Effects of making errors in learning a foreign language

Guzmán-Muñoz, Francisco Javier

Published in:
Journal of cognitive psychology

DOI:
10.1080/20445911.2020.1711766

Publication date:
2020

Journal of Cognitive Psychology

Routledge
Taylor & Francis Group

Journal of Cognitive Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pecp21

Effects of making errors in learning a foreign language

Francisco Javier Guzmán-Muñoz

To cite this article: Francisco Javier Guzmán-Muñoz (2020) Effects of making errors in learning a foreign language, Journal of Cognitive Psychology, 32:2, 229-241, DOI: 10.1080/20445911.2020.1711766

To link to this article: https://doi.org/10.1080/20445911.2020.1711766

Published online: 12 Jan 2020.

JOURNAL OF COGNITIVE PSYCHOLOGY
2020, VOL. 32, NO. 2, 229–241
https://doi.org/10.1080/20445911.2020.1711766

Routledge
Taylor & Francis Group

OPEN ACCESS

Check for updates

Effects of making errors in learning a foreign language

Francisco Javier Guzmán-Muñoz[a,b]

[a]Institute of Engineering and International Business School, Hanze University of Applied Sciences, Groningen, Netherlands; [b]Department of Psychology, University of Groningen, Groningen, Netherlands

ABSTRACT

Kornell, Hays, and Bjork ([2009]. Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998) showed that incorrect guesses do not necessarily harm and might even improve the retention of information on a subsequent test. We sought to replicate the finding using educationally relevant stimuli. In two experiments, our participants either translated sentences in a foreign language receiving immediate feedback (errorful condition), or copied and studied the correct translation (errorless condition). After this training phase, a final test with the same sentences showed that translating sentences wrongly during training did not lower the accuracy of the errorful as compared to the errorless condition. Overall there was evidence that errorful training produced superior learning of the meaning and grammar of the foreign language sentences. The results support the idea that search processes activate a greater network of related knowledge in the errorful than in the errorless condition.

ARTICLE HISTORY
Received 18 October 2018
Accepted 20 December 2019

KEYWORDS
Second language acquisition; testing effect; generation effect; errors in learning; errorless training

When trying to learn new information, testing yourself regularly instead of restudying the materials has been shown to improve performance. This effect of retrieval practice, or *testing effect*, is among the most robust findings in psychological and educational research and is supported by around 100 years of studies (for recent reviews see Delaney, Verkoeijen, & Spirgel, 2010; Roediger, Agarwal, Kang, & Marsh, 2010; Roediger & Karpicke, 2006a). Although the effect has been known for long, recent research has focused on its application to educationally relevant materials and on the consequences of giving a wrong answer. The present paper is concerned with both topics.

The testing effect with complex materials

The vast majority of research in the testing effect has traditionally used simple experimental tasks like learning paired-word associates or memorising lists of words. Interestingly, two of the earliest studies recorded (Gates, 1917; Spitzer, 1939) used ecological settings and educational materials. Gates compared reading with recitation (i.e. put away the material and try to recall from memory) and manipulated the time devoted to each method between groups. His results showed a positive relation between time spent in recitation and performance in a test of biographical facts. Spitzer had more than 3600 students learn 600-word essays and tested them according to different schedules during the following days, and once more after 63 days. His most interesting findings were that the later a first test was given the more forgetting occurred (a normal forgetting function) and that taking a test almost stopped forgetting. Whenever students took a test, performance in the final 63-day delayed test remained roughly at the same level as in that first test.

In the early 1980s, Duchastel and colleagues (Duchastel, 1981; Duchastel & Nungester, 1981; Nungester & Duchastel, 1982) embarked in a series of studies researching testing effects using historical passages as stimuli. The stimuli-text consisted of 1700-word essays describing 12 different topics of British history corresponding to the reign of Queen Victoria (i.e. the Crimean war, the Suez Canal, Living-

CONTACT Francisco Javier Guzmán-Muñoz f.j.guzman.munoz@pl.hanze.nl Institute of Engineering (location Assen), Hanze University of Applied Sciences, Industrieweg 1, Assen, Groningen 9402 NP, Netherlands

stone). In their different studies, they compared study-only (review in their terminology) conditions with test and control conditions as well as different test formats. For example, Nungester and Duchastel had all students read the passage for 15 minutes. Then, the participants in the test condition completed a five-minutes test (composed of multiple-choice and short-answer items), those in the review condition spent five minutes reviewing the material and those in the control group completed a filler task (a learning processes questionnaire). Performance in a two-week delayed test was superior in the test group. However, their results might have been influenced by a practice effect as the participants in the test group were tested with half of the items twice (albeit in a different format). This explanation is supported by the fact that the advantage of testing was confined to those previously used items. Glover (Experiment 1a, 1989) used a 300-word essay about a fictitious nation to compare a study-test group with a study only group. Four days later, the study-test group performed better in a free-recall test. The results of Glover might also be contaminated by the extra practice afforded to the study-test group which had the testing session two days after the initial learning session whereas the control group didn't have an equivalent study opportunity. McDaniel and Fisher (1991) also showed a testing effect, in this case using trivial facts extracted from a popular game. In their study, being tested (with feedback) about a fact resulted in better memory in a cued-recall test than equivalent study time. Interestingly, elaborative processing of the feedback during the test didn't increase, as compared to rote rehearsal, the efficacy of the test condition. The authors explained the finding as a consequence of the nature of the stimuli used. Trivial facts can be understood as arbitrary associations and they don't get enhanced by elaboration as more complex and integrated material probably could. This explanation, however, is not consistent with all available evidence. Pressley and colleagues for example, have shown that elaboration (promoted through questions) does have a positive effect in the learning of arbitrary facts (Pressley, Symons, McDaniel, Snyder, & Turnure, 1988; Woloshyn, Willoughby, Wood, & Pressley, 1990).

More recently, Roediger and colleagues have carried out a number of studies researching the testing effect using integrated and complex materials (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Chan, McDermott, & Roediger, 2006; Kang, McDermott, & Roediger, 2007; Roediger & Karpicke, 2006b). For example, in a very clear demonstration of the testing effect, Roediger and Karpicke (Experiment 1) had their participants study two different general science passages of approximately 260 words. After the first study session participants had a second study session about one of the passages (study-study) and a test about the other (study-test) and followed a final test either after five minutes, two days or one week. Participants who did the immediate test recalled more content of the study-study text, whereas participants who did the delayed tests recalled more from the study-test condition. Among other complex materials, the testing effect has also been shown in learning of maps (Carpenter & Pashler, 2007), teaching of clinical topics to medical students (Larsen, Butler, Lawson, & Roediger, 2012; Larsen, Butler, & Roediger, 2009), learning of face-name pairs (Tse, Balota, & Roediger, 2010; Weinstein, McDermott, & Szpunar, 2011), learning the meaning of Chinese characters (Kang, 2010) and learning of mathematical functions (Kang, McDaniel, & Pashler, 2011).

Logically, in the last years, the interest of researchers has moved into discovering the boundaries of the testing effect. For example, in order to investigate whether the format of the test will affect the size of the effect, Karpicke and Blunt (2011) compared students studying scientific texts and taking afterwards either extra study time, a free-recall test or elaborating a concept-map. When tested one week later, the participants in the test condition showed the highest levels of recall. Studies comparing more classical test formats as short-answer (SA) and multiple-choice (MC) have shown for example that taking a MC test can improve retention of related, non-tested information in comparison with taking a SA test (Little, Bjork, Bjork, & Angello, 2012). Taking a MC test however might also increase the number of incorrect intrusions (lures) in the final test (Fazio, Agarwal, Marsh, & Roediger, 2010; Roediger & Marsh, 2005). Similarly, Potts and Shanks (2014) also found that lures selected during training were more likely to be selected again during testing. This is obviously one of the unintended consequences of taking a test and it will be reviewed in more detail in the next section.

Consequences of giving a wrong answer

Every time participants take a test we risk that they give a wrong answer. From a theoretical point of view, classical learning theory recommends the avoidance of errors because they will create wrong stimulus-response associations (Skinner, 1958). This idea has been applied in a technique called *errorless learning* (Terrace, 1963) in which participants are prevented from making errors by giving them the correct response in advance. Errorless learning has been successfully applied in learning tasks with amnesics (Baddeley & Wilson, 1994; Tulving, Hayman, & Macdonald, 1991), older adults (Kessels & de Haan, 2003) and Alzheimer patients (Clare et al., 2000). On the other hand, there is evidence that the commission of errors can also have positive consequences on learning. For example, Frese and colleagues have developed the *error-management* approach in which learners are permitted to make errors during training and encouraged to learn from them (Dormann & Frese, 1994; Heimbeck, Frese, Sonnentag, & Keith, 2003). Errors can be promoted, for example, through the use of problems that exceed the level of expertise of the learners or by giving minimal instructions and encouraging learners to explore the system. As an illustration of the approach, Dorman and Frese found that errors promoted exploration in a task of learning to use the SPSS data analysis environment and exploration activity correlated positively with performance in a final test (for a review of the topic see Guzmán-Muñoz, 2009).

A key feature of studies in errorless learning and error management is that they simply compare groups of participants making errors with groups of participants avoiding them. The studies show that, overall, the benefits of making errors outweigh the drawbacks but they don't research the consequences of specific error responses. As described in the previous section, one negative consequence of making an error in a MC test is that the wrong response might re-appear in the final test. For example, Smith and Karpicke (2014) reported an increase in the number of incorrect responses from the training phase that re-appeared in the final test. The same finding has been reported by Roediger and Marsh (2005) using more complex educational materials.

In contrast, Kornell, Hays, and Bjork (2009) showed recently an advantage in a memory test of incorrect responses as compared to equivalent study time. In their first two experiments, participants studied obscure fictional and non-fictional matched facts and the main manipulation consisted of a comparison between study-study or guess-study conditions. In the guess-study condition, the participant is presented with the question and, after the time given to respond, the answer is presented. In the study-study condition, the participant studies the pairs of questions and answers for an equivalent amount of time. The fact of using obscure facts as stimuli guaranteed that the majority of responses in the guess-study condition were erroneous. However, the performance was not affected by wrong responses and answers in a cued-recall test were similar for guess-study and study-study items. Kornell et al. argued that trying to retrieve an answer activates a net of related knowledge which probably includes the right answer even when it was not given as a response on the first occasion. When feedback is presented, this activation of related knowledge makes encoding of the correct stimulus-response association more successful. This explanation was supported in the following experiments. When the authors changed the stimuli to weakly-related word pairs, cued-recall of guess-study items was superior to study-study items, clearly showing an advantage for retrieval failures as compared to study only conditions.

This idea was termed *search set theory* and further investigated by Grimaldi and Karpicke (2012). In a series of three experiments, they tested three corresponding hypotheses derived from search set theory. They argued that, if semantically related candidates become active during the retrieval attempt, then the advantage of failed retrievals should decrease by: using unrelated words as word pairs (Exp. 1), constraining the search set to a particular (wrong) candidate (Exp. 2) or delaying the presentation of the feedback, in which case, the activation of the set of candidates would diminish. All predictions were supported. Similarly, Hays, Kornell, and Bjork (2012) and Vaughn and Rawson (2012) have showed that failed retrieval attempts promote memory if feedback is immediate but that study only is more effective when feedback is delayed.

The previously reviewed studies show that giving a wrong answer increases cued-recall performance as compared to study-only conditions in tasks that involve simple stimuli. Kang, Pashler, et al. (2011) extended these findings to a task that involved learning richer conceptual information. In their third experiment, they asked participants to construct plausible explanations for a number of common phenomena. The authors used familiar phenomena (scientific or cultural) whose causal explanation is normally unknown to the average college student in order to guarantee a high rate of errors in the initial guesses. Their results revealed that making wrong guesses didn't harm performance in this task as long as feedback was given relatively soon after the response.

The present study

The studies reviewed above show that there is a growing literature that supports the positive effects of testing/retrieval practice even when this retrieval is not successful (Grimaldi & Karpicke, 2012; Hays et al., 2012; Kornell et al., 2009). However, most of these studies have used simple stimuli, often limited to learning paired-word associates (with the exception of Kang, Pashler, et al., 2011). The interest in the study of the testing effect with complex materials is illustrated by a relatively recent special issue in Educational Psychology Review. In that issue, we can find authors who argue for the existence of such effect (Karpicke & Aue, 2015) and authors who deem its size to be negligible (Van Gog & Sweller, 2015).

The present study was designed to test the effect of making wrong guesses using arguably more complex and educationally relevant materials. With that purpose, we asked our participants to engage in a task of foreign language learning. The task involved a training phase in which participants translated short sentences after being assigned to either an errorless or an errorful condition. In the errorful condition, each sentence was presented and the participant had to write the translation before feedback together with the correct response was given. In the errorless condition, sentences were presented concurrently with the translation and the participant simply had to copy the answer. We believe that the present paradigm offers an advantage over traditional study-study vs. test-study comparisons. In a study-study condition, we cannot know whether the students are indeed studying the materials. They might instead be looking at the computer but thinking of something else. In our errorless condition, students were required to copy down the correct response and their accuracy during the task could be recorded to check that they indeed did so accurately. In addition, the use of complete sentences allows testing the knowledge of more complex information (i.e. grammatical rules) than simple associations between pairs of stimuli. After the training phase, the participants were tested through a recall test (Exp. 1) and a recognition test (Exp. 2). The expectation, based on the results of Kang et al. was that making errors should not harm performance in a later test. More specifically and based on the literature reviewed previously, we hypothesised that the improvement in test performance would extend even to those items in which errors were made during training.

Experiment 1

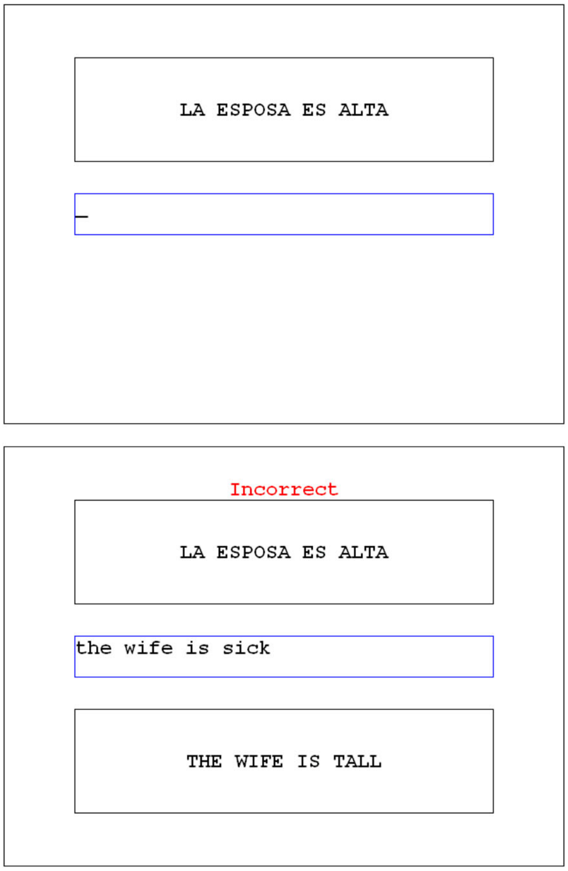
Method

Participants: Fifty-one undergraduate students enrolled at a Dutch public university took part in the study in exchange for course credit. The language of instruction at the university is English. Their average age was 20.5 years old (SD = 2.7) and 34 of them classified themselves as female. Fifteen participants were excluded from the analyses because they reported previous knowledge of Spanish so the final distribution between conditions consisted of 18 participants in the errorful and 18 participants in the errorless condition.

Materials: Forty-eight training sentences in Spanish were constructed from the combination of 12 nouns and 12 adjectives and used during the training and testing phases (see Appendix 1). The sentences followed the structure: *article – noun – verb – adjective*, and were built so every word (nouns and adjectives) was used four times in all combinations of gender and number (masculine or feminine and singular or plural). The verb used was "to be" which in Spanish has two forms, depending on how permanent is the characteristic being described. Half of the adjectives thus referred to permanent qualities, those paired with "ser", and the other half referred to changing qualities, in the case of those paired with "estar". All the nouns referred to entities with a specific sexual gender and had four different forms in English. Presentations of gender and number were counterbalanced between the two verbs.

Procedure: The study took place in a multi-station laboratory were up to eight participants could be tested simultaneously. Stimuli were presented and responses were registered with a programme written in E-prime (Schneider, Eschman, & Zuccolotto, 2002).

Participants sat at one of the computers and read and signed a consent form after which the researcher started the task on the computer. All instructions were given through the screen although they were informed that they could ask the researcher at any time. The training phase consisted of the presentation of the 48 training sentences in a different random order per participant. Each sentence was presented inside a black rectangle centred on the upper part of the screen (see Figure 1). Participants were instructed to write the English translation of the sentence in the blue rectangle situated underneath and to press the enter key in order to receive feedback. In the case of the errorless condition, they were provided with the correct translation inside a third rectangle on the lower part of the screen and asked to use that information to avoid errors in their own translation. All participants had a time limit of 15 s to write their translation and, if exhausted, the response was considered an error and feedback was presented. The feedback lasted for 5 s and presented the word "correct" in green or "incorrect" in red on the top of the screen plus a rectangle (already present in the errorless condition) at the bottom of the screen containing the correct translation (see Figure 1).



LA ESPOSA ES ALTA

Incorrect

LA ESPOSA ES ALTA

the wife is sick

THE WIFE IS TALL

Figure 1. The screenshots show how the interface looked in the training phase of the errorful condition, during Experiments 1 and 2. The top panel shows the beginning of a given trial. The bottom panel shows how feedback was given after a (wrong) response had been introduced. (To view this figure in color, please see the online version of this journal).

The testing phase started directly after the training and presented a very similar interface. As in training, a sentence in Spanish appeared on the screen and the participant was required to type the correct translation on the blue box beneath the sentence and press the enter key. There was no time limit during the testing phase and no feedback was given after the response.

Results

We divided the training phase into four blocks of 12 trials each and computed the accuracy per participant across training condition (errorless or errorful) and block of training. Although the errorless participants were provided with the correct translation, they made mistakes probably due to the time pressure and to not using their first language. Even minor spelling mistakes were treated as errors (and thus flashed as "incorrect" in the feedback) by the string-matching algorithm used by the computer. However, in the computation of the accuracy for the analyses, we inspected the error answers during training and testing and recoded them as accurate or inaccurate depending on the error made (see Appendix 2 for a description of the

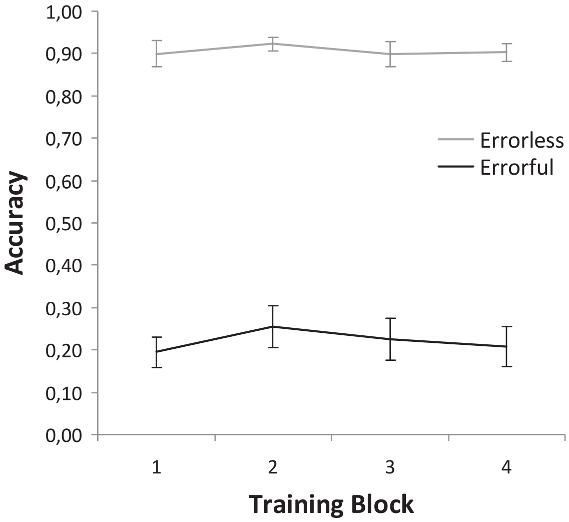


Figure 2. Average accuracy (proportion correct) during training in Experiment 1 as a function of training condition (errorless or errorful) and block of training (1–4). Error bars represent the standard error of the mean.

recoding procedure as well as examples of errors). Figure 2 shows that, on average participants in the errorless condition had a very high accuracy from the beginning of the training whereas the accuracy of the errorful group remained relatively low across blocks. Participants in the errorless condition were also faster completing their training than those in the errorful condition (means 392 (SD = 96.06) and 478 (SD = 67.40) seconds respectively, $t(34) = 3.09$, $p = 0.004$, $d = 1.03$).

Accuracy during the test was computed as a proportion of correct answers. The data met the assumptions for statistical analyses so we compared the conditions through a *t*-test which showed that although the errorful group was more accurate on average during the test (errorful: $M = 0.46$, SD = 0.215; errorless: $M = 0.35$, SD = 0.228), this difference did not reach statistical significance ($t(34) = 1.54$, $p = 0.131$).

We also compared performance during the test as a function of performance during training in order to check whether errors during training resulted in deteriorated performance during the test. That is, we put all sentences from training into groups of correct and incorrect translations and then determined the proportion from each group that was correctly translated during the test phase. We did this per participant, so we could compute the percentage of sentences correctly translated during the test as a function of their score during training and across conditions (see Figure 3). The two conditions differed in their test accuracy with sentences wrongly (errorful: $M = 0.40$, SD = 0.215; errorless: $M = 0.22$, SD = 0.272; $t(33) = 2.10$, $p = 0.043$, $d = 0.73$) and correctly (errorful: $M = 0.79$, SD = 0.195; errorless: $M = 0.36$, SD = 0.232; $t(34) = 5.99$, $p < 0.001$, $d = 2.00$) answered during training. These analyses, however, might suffer from item-selection problems. For example, participants in the errorless condition had an average accuracy of 90% during training. That means that their average accuracy on sentences wrongly answered during training is drawn from a pool of 4–5 sentences approximately. In contrast, the comparable average accuracy of participants in the errorful condition is drawn from a pool of around 38–39 sentences. A more balanced test would involve comparing the accuracy on sentences wrongly answered during training in the errorful condition with the total accuracy in the errorless condition. When we did this test we found that, although the errorful condition had a higher accuracy (errorful: $M = 0.40$, SD = 0.215; errorless: $M = 0.35$, SD = 0.228), the difference was not significant ($t(34) = 0.65$, $p = 0.518$). As Figure 3 shows, making errors during training didn't diminish test performance in the errorful condition as compared to the errorless condition even if we consider only those sentences wrongly answered during training.

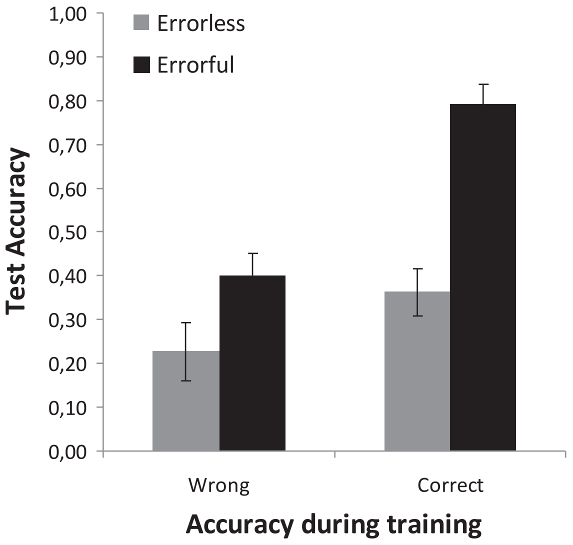


Figure 3. Average accuracy (proportion correct) during test in Experiment 1 as a function of training condition (errorless or errorful) and training accuracy (wrong or correct). Error bars represent the standard error of the mean.

We carried out a last analysis in Experiment 1 comparing vocabulary learning between the two conditions. As explained in the Methods section, the sentences followed the structure: *article – noun – verb – adjective.* When writing in English, as it was the test in Experiment 1, the article was always "the" and the verb was always "to be" which minimised the difficulty of the translation. For this reason, we concentrated in translation accuracy of nouns and adjectives only. We carried out a mixed ANOVA with the type of word (noun or adjective) as a within-subjects factor and type of training (errorful or errorless) as a between-subjects factor. The results are in agreement with the results of the general accuracy in the test reported in the manuscript and the means also point towards that direction (proportion accuracy with nouns: errorful $M = 0.59$, SD = 0.20; errorless $M = 0.48$, SD = 0.21; proportion accuracy with adjectives: errorful $M = 0.66$, SD = 0.18; errorless $M = 0.57$, SD = 0.22). However and also matching the general accuracy results, the effect of condition was not significant ($F(1, 34) = 2.45$, $p = .126$), and there was no interaction ($F(1, 34) = .23$, $p = .632$). There was an effect of type of word ($F(1, 34) = 12.99$, $p < .001$, $\eta_p^2 = .276$) which showed that both groups could translate more easily adjectives than nouns (mean accuracies with nouns and adjectives were 0.53 and 0.62 respectively). This effect of type of word is probably related to the greater variability among nouns than among adjectives, which can be seen in Appendix 1.

Experiment 2

The first experiment showed that making errors during a task of language learning doesn't decrease performance in a later test. The experiment, however, could not give information regarding the learning of more complex information (i.e. grammatical rules). The second experiment was designed to address this point and to replicate the main findings of Experiment 1.

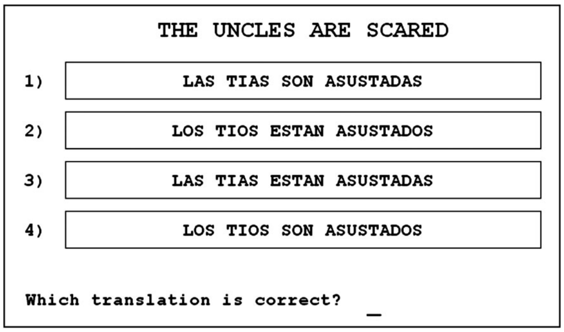
Method

Participants: The participants were 50 first-year students (38 female) from a Dutch public university who took part in the study in exchange for course credit. Their mean age was 20.62 (SD = 2.94). Participants could only join the study if they reported no previous knowledge of Spanish. After random assignment to the conditions, the distribution was 24 participants in the errorful and 26 in the errorless condition. The errorless group was further reduced to 25 after the exclusion of one participant who had a test accuracy 3.5 standard deviations below the group average.

Materials: In addition to the 48 sentences in Spanish used as stimuli in Experiment 1, we constructed three sets of 48 matching sentences showing errors as well as one set of correct translations into English. The error sentences belonged to one of three categories ("surface", "deep" or "both") depending on the type of grammatical error contained. The surface error was related with gender or number agreement between the noun and adjective in the sentence and it was considered such because the error can be detected by just attending to the endings of the words and without deep processing for meaning. The deep error was related with the use of the wrong verb (ser instead of estar or viceversa) and it is considered such because its detection requires processing the sentence at a deeper level. The third set contained sentences presenting both errors.

Procedure: The study took place in the same setting and followed the same procedure as Experiment 1. The training phase and set-up was identical to that of Experiment 1, the only difference was in the testing phase, which we proceed to describe now making a comparison with that of Experiment 1.

During a typical trial of the testing phase in Experiment1, a sentence in Spanish appeared on the screen and the participant was required to type the correct English translation on a box beneath the sentence, just as they did in the training phase. In contrast, during testing in Experiment 2, participants were presented with one sentence in English paired with four possible translations in Spanish and they had to choose the number corresponding to the correct translation. There was no time limit to respond and no feedback was given. The sentence in English appeared centred on top of the screen with the four translations directly beneath it and the prompt to introduce the response at the bottom (Figure 4). The same sentences were used for all participants but the order and the location on the screen where they appeared was randomised per participant and set so each type of sentence (correct, surface, deep or both) appeared an equal amount of times on each one of the four possible locations. The test, therefore, presented one sentence in English (translations of the original Spanish sentences used during training) and four possible translations into Spanish and the participant was asked to indicate the correct answer (see Figure 4). One of the answers contained the correct Spanish sentence and the other three



THE UNCLES ARE SCARED

1) LAS TIAS SON ASUSTADAS

2) LOS TIOS ESTAN ASUSTADOS

3) LAS TIAS ESTAN ASUSTADAS

4) LOS TIOS SON ASUSTADOS

Which translation is correct? _

Figure 4. Screenshot of the computer during the testing phase in Experiment 2. In the example, the correct option is number two and numbers one, three and four present surface, both and deep errors respectively.

alternatives showed one example of each type of error.

When the testing phase was finished, participants were asked whether they had learned or discovered any grammatical rules used in the construction of the sentences. If they answered affirmatively they were required to write a small description of the rule/s using the keyboard. No time limit was given for this task which also finished the study.

Results

As in Experiment 1, the training phase was divided into four blocks of 12 trials each in order to check the progression of accuracy through training. Figure 5 shows that participants in the errorless condition maintained a very high accuracy, although not perfect, throughout the training phase, whereas the accuracy in the errorful condition increased as training progressed. Again, participants in the errorless condition were also faster to finish training than participants in the errorful condition (means 370 (SD = 58.23) and 487 (SD = 61.07) seconds respectively, $t(47) = 6.87$, $p < 0.001$, $d = 1.96$).

Levene's test revealed that the data regarding accuracy in the test phase did not meet the assumption of equality of variances so the degrees of freedom were adjusted accordingly to perform the corresponding analyses. First, a *t*-test comparing accuracy during the test phase showed an advantage in accuracy for participants trained under errorful conditions ($t(28.209) = 2.87$, $p = 0.008$, $d = 0.77$; average test accuracies were 0.59 (SD = 0.206) and 0.47 (SD = 0.071) in the errorful and errorless conditions respectively). The advantage in accuracy for the errorful condition was also present when we compared sentences correctly answered during training (errorful: $M = 0.60$, SD = 0.239; errorless: $M = 0.46$, SD = 0.073; $t(27.092) = 2.84$, $p = 0.008$, $d = 0.66$) but not in sentences wrongly answered during training (errorful: $M = 0.57$, SD = 0.209; errorless: $M = 0.43$, SD = 0.372; $t(32.461) = 1.57$, $p = 0.124$). As in Experiment 1 and in order to offer a more balanced comparison in terms of base rates we compared accuracy in sentences wrongly answered during training in the errorful condition with total accuracy in the errorless condition. A *t*-test showed that the errorful condition had a higher accuracy (errorful: $M = 0.57$, SD = 0.209; errorless: $M = 0.47$, SD = 0.071; $t(28.047) = 2.46$, $p = 0.020$, $d = 0.46$) even when we limited the comparison to those sentences that were wrongly answered during training. As Figure 6 shows, making errors during training didn't seem to affect negatively the performance of the errorful condition during the test.

The structure of the test phase in Experiment 2 permitted the comparison between conditions of the error rate across type of error as an extra analysis in comparison with Experiment 1. For that purpose, we compared the number of errors through a series

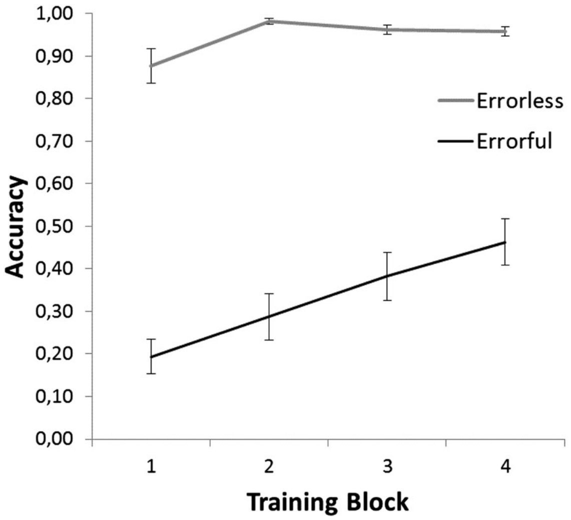


Figure 5. Average accuracy (proportion correct) during training in Experiment 2 as a function of training condition (errorless or errorful) and block of training (1–4). Error bars represent the standard error of the mean.

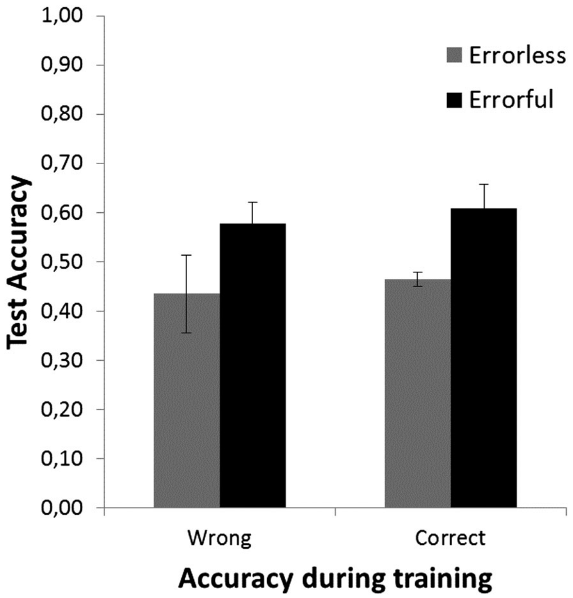


Figure 6. Average accuracy (proportion correct) during test in Experiment 2 as a function of training condition (errorless or errorful) and training accuracy (wrong or correct). Error bars represent the standard error of the mean.

of *t*-tests which showed that participants in the errorful condition made fewer surface errors (errorful: $M = 0.88$, SD = 1.11; errorless: $M = 1.96$, SD = 1.85; $t(47) = 2.46$, $p = 0.017$, $d = 0.70$) than participants in the errorless condition. After adjusting the degrees of freedom (Levene's test indicated unequal variances, $F = 28.37$, $p < 0.001$) a *t*-test also revealed differences between conditions in the number of deep errors (errorful: $M = 17.3$, SD = 9.06; errorless: $M = 21.6$, SD = 2.41; $t(26.126) = 2.23$, $p = 0.034$, $d = 0.64$).

The last analysis involved the answers to the explicit questions about knowledge of grammatical rules. The answers were coded as representing awareness of the surface or the deep rule and the proportion of participants who reported awareness was compared between conditions through a *z* test based on the normal approximation to the binomial distribution. In total, 19 participants reported awareness of the surface rule in the errorful condition and 15 did in the errorless condition (proportions 0.79 and 0.6 respectively). The result of the test ($z = 1.44$, $p = 0.074$, one-tailed) gave evidence of a marginally significant difference in awareness of the surface rule between conditions. Only one participant reported awareness of the deep rule so no test could be carried out to check for differences in awareness between the conditions.

General discussion

In two experiments we showed that making errors during a task of foreign language learning did not seem to harm performance in general. In both experiments, test accuracy in items wrongly answered during training in the errorful condition was equal or higher than the general accuracy in the errorless condition. Analyses of vocabulary learning in Experiment 1 revealed a similar pattern of results. Finally, participants in the errorful condition also made fewer mistakes during the test phase involving the deep rule and the surface grammatical rules. These last results might indicate a beneficial effect of making errors in the learning of grammar.

Our analysis of vocabulary learning in Experiment 1 showed that all participants learned more easily adjectives than nouns. This finding might be due to the greater variability among nouns than among adjectives. Some nouns were very similar in Spanish for all forms of masculine, feminine, singular and plural (e.g. niño, niña, niños, niñas, which respectively translate to boy, girl, boys and girls) whereas other nouns showed more variability (e.g. marido, esposa, maridos, esposas, which were translated as husband, wife, husbands, wives). This difference was not present among adjectives because the morphological distinction here was always limited to the endings (e.g. honesto, honesta, honestos, honestas). In addition, the association of adjectives was to one English word only (honest in the last example) which also simplified the learning.

There are now numerous studies showing that the commission of errors during a learning task does not necessarily have a detrimental effect on later test performance (Hays et al., 2012; Kang, Pashler, et al., 2011; Kornell et al., 2009). In fact, recent research shows that the advantage of retrieval practice extends even to clinical settings (Middleton, Schwartz, Rawson, & Garvey, 2015; Middleton, Schwartz, Rawson, Traut, & Verkuilen, 2016), a field where errorless learning used to be the most successful approach (Clare et al., 2000). Our results show that this effect can also be found using educationally relevant materials. If feedback is given immediately (Pashler, Cepeda, Wixted, & Rohrer, 2005) making an error does not have a negative influence in the re-test of that specific item. Guessing answers and making errors can be considered *desirable difficulties* (Bjork, 1994; Schmidt & Bjork, 1992) in the sense that they produce a deterioration of performance during the learning phase but retention and transfer are promoted. The present study clearly illustrates the idea: The errorful groups showed a lower performance during the learning phase than the errorless groups but outperformed them during the testing phase.

Bjork (1994) explains that desirable difficulties should help the learner to achieve an encoding of information that is " … part of a broader framework of interrelated concepts and ideas" (p. 188). This idea is in line with search set theory proposed by Kornell et al. (2009) and further supported by Grimaldi and Karpicke (2012; see also Kang, Gollan, and Pashler (2013) for supporting evidence using a measure of spoken vocabulary). When participants try to give an answer, a network of alternatives and related knowledge is activated. This network by definition must be greater in the errorful condition because it includes different alternatives (among them maybe the right response), whereas in the errorless condition, only the correct response is processed. Activation of multiple cues should facilitate the detection of commonalities and the inference of rules through the declarative language system, which is the main responsible for acquisition of grammar in adult language learning (Ullman, 2001, 2004).

In addition, simply making errors might also increase explicit language and learning of grammar. Krashen (1981) argued that explicit language learning develops through error correction and presentation of rules. Ellis (2005) proposed that the commission of errors during language learning makes the learner think about the language itself, rather than the content of the message. Meta-knowledge or knowledge about the knowledge is one of the features characterising explicit knowledge, as opposed to lack of meta-knowledge in implicit knowledge (Dienes & Berry, 1997). The analyses of the types of errors and the answers to the explicit questions in our Exp. 2 seem to support this view: Participants in the errorful condition showed fewer surface and deep errors and became more generally aware of the surface rule than participants in the errorless condition.

Two methodological issues have implications for our results and need to be addressed in this discussion. First, the sample size of Experiment 1 is not very large. This is due to the exclusion of a number of participants because they had previous knowledge of Spanish. This issue was avoided to some extent in Experiment 2 by preventing participants with knowledge of Spanish from joining the study in the first place. Power calculations[1] show that the achieved power in Experiment 1 was not very high (0.37). This problem was remedied in Experiment 2 which achieved a power (0.79) more in line with the values reported in the literature.[2] Nevertheless, we think that the issues of sample size and power must be taken into consideration when interpreting the described results. For example, Rowland (2014) shows that recognition tests normally produce smaller testing effects than recall tests. The fact that we found the opposite in our study could be related to the smaller power of the test in Experiment 1. It could be however that the contrast in results between our two experiments was not due to differences in power or type of test (recognition or recall) but to a difference in the direction of translation. In Experiment 1, participants did a recall test that involved translating from Spanish into English. In Experiment 2, they did a recognition test that involved translating from English into Spanish. The main reason to design Experiment 2 was to test Spanish grammar, which was not possible in Experiment 1. The unintended consequence, however, was the introduction of this confound between the two experiments.

In addition, although we controlled for the effect of previous knowledge of Spanish in Experiment 2, we think that this control could have been more exhaustive. For example, it was possible that mono-lingual English speakers took part in the study. As learning a second language seems to have a positive effect in the learning of any subsequent languages (Hirosh & Degani, 2017), these participants could have been at a disadvantage. Furthermore, being a language from the Romance family, Spanish is easier to learn for participants who already speak another related language (i.e. Italian, French, Portuguese) than for those who do not. Thus, previous knowledge of any language and especially Romance languages should be controlled more carefully in any follow-up study.

An interesting point to be addressed in future research could be the effect of delaying feedback. Grimaldi and Karpicke (2012) showed that delaying feedback was less effective than giving immediate feedback in a word-pairs associate task. However, Kornell (2014) showed that, when the retrieval attempt is meaningful, there is a benefit in delaying feedback. In a similar line, Guzmán-Muñoz and Johnson (2008) showed in a task of map learning that overall, delayed feedback can improve the development of relational knowledge. Considering that the paradigm used in the present study contains higher order knowledge (learning of grammatical rules), it could be argued that making errors and giving delayed feedback might promote its acquisition better than making errors and giving immediate feedback. More recent work by Krishnan, Sellars, Wood, Bishop, and Watkins (2018), who found a positive effect of evaluative feedback on vocabulary learning in a one-week delayed test, also seems to support this hypothesis.

In conclusion, our study showed that testing with immediate feedback is superior to just studying in a task of foreign language learning. This superiority

[1] We computed power of both two-tailed *t*-tests through an approximation to the Standard Normal Distribution (as suggested in Moore, McCabe, & Craig, 2009) and using the following parameters: Experiment 1: difference of means = 0.1, pooled *SD* = 0.22, $n_1 = 18$ & $n_2 = 18$, and $\alpha = 0.05$; Experiment 2: difference of means = 0.12, pooled *SD* = 0.15, $n_1 = 24$ & $n_2 = 25$; and $\alpha = 0.05$.

[2] A short literature review based on our own reference list and selecting studies reporting a similar test (independent samples t-test comparing study vs. test conditions) showed that the sample size, effect size and power of Experiment 2 were well within the values reported in previous studies.

comes from the activation of a greater network of related knowledge in the case of wrong answers. The greater activation translates into an increased facility to detect patterns among the knowledge, which, in the case of language learning might result in better learning of grammatical rules.

Disclosure statement

No potential conflict of interest was reported by the author.

References

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

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*, 53–68.

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

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478.

Chan, J. C. K., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553–571.

Clare, L., Wilson, B. A., Carter, G., Breen, K., Gosses, A., & Hodges, J. R. (2000). Intervening with everyday memory problems in dementia of Alzheimer type: An errorless learning approach. *Journal of Clinical and Experimental Neuropsychology*, *22*, 132–146.

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. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). San Diego, CA: Elsevier Academic Press.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, *4*, 3–23.

Dormann, T., & Frese, M. (1994). Error training: Replication and the function of exploratory behavior. *International Journal of Human-Computer Interaction*, *6*, 365–372.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, *6*, 217–226.

Duchastel, P., & Nungester, R. (1981). Long-term retention of prose following testing. *Psychological Reports*, *49*, 470.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*, 305–352.

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*, 407–418.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *40*, 104.

Glover, J. A. (1989). The 'testing' phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513.

Guzmán-Muñoz, F. J. (2009). *Errors, feedback and attentional load: differential involvement of memory systems as a function of conditions of learning* (Doctoral dissertation). University of Groningen, Groningen, The Netherlands. Retrieved from http://www.rug.nl/research/portal/publications/pub(9a43a367-b893-4a49-9b02-1e7dba854f93).html

Guzmán-Muñoz, F. J., & Johnson, A. (2008). Error feedback and the acquisition of geographical representations. *Applied Cognitive Psychology*, *22*, 979–995.

Hays, M. J., Kornell, N., & Bjork, R. A. (2012). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 290–296.

Heimbeck, D., Frese, M., Sonnentag, S., & Keith, N. (2003). Integrating errors into the training process: The function of error management instructions and the role of goal orientation. *Personnel Psychology*, *56*, 333–361.

Hirosh, Z., & Degani, T. (2017). Direct and indirect effects of multilingualism on novel language learning: An integrative review. *Psychonomic Bulletin & Review*, *25*, 892–916.

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, *38*, 1009–1017.

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is more effective than imitation for foreign language learning. *Psychonomic Bulletin & Review*, *20*, 1259–1265.

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*, 998–1005.

Kang, S. H. K., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*, 48–59.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*, 317–326.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science*, *331*, 772–775.

Kessels, R. P. C., & de Haan, E. H. F. (2003). Implicit learning in memory rehabilitation: A meta-analysis on errorless learning and vanishing cues methods. *Journal of Clinical and Experimental Neuropsychology*, *25*, 805–814.

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 106–114.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998.

Krashen, S. D. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon.

Krishnan, S., Sellars, E., Wood, H., Bishop, D. V. M., & Watkins, K. E. (2018). The influence of evaluative right/wrong feedback on phonological and semantic processes in word learning. *Royal Society Open Science*, *5*, 171496.

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger III, H. L. (2012). The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education*, *18*, 409–425.

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*, 1174–1181.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

Middleton, E. L., Schwartz, M. F., Rawson, K. A., & Garvey, K. (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1253–1261.

Middleton, E. L., Schwartz, M. F., Rawson, K. A., Traut, H., & Verkuilen, J. (2016). Towards a theory of learning for naming rehabilitation: Retrieval practice and spacing effects. *Journal of Speech, Language, and Hearing Research*, *59*, 1111–1122.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics*. New York, NY: Freeman & Co.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22.

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*, 3–8.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667.

Pressley, M., Symons, S., McDaniel, M. A., Snyder, B. L., & Turnure, J. E. (1988). Elaborative interrogation facilitates acquisition of confusing facts. *Journal of Educational Psychology*, *80*, 268–278.

Roediger III, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *Current issues in applied memory research* (pp. 13–49). New York, NY: Psychology Press.

Roediger III, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger III, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159.

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.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime Reference Guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Skinner, B. F. (1958). Teaching machines. *Science*, *128*, 969–977.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*, 784–802.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.

Terrace, H. S. (1963). Discrimination learning with and without 'errors'. *Journal of the Experimental Analysis of Behavior*, *6*, 1–27.

Tse, C., Balota, D. A., & Roediger III, H. L. (2010). The benefits and costs of repeated testing on the learning of face–name pairs in healthy older adults. *Psychology and Aging*, *25*, 833–845.

Tulving, E., Hayman, C. A., & Macdonald, C. A. (1991). Long-lasting perceptual priming and semantic learning in amnesia: A case experiment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 595–617.

Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, *30*, 37–69.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231–270.

Van Gog, T., & Sweller, J. (2015). Not new but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*, 247–264.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*, 899–905.

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review*, *18*, 518–523.

Woloshyn, V. E., Willoughby, T., Wood, E., & Pressley, M. (1990). Elaborative interrogation facilitates adult learning of factual paragraphs. *Journal of Educational Psychology*, *82*, 513–524.

Appendices

Appendix 1

List of sentences used as stimuli during training in Experiments 1 and 2:

EL MARIDO ES HONESTO
LA ESPOSA ES ALTA
LOS MARIDOS SON BAJOS
LAS ESPOSAS SON GUAPAS
LAS NIÑAS SON HONESTAS
LOS NIÑOS SON ALTOS
LA NIÑA ES BAJA
EL NIÑO ES GUAPO
EL MONJE ES TIMIDO
LA MONJA ES FEA
LOS MONJES ESTAN PREOCUPADOS
LAS MONJAS ESTAN ENFADADAS
LAS DAMAS SON TIMIDAS
LOS CABALLEROS SON FEOS
LA DAMA ESTA PREOCUPADA
EL CABALLERO ESTA ENFADADO
EL TIO ESTA CANSADO
LA TIA ESTA SORPRENDIDA
LOS TIOS ESTAN ASUSTADOS
LAS TIAS ESTAN ABURRIDAS
LAS SOBRINAS ESTAN CANSADAS
LOS SOBRINOS ESTAN SORPRENDIDOS
LA SOBRINA ESTA ASUSTADA
EL SOBRINO ESTA ABURRIDO
LA ABUELA ES HONESTA
EL ABUELO ES ALTO
LAS ABUELAS SON BAJAS
LOS ABUELOS SON GUAPOS
LOS PRINCIPES SON HONESTOS
LAS PRINCESAS SON ALTAS
EL PRINCIPE ES BAJO
LA PRINCESA ES GUAPA
LA MADRE ES TIMIDA
EL PADRE ES FEO
LAS MADRES ESTAN PREOCUPADAS
LOS PADRES ESTAN ENFADADOS
LOS HERMANOS SON TIMIDOS
LAS HERMANAS SON FEAS
EL HERMANO ESTA PREOCUPADO
LA HERMANA ESTA ENFADADA
LA CAMARERA ESTA CANSADA
EL CAMARERO ESTA SORPRENDIDO
LAS CAMARERAS ESTAN ASUSTADAS
LOS CAMAREROS ESTAN ABURRIDOS
LOS HIJOS ESTAN CANSADOS
LAS HIJAS ESTAN SORPRENDIDAS
EL HIJO ESTA ASUSTADO
LA HIJA ESTA ABURRIDA

Appendix 2

Answers during training and testing which missed entire words were considered incorrect by the computer and were left as such in our recoding of the data. Minor spelling mistakes were recoded as correct unless they affected the meaning of the sentence. In addition, some sentences were classified as incorrect by the computer due to the use of a wrong word which was in fact a synonym of the right word. For example, participants used “timid” instead of “shy”, “small” instead of “short”, and “grandpa” instead of “grandfather” or “brides” instead of “wives”. When this was the only mistake present, the sentence was recoded as correct.