; the other students had incomplete data sets because they were absent during one of the lessons or experienced technical problems that led to incomplete or repeated study blocks. The students who participated in Experiment 2 had not participated in Experiment 1.

Design and experimental control. All students first practiced in the show-answer condition and then in the hints condition to prevent differential transfer from the block with mnemonic hints to the block without hints. This is common in studies on mnemonic techniques (e.g., Fritz, Morris, Acton, Voelkel, & Etkind, 2007).

Feedback with mnemonic hints. In case of an incorrect response, students in the hints feedback condition were presented with a mnemonic hint in the students' native language Dutch. This hint contained a keyword and a sentence that linked the keyword to the Dutch translation. For example, if a student failed to fill in the Dutch translation of the English word *vain* [Dutch: *ijdel*], the hint was, "*vain* . . . fijn. *Als je er altijd fijn wilt uitzien, dan ben je _____*" ["*vain* . . . pretty. If you always want to look pretty, you are _____"]. The word *fijn* [*pretty*] thus functioned as keyword to link the phonologically similar word form *vain* to its translation *ijdel*. The show-answer condition was identical with Experiment 1.

Test of learning outcomes. The students first took the same translation test as in Experiment 1, to measure overall recall and need for prompts. Afterward, the students took an additional separate test during which the words were presented one by one together with the mnemonic hint used in the practice phase of the hints condition. Performance on this test is called *recall with mnemonic prompts.*

Procedure. The procedure was identical with Experiment 1, but the students did no sustained attention task at the beginning of Session 2 and did an extra recall test with mnemonic prompts at the end of Session 2. Between the first recall test and the extra recall test with mnemonic prompts, students filled in a short questionnaire.

Results

Feedback effects on later recall. The overall number of words that students recalled was not significantly different in the two conditions, $t(89) = -0.65$, $p = .52$, $d = 0.05$, with a Bayesian t test indicating moderate support for the null hypothesis that no difference exists between the conditions, $BF_{10} = 0.14$ ($BF_{01} = 7.14$). The need for recall prompts on the test did not differ significantly between conditions, $t(84) = -2.30$, $p = .02$, $d = 0.28$ (alpha corrected for multiple comparisons = .008), although numerically, students used more orthographic prompts to recall the words from the hints condition. Bayesian analyses indicated that the evidence concerning the (lack of) difference between the conditions on the need for recall prompts was inconclusive, $BF_{10} = 1.44$.

The two feedback conditions were also compared on recall on an additional separate test with mnemonic prompts. This recall measure was higher in the hints condition than in the show-answer condition, $t(87)$[3] $= -4.50$, $p < .001$, $d = 0.35$. A Bayesian t test showed that the evidence for this difference between the feedback conditions was very strong, $BF_{10} = 809.80$.

Feedback effects during the practice phase. The number of trials that participants went through was approximately the same in the two conditions, $t(89) = 1.74$, $p = .09$, $d = 0.13$. However, students practiced 1.9 more words in the hints condition than in the show-answer condition, $t(89) = -3.28$, $p = .001$, $d = 0.26$. Bayesian t tests indicated inconclusive evidence regarding the

[3] The degrees of freedom differ from the other recall measures because two participants in Experiment 2, and one participant in Experiment 3, did not complete the last test with prompts from practice within the lesson.

(absence of) differences in the number of trials, $BF_{10} = 0.49$, but strong evidence for the higher number of words practiced in the hints condition, $BF_{10} = 16.3$. The number of errors was also significantly lower in the hints condition ($M = 1.39$) than in the show-answer condition ($M = 2.21$), $t(89) = 6.42$, $p < .001$, $d = 0.60$, with $BF_{10} = 1720000$, indicating very strong evidence. When students made an error, the chance that they translated the word correctly the next time it came up was higher when they had received hints feedback than when they had received show-answer feedback, $t(89) = -6.31$, $p < .001$, $d = 0.67$, $BF_{10} = 1070763$.

Discussion

In Experiment 2, we tested whether retrieval practice with feedback with mnemonic hints is more efficient than retrieval practice with standard show-answer feedback. Differences between the two feedback conditions were found on the recall test with hints as well as during practice. As in Experiment 1, the overall number of words that were recalled on the final test one week after practice was not significantly different in the two feedback conditions. The need for recall prompts was also not significantly different, although there was a trend that students used more prompts to recall the words from the hints condition. Consistent and strong evidence for a difference in recall between the conditions was only found on a separate test on which words were presented with the mnemonic hints from practice. On this test, students showed better recall for the words from the hints condition than for the words from the show-answer condition. Likely, students were better able to find the correct translation with the mnemonic hints if they had seen those hints already during practice. This suggests that students recognized the hints from practice on the test, but their knowledge of the hints did not enhance their recall performance on the test without hints. Benefits of practice with hints did not transfer to a recall situation without hints, as in Experiment 1.

During practice, the total number of trials was similar in the two conditions, but a higher number of words were practiced in the hints condition than in the show-answer condition. This was the case because students made fewer errors in the hints condition. Students were also more likely to answer correctly on the next repetition of a word when they got hints feedback after an error than when they got show-answer feedback. These results are promising because they suggest that students may have benefited from the mnemonic hints feedback and were less likely to repeat errors during practice. However, differences during practice did not lead to differences in recall on the final test.

A limitation of Experiment 2 is that we could not counterbalance the order in which the two conditions were presented. The order was fixed to avoid that students who first practiced with mnemonic hints would then exploit the keyword strategy during practice in the show-answer condition. Differential transfer has also been avoided with such a fixed presentation order in other within-subject studies on mnemonic techniques (e.g., Fritz et al., 2007). To get an estimate of possible order effects in Experiment 2, the data of Experiment 1—in which the practice blocks were counterbalanced—were reanalyzed. Averaged across feedback conditions, learning outcomes and all measures of practice efficiency except for the chance to learn from errors were better for words practiced in Block 2 than in Block 1 in Experiment 1. These control analyses suggest that the order of the two practice blocks in Experiment 2 might have caused higher learning outcomes in the hints condition (always Block 2) compared to the show-answer condition (always Block 1). However, in spite of these possible order benefits for the hints condition, later learning outcomes did not differ significantly between the two feedback conditions in Experiment 2. This further strengthens the conclusion that hints did not enhance later recall.

Experiment 3

In Experiment 3, we presented students with cross-language hints. These hints were similar to Experiment 2 in that they contained a keyword to help the learners associate the vocabulary words to their translation. This time, however, the keyword was a cognate. Cognates are words from different languages that have similar phonological and/or orthographical forms and are often semantically related, like "vest" (English) and "vestis" (Latin).

Cognates tend to be recognized and learned more efficiently than noncognates (e.g., Helms-Park & Dronjic, 2015; Rogers, Webb, & Nakata, 2015). However, learners often fail to recognize cognates (Moss, 1992) and increasing learners' cognate awareness might be beneficial for word learning (e.g., White & Horst, 2012). We therefore used cross-language hints in Experiment 3, which drew students' attention to the cognate status of the to-be-learned words. The target language that students practiced in Experiment 3 was Latin and the hints contained cognates from the students' second language, English. For example, when students were trying to translate the Latin word *vestis* into their first language German (de. *Kleidung* [clothes]), the hint was: "Try again! Think of the English word 'vest'!". We expected that these hints would help students associate the Latin vocabulary to their prior knowledge, and thereby enhance retention.

Method

Participants. A total of 88 students of a German high school took part in the experiment. The data of 74 students (59.5% female, $M_{age} = 13.2$ years, $SD_{age} = 0.6$) were analyzed; the other students had incomplete data sets because they were absent during one of the lessons or experienced technical problems that led to incomplete or repeated study blocks. The students were in Grades 7 and 8 of a grammar school that prepared them for university education. Students had learned English at school for 9 years on average ($SD = 1.17$) and had learned Latin for less than three years ($M = 2.2$, $SD = 0.6$).

Stimuli. Seventy-two Latin words were selected from vocabulary lists from the last chapters of the schoolbook of Grade 8, which the students had not yet studied according to their teachers. This was confirmed by a low proportion of unpracticed experimental words that were correctly translated on the final test ($M = 0.18$, $SD = 0.14$) and subjective ratings that the students made about the proportion of words they already knew ($M = 0.06$, $SD = 0.07$).

Differences Between Experiment 2 and Experiment 3.

Design and experimental control. In Experiment 3, the order of the two practice blocks (hints feedback or show-answer condition) was counterbalanced across participants.

Cross-language hints during retrieval practice. Students in the hints feedback condition were presented with a cross-language

hint when they made an error. The target language that students practiced in Experiment 3 was Latin and the hints contained cognates from the students' second language, English. For example, when students were trying to translate the Latin word *procedere* [to proceed] into German (fortfahren), the hint was, "Versuch es noch einmal! Denk an das englische Wort 'to proceed' (vorgehen, fortfahren)!" ["Try again! Think of the English word 'to proceed' (to continue, to proceed)!"]. To ensure that students understood the English cognate, the hints contained either German translations of the cognate or a brief phrase from which its meaning could be derived. For example, for the word *honestus* (*ehrlich*, or honest], the hint was,"Think of the English word 'honest' (e.g., 'an honest answer'!" The hints were designed in such a way that students could not just type over the only German translation in the hint but had to think about the meaning of the cognate. The cognates were chosen from a database of etymological relations between Latin and English (Gerbrandy et al., 2014).

Test of learning outcomes. For practical reasons, the delayed test took place three days after learning instead of a week later, as in Experiment 1 and 2. During Session 2, students first took the same recall test as in Experiment 1 and 2, and then a test that presented the vocabulary words with the cross-language hints from practice.

Results

Feedback effects on later recall. The overall number of words that students recalled on the final test, was not significantly different in the show-answer condition and in the hints condition, $t(73) = 2.60$, $p = .01$, $d = 0.24$ ($\alpha = .008$), but was numerically higher in the show-answer condition. The proportion of words for which students needed recall prompts on the test was not significantly different in the two conditions, $t(73) = -1.37$, $p = .18$, $d = 0.19$, $BF_{10} = 0.31$. On the separate test with the cross-language hints from practice, students did not perform significantly different for words from the show-answer condition and the hints-condition, $t(72) = 2.46$, $p = .016$, $d = 0.22$ (see Footnote 3), but again showed numerically higher results in the show-answer condition. Bayesian *t* tests indicated anecdotal evidence for differences between the conditions on overall recall and recall on the test with mnemonic hints (BF_{10}, respectively, 2.9 and 2.14).

Feedback effects during the practice phase. On average, students went through significantly more trials in the show-answer condition than in the hints condition, $t(73) = 5.73$, $p < .001$, $d = 0.60$, $BF_{10} = 65{,}493$ (very strong evidence). Students numerically practiced more words in the show-answer condition than in the cross-language hints condition, $t(73) = 2.52$, $p = .01$, $d = 0.25$, $BF_{10} = 2.4$, but this difference was not significant at $\alpha = .008$ and Bayesian analyses indicated only anecdotal evidence. The number of errors that students made per word, was relatively low ($M_{\text{hints}} = 1.36$ and $M_{\text{ShowAns}} = 1.41$), and not significantly different between conditions, $t(73) = -0.35$, $p = .73$, $d = 0.04$, $BF_{10} = 0.14$ (moderate evidence for H_0). The chance that students gave a correct response on the next repetition of a word after a previous error, was higher in the show-answer condition than in the cross-language hints condition, $t(73) = 2.75$, $p = .0074$, $d = 0.30$, $BF_{10} = 4.19$.

Discussion

Experiment 3 showed no significant difference in learning outcomes of retrieval practice with cross-language hints feedback and show-answer feedback. On the recall tests three days after practice, overall recall and recall with the cross-language hints from practice were not significantly different. Performance was even numerically higher in the show-answer condition than in the hints condition, but feedback effects were not significant after correction for multiple comparisons. Test-taking behavior did not differ significantly between conditions; students needed recall prompts on the final test about equally often for words practiced with cross-language hints and for words practiced with show-answer feedback.

During practice, students completed significantly more trials in the show-answer condition than in the cross-language hints condition, as in Experiment 1. Surprisingly, students were also more likely to respond correctly on the next repetition of a word if they had received show-answer feedback after an error than if they had received cross-language hints. A possible explanation for this could be that students changed their response behavior during practice when they received hints feedback and more readily submitted incorrect responses because they knew that they would get a second chance with a hint. There was, however, no significant difference in the proportion of correct answers overall during the practice blocks with and without hints-feedback.

General Discussion

In spite of a large literature on feedback effects in general (Narciss & Huth, 2004; Shute, 2008), relatively little is known about specific elaborations that make feedback more efficient (Van der Kleij et al., 2015). The present study provides information about one possible elaboration, namely hints that are designed to evoke memory retrieval. We conducted three experiments to compare how well high school students learned from a computerized vocabulary training that included either standard show-answer feedback or hints feedback in case of an error. Overall, learning outcomes were similar in the two feedback conditions in the three experiments except when the hints from practice were available again during the recall test. During practice, hints feedback led to a shift of time from further repetitions to longer feedback processing after errors in all three experiments, but only the mnemonic hints feedback in Experiment 2 also reduced errors during practice.

Both the extensive literature on the benefits of retrieval practice (e.g., Karpicke, 2017; Rowland, 2014) and the long-standing tradition in education to work with scaffolds (Wood et al., 1976), suggest that hints feedback is beneficial for later recall. However, neither orthographic (Experiment 1), mnemonic (Experiment 2), nor cross-language hints (Experiment 3) led to higher overall recall compared with standard show-answer feedback in the present study. Experiment 3 even showed numerically higher recall in the show-answer condition than in the cross-language hints condition. The only significant differences between conditions were found when the hints from practice were available again during the recall test. Students used significantly more orthographic recall prompts on the final test for the words from the orthographic hints condition than for the words from the show-answer condition (Experiment 1) and recall on a separate test with mnemonic prompts was higher for words from the mnemonic hints condition than for

words from the show-answer condition (Experiment 2). These could be examples of transfer-appropriate processing (Morris, Bransford, & Franks, 1977), the phenomenon that memory performance depends on the match between practice and test. Specifically, recall might have become dependent on the hints if an association between the hints and the correct response was formed but this association could not be retrieved when the hints were later absent (for more information on context-dependent memory, see Smith & Handy, 2016). Tentatively, this could mean that effects of practice with hints do not transfer to later recall without hints.

The lack of general benefits of hints feedback replicates the results of three prior studies which found no difference in recall after practice with (orthographic) hints and show-answer feedback (Hall et al., 1968; Kornell et al., 2015; Kornell & Vaughn, 2016). For the earlier studies, this lack of effects could have been due to the use of hints that were completed too easily (Kornell et al., 2015; Kornell & Vaughn, 2016) or that triggered only processing of orthographic features (Hall et al., 1968). However, even with orthographic hints that required effortful processing of a new word form and with carefully constructed mnemonic and cross-language hints, the present study did not show benefits of hints feedback compared to show-answer feedback. Clearly, adding hints feedback to repeated retrieval practice does not in every case enhance later recall. To understand this result, it is necessary to take into account the effect of hints feedback during practice under time constraints.

In the present study, we controlled the total practice time to ensure that the feedback manipulation was not confounded by longer practice times in the hints condition than in the show-answer condition (as, e.g., in Hall et al., 1968). As a consequence, hints feedback could influence mainly two aspects of practice: the duration of feedback processing after errors (which was by definition longer in the hints condition than in the show-answer condition) and the chance that students learned from feedback and did not repeat errors. These effects, in turn, influenced the number of trials that students could go through in the available practice time and thereby the number of words in practice (see also Hays et al., 2010). In Experiments 1 and 3, the hints feedback did not reduce (repeated) errors, and therefore resulted in a lower number of practiced words. In Experiment 2, students made significantly fewer (repeated) errors during practice with mnemonic hints feedback, and this led to a higher number of words practiced in the hints condition than in the show-answer condition. These differences between the number of words practiced in the two feedback conditions could have influenced later recall and explain the lack of benefits of hints feedback. Therefore, we replicated all analyses of learning outcomes using the proportion of practiced words that were recalled on the test (e.g., 3 recalled out 10 practiced words = 0.30) as dependent variable (see the Appendix). These control analyses led to the same conclusions: show-answer and hints feedback conditions only differed significantly on test trials on which the hints from practice were again available. The (lack of) feedback effects in the present study is thus unlikely to be a mere consequence of differences in the number of practiced words.

A number of characteristics of this study need to be taken into account when generalizing conclusions to other learning situations. First, the learning system that controlled the spacing of repetitions during practice (Sense et al., 2016) was changed to produce a relatively high error rate during practice, compared to its default settings. The error rates of 30% to 40% in the present study were, however, still lower than the error rate in the single previous study that found positive effects of hints feedback (Finn & Metcalfe, 2010, who reported an initial error rate of 70% to 78% before feedback). We used these adjusted settings to evoke enough errors during practice to observe differences between the conditions, because the feedback conditions differed only on error trials. Feedback effects may be different if fewer errors occur during practice so that the few feedback moments draw more attention. This could, for example, motivate students to process mnemonic or cross-language hints more thoroughly. Similarly, the effect of hints feedback might be different if hints are only given after certain incorrect responses, for example, only for words that were repeatedly translated incorrectly. These could be starting points for further research, although based on the present results it is questionable how large the benefits of hints feedback would be. Second, we used an effective baseline practice condition—adaptive, repeated, spaced retrieval (Delaney, Verkoeijen, & Spirgel, 2010; Pavlik & Anderson, 2008). The trade-off between longer feedback processing and further practice trials may be different when the baseline is not as effective. However, a less effective baseline is also less relevant for practical purposes. Third, the effect of hints feedback might depend on the retrieval task and materials. Benefits of elaborated feedback are stronger for complex than simple learning tasks (Van der Kleij et al., 2015), and this might also be the case for hints feedback. In the present experiments, students responded to newly learned vocabulary words by typing in words from their native language. Hints might, for example, have more benefits when the translation direction is reversed and learners must retrieve a newly learned foreign word. In this case, hints could possibly trigger more useful, deep processing of the word form.

Conclusion

We found no clear benefits of hints feedback that created an extra retrieval opportunity compared to show-answer feedback. This was an unexpected finding given the large support for benefits of retrieval practice in general (Adesope et al., 2017; Rowland, 2014). A manipulation that otherwise enhances learning outcomes (i.e., retrieving a word from memory instead of restudying the complete word) is not automatically a beneficial addition to feedback after a failed retrieval attempt. More is not always more: even a simple addition to a training, such as replacing show-answer feedback with hints feedback, incurs costs because it takes time away from other forms of practice, such as further repetitions. Moreover, in the present study, students did not transfer what they practiced with hints to a later recall situation without hints. It is important to consider carefully whether hints that support retrieval during practice lead to associations that can also be used in situations when the hints are not available.

References

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

Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American Psychologist, 30,* 821–828. http://dx.doi.org/10.1037/h0077029

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education, 110,* 154–169. http://dx.doi.org/10.1016/j.compedu.2017.03.012

Beaton, A. A., Gruneberg, M. M., Hyde, C., Shufflebottom, A., & Sykes, R. N. (2005). Facilitation of receptive and productive foreign vocabulary learning using the keyword method: The role of image quality. *Memory, 13,* 458–471. http://dx.doi.org/10.1080/09658210444000395

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,* 268–276. http://dx.doi.org/10.3758/BF03193405

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

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294. http://dx.doi.org/10.1037/0096-3445.104.3.268

Deconinck, J., Boers, F., & Eyckmans, J. (2017). 'Does the form of this word fit its meaning?' The effect of learner-generated mapping elaborations on L2 word recall. *Language Teaching Research, 21,* 31–53. http://dx.doi.org/10.1177/1362168815614048

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53,* 63–147. http://dx.doi.org/10.1016/S0079-7421(10)53003-2

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

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,* 289–298. http://dx.doi.org/10.1016/j.jml.2011.01.006

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition, 38,* 951–961. http://dx.doi.org/10.3758/MC.38.7.951

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology, 21,* 499–526. http://dx.doi.org/10.1002/acp.1287

Gerbrandy, P., Castricum, J., Hermsen, C., Hupperts, C., Raiijmakers, M., Risselada, R., . . . Wolsing, I. (2014). *Janus. Connecties tussen het Latijn en moderne Europese talen* [Janus. Connections between Latin and modern European languages]. Retrieved from http://janus.humanities.uva.nl/

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. http://dx.doi.org/10.1037/0033-295X.91.1.1

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 801–812. http://dx.doi.org/10.1037/a0023219

Hall, K. A., Adams, M., & Tardibuono, J. (1968). Gradient- and full-response feedback in computer assisted instruction. *The Journal of Educational Research, 61,* 195–199. http://dx.doi.org/10.1080/00220671.1968.10883643

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review, 17,* 797–801. http://dx.doi.org/10.3758/PBR.17.6.797

Helms-Park, R., & Dronjic, V. (2015). Crosslinguistic lexical influence: Cognate facilitation. In R. Alonso (Ed.), *Crosslinguistic influence in second language acquisition* (Vol. 95, pp. 71–92). Bristol, UK: Multilingual Matters.

JASP Team. (2016). JASP (Version 0.8.0.0) [Computer software]. Retrieved from https://jasp-stats.org/

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.) & J. H. Byrne (Series Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (pp. 487–514). Oxford, UK: Academic Press. http://dx.doi.org/10.1016/B978-0-12-809324-5.21055-9

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254

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,* 85–97. http://dx.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,* 283–294. http://dx.doi.org/10.1037/a0037850

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

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science, 25,* 639–647. http://dx.doi.org/10.1177/0956797613504302

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

Moss, G. (1992). Cognate recognition: Its importance in the teaching of ESP reading courses to Spanish speakers. *English for Specific Purposes, 11,* 141–158. http://dx.doi.org/10.1016/S0889-4906(05)80005-5

Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language, Learning and Technology, 11,* 107–129. Retrieved from http://llt.msu.edu/vol11num3/murphy/

Murphy, P. (2010). Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer-mediated communication. *ReCALL, 22,* 112–134. http://dx.doi.org/10.1017/S0958344010000030

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegeman, D. Leutner, & R. Brünken (Eds.), *Instructional design for multimedia learning* (pp. 181–195). Berlin, Germany: Waxmann.

OECD. (2016). *Netherlands 2016: Foundations for the Future, Reviews of National Policies for Education*. Paris, France: OECD Publishing. http://dx.doi.org/10.1787/9789264257658-en

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

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,* 437–447. http://dx.doi.org/10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. http://dx.doi.org/10.1037/a0023956

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

Rogers, J., Webb, S., & Nakata, T. (2015). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research, 19,* 9–27. http://dx.doi.org/10.1177/1362168814541752

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432–1463. http://dx.doi.org/10.1037/a0037559

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science, 8,* 305–321. http://dx.doi.org/10.1111/tops.12183

Shapiro, A. M., & Waters, D. L. (2005). An investigation of the cognitive processes underlying the keyword method of foreign vocabulary learning. *Language Teaching Research, 9,* 129–146. http://dx.doi.org/10.1191/1362168805lr151oa

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189. http://dx.doi.org/10.3102/0034654307313795

Smilek, D., Carriere, J. S. A., & Cheyne, J. A. (2010). Failures of sustained attention in life, lab, and brain: Ecological validity of the SART. *Neuropsychologia, 48,* 2564–2570. http://dx.doi.org/10.1016/j.neuropsychologia.2010.05.002

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,* 1582–1593. http://dx.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,* 1134–1141. http://dx.doi.org/10.1080/09658211.2015.1071852

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85,* 475–511. http://dx.doi.org/10.3102/0034654314564881

van Rijn, D., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving Learning Gains by Balancing Spacing and Testing Effects. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 108–114). [187] Manchester.

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

White, J. L., & Horst, M. (2012). Cognate awareness -raising in late childhood: Teachable and useful. *Language Awareness, 21,* 181–196. http://dx.doi.org/10.1080/09658416.2011.639885

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Child Psychology & Psychiatry & Allied Disciplines, 17,* 89–100. http://dx.doi.org/10.1111/j.1469-7610.1976.tb00381.x

Appendix

Feedback Effects on the Proportion of Recalled Words

We measured learning outcomes as the number of words that were recalled on the final test. As a control analysis, we here replicated all analyses using a different dependent variable: the proportion of the practiced words that were recalled on the final test. For instance, a student who recalled 12 words out of 24 practiced words had a proportion correct on the test of 12/24 = 0.5. These extra analyses were done to compare the test results between feedback conditions after controlling for differences in the number of practiced words.

To foreshadow: The results reported hereafter all lead to the same conclusions as the analyses reported in the main text. The (lack of) feedback effects is thus not likely due to differences in the number of practiced words.

Experiment 1

Students practiced on average 2.8 fewer words in the hints condition than in the show-answer condition. Relative to this smaller number, overall, a numerically larger proportion of words were recalled correctly from the hints-condition than from the show-answer condition, $t(84) = 2.70$, $p = .008$, $d = 0.23$. This difference did not reach the corrected alpha of 0.008, but a Bayes factor BF_{10} of 3.5 ($BF_{01} = 0.28$) indicated moderate evidence against the null hypothesis. The overall recall on the test was calculated as the proportion of practiced words that were recalled correctly on the test either directly or on a second attempt with orthographic prompts. Further analyses of test sub scores revealed that students recalled a similar proportion of the words from both conditions on the first attempt, $t(84) = 1.38$, $p = 0.17$, $BF_{10} = 0.30$ (moderate evidence for H_0), but a larger proportion of the words from the hints condition with orthographic prompts, $t(80) = -2.90$, $p = .005$, $BF_{10} = 5437$ (strong evidence against H_0).

Comparison With Main Results

As in the results reported in the main text, a significant difference between the two feedback conditions was only found when the hints from practice were available also on the test.

(*Appendix continues*)

Experiment 2

Students practiced on average 1.9 more words in the hints condition than in the show-answer condition. The proportion of words that were recalled overall was not significantly different between the two conditions, $t(89) = 0.88$, $p = .38$. $BF_{10} = 0.17$ ($BF_{01} = 5.89$) indicated moderate evidence for the null hypothesis. Further analyses revealed that students recalled a numerically higher proportion of the words from the show-answer condition on the first attempt, $t(89) = 2.17$, $p = .03$, $BF_{10} = 1.08$, and a numerically larger proportion of the words from the hints condition on the second attempt with orthographic prompts, $t(84) = -2.30$, $p = .24$, $BF_{10} = 1.44$. However, Bayesian *t*-tests indicated that the evidence was inconclusive for both of these contrasts. On the test with the semantic prompts from practice, students recalled a significantly higher proportion of the words from the hints condition than from the show answer condition, $t(87) = -4.99$, $p < .0001$, $BF_{10} = 4949$.

Comparison With Main Results

The only significant difference between the feedback conditions in the proportion of words recalled was found on the test on which the semantic hints from practice were available again. On this test, students performed better after practice with hints feedback. The same is reported in the main text based on analyses of the number of words recalled.

Experiment 3

None of the outcome measures showed a significant difference between the two feedback conditions in Experiment 3: The proportion of words that were recalled overall, $t(73) = 1.70$, $p = .09$, $BF_{10} = 0.5$ (inconclusive evidence), on the first attempt, $t(73) = 2.12$, $p = 0.04$ (numerically higher in the show-answer condition but $BF = 1.05$ indicated inconclusive evidence), and with orthographic prompts, $t(73) = -0.566$, $p = 0.57$, $BF_{10} = 0.15$ (moderate evidence for H_0). On the test with the ethymological prompts from practice, students also recalled a similar proportion of the words from both conditions, $t(72) = -0.02$, $p = 0.98$, $BF_{10} = 0.13$ (moderate evidence).

Comparison With Main Results

As in the analyses of the number of recalled words reported in the main text, none of the outcome measures differed significantly between the two feedback conditions.

Received July 22, 2018
Revision received November 21, 2018
Accepted November 30, 2018 ■