Mem Cogn
DOI 10.3758/s13421-014-0408-z

Errorful and errorless learning: The impact of cue–target constraint in learning from errors

Emma K. Bridger · Axel Mecklinger

© Psychonomic Society, Inc. 2014

Abstract The benefits of testing on learning are well described, and attention has recently turned to what happens when errors are elicited during learning: Is testing nonetheless beneficial, or can errors hinder learning? Whilst recent findings have indicated that tests boost learning even if errors are made on every trial, other reports, emphasizing the benefits of errorless learning, have indicated that errors lead to poorer later memory performance. The possibility that this discrepancy is a function of the materials that must be learned—in particular, the relationship between the cues and targets—was addressed here. Cued recall after either a study-only errorless condition or an errorful learning condition was contrasted across cue–target associations, for which the extent to which the target was constrained by the cue was either high or low. Experiment 1 showed that whereas errorful learning led to greater recall for low-constraint stimuli, it led to a significant decrease in recall for high-constraint stimuli. This interaction is thought to reflect the extent to which retrieval is constrained by the cue–target association, as well as by the presence of preexisting semantic associations. The advantage of errorful retrieval for low-constraint stimuli was replicated in Experiment 2, and the interaction with stimulus type was replicated in Experiment 3, even when guesses were randomly designated as being either correct or incorrect. This pattern provides support for inferences derived from reports in which participants made errors on all learning trials, whilst highlighting the impact of material characteristics on the benefits and disadvantages that accrue from errorful learning in episodic memory.

Keywords Testing effect · Cued recall · Errorless learning · Errorful learning

The significant boost in memory retention for items that are tested rather than restudied during learning is one of the best characterized memory phenomena to date (Carrier & Pashler, 1992; Karpicke & Roediger, 2008). The retention advantage that this incurs, known as the testing effect, has been replicated across numerous materials including simple word lists (Carpenter & DeLosh, 2006), foreign language associates (Carrier & Pashler, 1992) and general knowledge facts (Carpenter, Pashler, Wixted, & Vul, 2008). The robustness of this phenomenon has led to repeated calls from empirical researchers for testing to be employed more frequently as a tool for boosting retention in educational settings (e.g., McDaniel, Roediger, & McDermott, 2007), calls that are supported by evidence of testing effects elicited in real classroom and learning environments (Carpenter, Pashler, & Cepeda, 2009; Carpenter, Sachs, Martin, Schmidt, & Looft, 2012; Larsen et al. 2009). One of the key facets of the argument for pushing testing as an instrument for learning as well as assessment (Metcalfe & Kornell, 2007) is the claim that the advantages that arise from recall during learning outweigh the losses that might arise from any errors that this could elicit. Put another way, this is the perspective that tests boost retention, even when they are errorful.

In one report, Kornell, Hays, and Bjork (2009) provided a degree of evidence in support of this assertion. Across a series of experiments, they compared the mnemonic consequences of two learning conditions: one in which participants incorrectly guessed items on (almost) every trial before they were told the correct item, and a second condition in which items were simply studied. By employing a condition in which testing principally

E. K. Bridger · A. Mecklinger
Department of Psychology, Experimental Neuropsychology Unit, Saarland University, Saarbrücken, Germany

E. K. Bridger (✉)
Department of Psychology, Saarland University, 66123 Saarbrücken, Germany
e-mail: e.bridger@mx.uni-saarland.de

Published online: 11 March 2014

elicited errors, the authors could determine the influence of making an error without concern for specific item characteristics that might have influenced the memorability of an item in the first place. In one representative task, Kornell and colleagues presented participants with a series of word cues and asked them to generate a semantic associate for each, before showing them the associate that they should actually learn for that item. In the overwhelming majority of instances, participants failed to correctly generate the to-be-learned word, but were nonetheless more likely to recall the correct answer in a final cued-recall phase than for pairs that they had studied for the same amount of time. Guessing with immediate corrective feedback thus appeared relatively beneficial for learning, even if it elicited a very high proportion of errors.

Given the practical implications this finding has for the endorsement of retrieval-based learning strategies even when the likelihood of making an error is high, subsequent reports have sought to define the boundary conditions under which an errorful learning benefit can be observed. These reports have addressed the impact of a variety of factors, including the number of guesses (Vaughn & Rawson, 2012), the timing of feedback (Kang et al., 2011; Kornell, 2014; Vaughn & Rawson, 2012), the presence or absence of a relationship between cue and target (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012), the plausibility of the self-generated error (Carpenter et al., 2012), the level at which a cue is processed and whether retrieval is semantic or episodic (Knight et al., 2012). In the present report, we built upon this work to extend the understanding of the impact of incorrect guesses on memory retention in two ways. First, we addressed the discrepancy between recent findings that have shown the relative benefits of incorrect guessing and another body of work in which errors made during learning were associated with a reduction in mnemonic performance (see Clare & Jones, 2008, for a review). In particular, we focused on the possibility that this discrepancy may in part be a consequence of the particular materials employed in the two literatures. In a second experiment, a possible limitation of the present paradigm's capacity to provide insight into the impact of making an error was addressed by modifying the task to ensure that testing did not always elicit an "incorrect" response. This allowed for a more accurate assessment of the impact of incorrect guessing, by ensuring that participants could not ignore their own self-generated responses, but instead had to learn to distinguish correct and incorrect guesses. In a final experiment, both of these aspects were brought together in one design.

Errorful versus errorless learning

According to some models of associative learning theory, learning comes about following the correction of an error signal (the difference between a predicted and actual outcome) over time (Rescorla & Wagner, 1972) and as such is most powerful when an error signal is present during learning. In line with this are data that show that the physiological correlates of error signals such as the error-related negativity (ERN: Holroyd & Coles, 2002) and the similar feedback-related negativity (FRN), which operate in line with basic principles of associative learning theory (Luque, López, Marco-Pallares, Càmara, & Rodriguez-Fornells, 2012) are associated with a reduction in the repetition of errors in some tasks (van der Helden, Boksem, & Blom, 2010). This perspective might lead one to predict that errorful learning conditions, should lead to better later memory performance. The extent to which these mechanisms can be applied to long-term episodic memory performance remains an open question, however, and alternative accounts state that retrieving during learning strengthens the associative links between cues and associated representations (e.g., Bjork, 1988; McDaniel, Kowitz, & Dunay, 1989). This viewpoint would predict that retrieving an incorrect answer should be detrimental to learning because it will enhance retrieval routes to erroneous responses. Related to this perspective are training procedures that emphasize the importance of errorless learning conditions, particularly for individuals with memory impairments such as in the early stages of Alzheimer's disease (Clare & Jones, 2008).

In the first influential report of this kind, Baddeley and Wilson (1994) compared final cued recall between amnesic patients with a variety of etiologies and younger and older control participants. During learning, participants were told either to write down words to learn (errorless learning) or to generate a number of words (i.e., "brother," "broom," and "brown") when given a word stem ("B–R–O") before being told which item to write down and learn (errorful learning). Whilst all participants were more likely to later correctly recall words from the errorless condition, amnesic patients' performance suffered especially in the errorful learning condition. The vulnerability of this group to errors during learning is thought to come about, at least in part, because of their inability to use explicit recollection mechanisms to identify and reject errors at test (Anderson & Craik, 2006). Although little investigation has focused on the impact of errorless learning conditions in normal populations (Kang et al., 2011), the data from Baddeley and Wilson's control participants, as well as from more recent reports that have borrowed heavily from their original paradigm (Hammer, Mohammadi, Schmicker, Saliger, & Münte, 2011; Heldmann, Markgraf, Rodríguez-Fornells, & Münte, 2008; Rodriguez-Fornells, Kofidis, & Münte, 2004), indicate that errorless learning conditions may also be relatively beneficial for young healthy participants. These findings clearly diverge with the advantage for errorful learning reported by Kornell et al. (2009). Whilst a close look at the particular tasks employed reveals a variety of factors that might account for this discrepancy, we

have good reasons to presume that this is in fact a consequence of the particular materials employed in the two cases.

Reports in which errorful learning conditions have led to a relative performance *decrease* in episodic memory performance of healthy individuals have used stem-completion tasks, in which word-stem cues are used to generate words. There are several reasons to suspect that this would lead to a difference in recall performance, as compared to the word pairs employed by Kornell et al. (2009). First, it appears that the presence of a semantic relationship between cue and target(s) predicts the extent to which errorful conditions will lead to a learning advantage over errorless read-only conditions (Grimaldi & Karpicke, 2012; Knight et al., 2012), and this relationship is necessarily absent between stem cues and their word targets. A stem-completion task also differs in the size of the set of possible answers through which one can search to provide an answer. Constraining the number of potential responses may disproportionately strengthen the representation of self-generated items, leading to greater later interference from these items if they are designated as errors. In line with this is one experiment reported by Grimaldi and Karpicke, in which later recall for a condition in which guessing was constrained by the stem of the 2nd word (i.e., TIDE–WA__) was significantly reduced relative to a read-only study condition. This decrease in performance came about because participants were more likely to remember "incorrect" guesses they had originally made at study during the final recall test. Whereas word-stem cues can elicit only a highly select group of words, the requirement to generate an associate for an arbitrary word, as is the case for weak semantic pairs, is constrained only by the limits of the participant's semantic knowledge. Changes in stimulus and task characteristics of this kind may determine whether incorrect guessing will be relatively advantageous or detrimental for learning at a given time.

It is not possible to establish the extent to which these characteristics determine the presence or absence of an errorful learning advantage in existing studies, because of changes in a variety of parameters across reports that, cumulatively, could have contributed to discrepancies of this kind. For example, one illustrative report from Rodriguez-Fornells et al. (2004) in which errorful learning led to poorer later memory performance, utilized a recognition memory test in which participants had to discriminate between old and new/lure items (see also Hammer et al., 2011; Heldmann et al., 2008, for comparable results from the same task). The presence of a final recognition rather than cued-recall task could influence the results in two ways. Firstly, the testing advantage has been shown to selectively boost recollection-based retrieval whilst leaving familiarity relatively unaffected (Chan & McDermott, 2007). The testing effect is thus observed primarily in recall tasks that are reliant upon recollection, rather than recognition tasks, in which familiarity may buffer the retrieval benefit. Secondly, the measures of recognition discrimination for the errorful and errorless conditions were confounded by difficulty: The errorless contrast entailed simple old/new discrimination, which is markedly easier than discriminating between words that have been generated by the participant but only one of which has been designated as "correct." Discrimination requirements were thus intrinsically more difficult in the errorful than in the errorless recognition condition and this alone could account for the accuracy advantage for errorless learning in these studies.

Experiment 1 represents an explicit test of the impact of errorful and errorless learning conditions on cued recall of cue–target associations that were either high in constraint (words generated from word-stem cues) or low in constraint (semantically associated word pairs). This was achieved using a paradigm based upon that originally employed by Kornell et al. (2009) and comprised randomly intermixed errorful and errorless trials during an intentional encoding phase. On errorless trials, participants saw cues and associated targets for just over 10 s, whereas on errorful trials the cue was presented for the initial 6 s whilst the participant provided an associated response. For the final 4 s, the correct to-be-learned cue–target combination was presented. The construction of the two stimulus sets was designed to reflect those employed in previous studies: These were either word pairs that were weakly semantically associated (e.g., Doktor–Pflaster; cf. Kornell et al., 2009) or word stems that had been selected via pretesting because they elicited two nouns with relatively high and approximately equal probabilities (e.g., Bir–Birne/Birke; cf. Rodriguez-Fornells et al., 2004). If testing is beneficial for learning even if it leads to an error, errorful learning should always lead to a relative cued-recall advantage over errorless learning, regardless of how constrained target retrieval is. If the learning outcomes of errorful conditions depend upon how constrained the association between cue and target is, however, the relative advantage for errorful learning will be seen for low- but not for high-constraint cue–target stimuli.

Experiment 1

Method

Participants and design A group of 48 native German speakers (29 female, 19 male; age range = 18 – 30 years) were recruited from the student population of Saarland University. Informed consent was required, payment was provided at a rate of €8/h or course credit, and participants were debriefed after the experiment. A 2 × 2 mixed design was employed with cue–target constraint as a between-subjects variable and learning condition (errorless vs. errorful learning) within subjects. One participant from the low-constraint group

was excluded because he or she failed to provide at least four correct answers in each condition.

Stimuli Word pairs were 60 weakly associated semantic pairs, each comprising two German nouns with a range of 4–11 letters in length. The strengths of associations for 40 of the pairs were quantified using the Noun Associates for German database (Melinger & Weber, 2006). This database represents the three associate responses provided by 50 native German speakers when presented with a list of approximately 400 German words. For each cue of a pair, a target was selected from this database if the likelihood with which it was generated from the first was less than .03. The mean proportion of occurrence for targets from this database was .008 (range = 0 –.03). For the remaining pairs, association strengths were taken from English translations of German words entered into the Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973). Pairs were only taken if the forward association strengths were less than .05 (mean = .014, range = 0 –.05). The frequency (Mannheim frequency per million: Baayen, Piepenbrock, & van Rijn, 1993) range of stimuli was variable (0–304) but did not significantly differ between cues (mean = 45) and targets (mean = 57), $t(59) = 1.27$, $p = .21$.

Word-stem stimuli were generated on the basis of the modal responses given by native German speakers to a series of three-letter word stems (e.g., BIR). A total of 86 raters were randomly allocated one of two lists each containing 140 word stems and were required to write down the first two German nouns beginning with these letters that came to mind. For each word stem, the probability of generating the first and second most common response (hereafter referred to as targets) was compared and only those stimuli for which the difference between these probabilities was no greater than .10 were selected. The 60 word stems whose modal responses had the highest probability were selected for Experiment 1. The mean probability of retrieving the most likely response was .33 (range = .15 –. 49) and the mean probability of the second most likely response was .26 (.10 – 46). First and second responses did not differ in word length, $t(59) = 1.39$, $p = .17$. The frequency (Mannheim per million) range of these words was variable (0–1,041) but did not significantly differ for the first (mean = 28) and second response (mean = 36), $t(59) = 0.46$. We found no significant differences in frequency between stimuli in the word-pair and word-stem tasks (all ps > .238).

The two classes of stimuli thus differed primarily in the degree to which targets were likely to be generated during initial retrieval. The high constraints of the word-stem stimuli inevitably increased the likelihood that participants would give a correct answer to these stimuli during the initial retrieval phase. In order to ensure that guesses were thus truly errorful, the word-stem task was programmed to replace correct answers with the 2nd word that preratings indicated was equally likely to be generated. Thus it was possible to ensure that all errorful trials led to an error for the high-constraint stimuli. Piloting indicated that on a minority of trials in the low-constraint condition, however, participants answered correctly (mean = .03, $SD = .03$). This proportion is comparable to that reported in previous studies (Grimaldi & Karpicke, 2012; Knight et al., 2012; Kornell et al., 2009). In order to ensure that the difference between the two stimuli was not a consequence of the likelihood of giving a correct answer at study, five additional "correct" filler trials were included in the high-constraint experiment, in which the to-be-learned item provided was identical to the participant's input. Later memory for these items was not tested.

Both the word pairs and word-stem stimuli were allocated to one of two lists of 30 items, matched for word length, frequency, strength of association/probability of word generation. The allocation of these lists to the errorful and errorless conditions was counterbalanced across participants.

Procedure Each testing session comprised a study, distractor and test phase and lasted approximately 40 min. The distractor task was an automated version of Unsworth's Ospan task (Unsworth, Heitz, Schrock, & Engle, 2005). This task was included both to maintain a meaningful interval (15 – 20 min) between study and test and to address the possibility that the sensitivity to the two learning conditions might interact with working memory capacity (Unsworth, 2009). No interactions of this kind were observed in any of the experiments reported here and these data will not be discussed further.

The study phase began with five practice trials in which the participant familiarized themselves with the timing parameters of the task. All trials began with a 1,500-ms blank screen. On errorless trials, this was followed by the cue and target presented vertically above one another, in the center of the screen for 10,300 ms. Errorful trials began with the presentation of the cue alone for 6,000 ms, during which participants were required to generate their own input by typing on a standard keyboard. For word pairs, participants were asked to create their own word pairs when they saw a single item on the screen and they were told that these word pairs should be semantically related (e.g., Whale–Mammal) but not strongly semantically related (e.g., Dog–Cat). For word stems, participants were told that they should type in a word that they thought most German students would give when shown this word stem. The cue and participant input were replaced by a blank screen for 300 ms, before being replaced by the correct cue–target combination for 4,000 ms. All words were presented in capital letters and errorless and errorful trials were randomly intermixed. In the final cued-recall test, participants were presented with each cue for 500 ms and were then presented with a blank screen for 9,000 ms during which they were required to type the target.

Results

Participants responded correctly on an average of .78 (2.6%) study trials of the word pairs (hereafter referred to as *low-constraint* cue–targets). These correct answers were excluded from further analysis for each participant, in line with the approach originally taken by Kornell et al. (2009; see also Grimaldi & Karpicke, 2012; Vaughn & Rawson, 2012). Answers correctly given at study were always correct at test, and thus removing these items led to an apparent decrease in final performance for the errorful condition. Figure 1 shows the mean final correct recall for the critical conditions in Experiment 1. A mixed ANOVA with the between-group factor Cue–Target Constraint and the within-group factor Learning Condition elicited a main effect of cue–target constraint [$F(1, 45) = 14.54$, $MSE = .038$, $p < .001$, $\eta_p^2 = .243$], following the higher performance level in the high-constraint condition, alongside a significant interaction [$F(1, 45) = 25.33$, $MSE = .009$, $p < .001$, $\eta_p^2 = .360$]. The reason for the interaction is clear; whereas errorful study conditions led to better performance than did errorless study conditions for low-constraint stimuli [$t(22) = 2.91$, $p < .01$, $d = 0.433$], errorful study led to relatively poorer performance for high-constraint stimuli [$t(23) = 4.21$, $p < .001$, $d = 0.956$]. Whilst the two levels of cue–target constraint did not significantly differ on accuracy for the errorful condition ($p = .19$), accuracy was significantly higher in the errorless high-constraint than in the errorless low-constraint condition [$t(45) = 5.44$, $p < .001$, $d = 1.62$]. Thus, incorrect guessing at study enhanced learning for low-constraint, but diminished learning for high-constraint, stimuli.

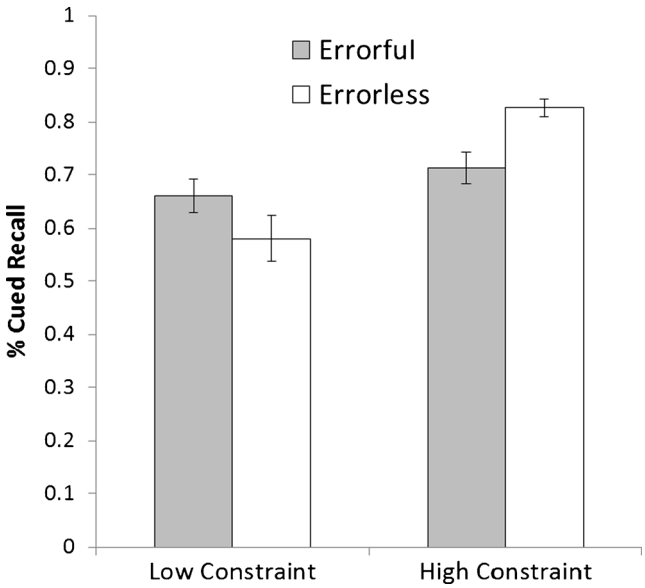

Fig. 1 Mean proportions correct in the final cued-recall test for the low and high cue–target constraint stimuli in Experiment 1. Error bars reflect ±1 standard error of the mean.

Errors made during the final recall test were also examined. For the errorless condition, these comprised errors of either omission or commission. In the errorful condition, commissions were further categorized as those that were self-generated (i.e., the same "error" given at study for that item) or other-commissions. Table 1 shows the proportions of these errors for each stimulus type, although subsequent analyses were limited to errors in the errorful condition. A mixed ANOVA with the between-group factor Cue–Target Constraint and the within-group factor Error Type (omissions, self-commissions, other-comissions) was conducted. Only interactions containing the factor Error Type are of interest here. We found a Cue–Target × Error Type interaction [$F(2, 90) = 18.55$, $MSE = .009$, $p < .001$, $\eta_p^2 = .292$]. This reflects the fact that participants were significantly more likely to make an omission [$t(45) = –2.78$, $p = .024$, $d = 0.83$; p values for all follow-up $t$tests are Holm–Bonferroni corrected: Holm, 1979] or an other-commission [$t(45) = –3.32$, $p = .004$, $d = 0.99$] for the low- than for the high-constraint stimuli at test. Self-commissions were significantly more likely for high- than for low-constraint stimuli, however [$t(45) = 4.35$, $p < .001$, $d = 1.30$]. These analyses reveal an important pattern: Participants were much more likely to make the same self-generated errors at test that they had made at study when stimuli were high-constraint than when they were low-constraint.

Discussion

The results from Experiment 1 replicate and extend the pattern reported in recent articles that have investigated the impact of errorful learning on later cued recall of semantically associated word pairs. For these items, guesses that were deemed errorful and replaced by a to-be-learned item were associated with higher later recall performance than errorless trials in which learned items were simply read on-screen. Errorful learning was associated with a significant decrease in performance, however, if the to-be-learned items were targets that were highly constrained by the associated cue. Examination of the pattern of errors made during test might provide some

Table 1 Proportions of omissions and total commissions (also separated into self- and other commissions) that were made at test in Experiment 1

			Low-Constraint	High-Constraint
Errorless	Omissions		.17 (.13)	.02 (.04)
	Commissions		.26 (.20)	.16 (.08)
Errorful	Omissions		.14 (.10)	.07 (.08)
	Commissions		.17 (.14)	.22 (.14)
		Self	.04 (.07)	.16 (.12)
		Other	.13 (.10)	.06 (.05)

Standard deviations are shown in parentheses.

insight into why this interaction comes about. In the high-constraint errorful condition, errors at test were more likely to be the same responses that participants initially gave during study, and this increase in self-generated errors accounts for the relative disadvantage in performance for the errorful condition for these stimuli.

Why might participants be able to exclude their own self-generated responses for low- but not for high-constraint stimuli? One reason is likely to be the number of possible responses that the respective cue types are associated with before learning. The very small set of possible targets that word-stem cues are associated with, by virtue of the way in which these stimuli are constructed, decreases considerably the number of possible answers that can be given at test. This small set size may strengthen the representations of erroneous responses and increase the likelihood that an incorrect answer given at test will be the same incorrect answer that was given at study. The high constraint on possible guesses is also likely to explain the main effect of performance between the two stimulus types: Guessing within a very small set of highly constrained responses is much more likely to lead to the correct answer.

Another possibility, and one that is not necessarily mutually exclusive, comes from the proportion of correct responses participants made during learning for the two stimulus types. In an effort to match testing conditions for the two constraint types, five additional "correct" trials occurred during learning for the word-stem condition. Although participants in the low-constraint task did make a small proportion of correct answers (2.6%), this was still somewhat less than the 14% (5/35) of test items that were correct at study in the word-stem condition. It is possible that the larger proportion of answers that were deemed correct in the high-constraint group may have led participants to be more likely to repeat self-generated answers at test because participants could remember during the test phase that their own responses had at times been correct. A necessary extension of this line of reasoning is that participants in the low-constraint task may have known not to repeat their own responses in the final test, because they learned that these were almost always wrong. In Experiment 2, the extent to which the errorful learning advantage might be a consequence of participants' ability to ignore their own responses was investigated.

Experiment 2

Examination of test errors made for low-constraint cue–target associations in Experiment 1 indicated that participants very rarely gave the same error at test that they made at study. It is not appropriate to assume that this came about because participants failed to remember these items, given what is known about the mnemonic advantages for self-generated responses. In line with this, Vaughn and Rawson (2012) have shown that participants remember their own guesses at test with high accuracy. If participants can remember their responses, why do they not tend to report them in the current task? The most obvious answer is that participants are aware that their own answers are never correct and explicitly withhold them during the final test. This possibility comes about because the task used in Experiment 1, and in previous reports, employs a condition in which almost all responses that participants make are deemed incorrect, in order to circumvent the problem of item characteristics (Pashler, Zarow, & Triplett, 2003). The outcome of this is that self-generated errors can never be confused with correct self-generated items. Instead, these responses could be used as an additional—and, presumably, potent—retrieval cue.

Experiment 2 comprises an explicit investigation of whether errorful learning of low-constraint cue–target associations still leads to a recall advantage over errorless trials when it occurs amongst a large number of trials on which guessing can also be "correct." To this end, the paradigm employed in Experiment 1 was modulated to ensure that on 50% of the test trials, participants were shown their own responses to retain for later learning. These trials are hereafter referred to as *errorless-generate* trials. If the errorful advantage previously observed for low-constraint stimuli is not a consequence of participants' ability to use a strategy in which they can discount all self-generated responses, and genuinely provides a learning benefit over errorless learning, then the errorful learning conditions should also lead to higher later recall relative to errorless study trials in Experiment 2.

Method

Participants A group of 27 native German speakers successfully completed Experiment 2a (12 female, 15 male), and 28 participants (22 female, six male) completed Experiment 2b for course credit or monetary payment (€8 per hour). One additional participant was excluded from the final analysis in Experiment 2a for failing to provide at least four correct answers in each experimental condition.

Stimuli and design Two versions of Experiment 2 were conducted, and the procedural difference between Experiments 2a and 2b is delineated further below. Stimuli in Experiment 2a comprised 90 weakly semantically associated word pairs, whilst in Experiment 2b they comprised 90 semantic triplets. Word pairs were constructed under identical constraints as those employed in Experiment 1. Triplets were created by adding a 2nd weak semantic associate to each cue–target pair employed in Experiment 2a, so that each cue had two targets with comparable association strengths. Word pairs and triplets were allocated to one of three lists of 30 items matched for

word length, frequency and strength of association, and these lists were counterbalanced across experimental conditions.

Procedure The trial parameters and task instructions in both Experiments 2a and 2b were identical to those used for low-constraint stimuli in Experiment 1 with the exception that 30 errorless-generate trials were added to both study and test. For these study trials, participants were required to input an associated word within the 6,000-ms input screen identical to the requirements during Errorful trials. For the final 4,000 ms of these trials, both the cue and participant's input were presented. The remaining 60 trials comprised 30 errorless and 30 errorful trials and all three trial types were randomly intermixed at study and test. As in Experiment 1, it remained possible for participants to occasionally guess correctly on errorful trials in Experiment 2a. Experiment 2b was identical, with the exception that the experiment was programmed such that, on errorful trials, the participants' input was compared with one of the two targets. If the participants' input matched this target (i.e., the participant made the correct response), the 2nd target would be presented to be learned. Which of the two targets was compared with the participants' input was counterbalanced across participants. In this way, it was possible to ensure that *all* errorful trials in Experiment 2b were truly errorful.

Results and discussion

Participants responded correctly on 0.93 (3.1%) errorful trials in Experiment 2a, and these were excluded from further analysis. Figure 2 shows the mean final correct cued recall for the three learning conditions for Experiments 2a and 2b. As would be expected, the final cued recall was greatest for the errorless-generate condition in both experiments. A mixed ANOVA with three levels of learning condition and the between-subjects factor Experiment (2a, 2b) revealed a main effect of learning condition [$F(2, 106) = 72.12$, $MSE = .013$, $p < .001$, $\eta_p^2 = .576$], and planned $t$tests revealed that recall in the errorless-generate condition was significantly higher than that in the other conditions [ts(54) > 8.88, ps < .001, ds > 1.33]. We also observed a significant increase in cued-recall performance for errorful learning relative to errorless learning [$t(54) = 3.08$, $p = .003$, $d = 0.378$]. No interactions with the Experiment factor emerged.

Table 2 shows the errors made at test for all conditions in Experiments 2a and 2b. The pattern of test errors in the errorful condition differs from that observed in Experiment 1 for low-constraint stimuli in the errorful condition in which self-commissions were significantly less likely to occur than other-commissions. In order to show this directly, test errors for low-constraint stimuli learned in the errorful condition were directly compared across the three experiments, using a mixed ANOVA with the factors Error Type (omissions, other-commission, and self-commission) and Experiment (1, 2a, 2b). A main effect of error type [$F(2, 150) = 4.63$, $MSE = .008$, $p = .013$, $\eta_p^2 = .058$] was moderated by an interaction [$F(4, 150) = 8.86$, $MSE = .008$, $p < .001$, $\eta_p^2 = .191$]. One-way ANOVAs (Exp. 1, 2a, 2b) conducted separately on each class of test error revealed that whereas the proportion of other-commissions did not interact with the Experiment factor [$F(2, 75) = 0.129$, $p = .879$), both omissions [$F(2, 75) = 6.145$, $MSE = .007$, $p = .003$] and self-commissions [$F(2, 75) = 13.117$, $MSE = .007$, $p < .001$] did interact with experiment type. Follow-up $t$tests showed that omissions were more likely [ts(>48) > 2.522, ps < .015, ds> 0.73] and self-commissions were less likely [ts(>48) > 4.051, ps < .001, ds > 1.17] in Experiment 1 than in Experiment 2a or 2b.

The data from Experiment 2 indicate that the relative mnemonic advantage for errorful learning is reduced when testing does not always lead to an error, although an advantage was nonetheless present. Participants were more likely to provide their own incorrect study answers in the final test phase when self-generated answers were no longer consistently deemed incorrect during learning. This is consistent with the idea that participants in Experiment 1 were able to retrieve their own generated responses and, being aware that they were always incorrect, they tended to withhold these responses. Important to note is that although the pattern of errors changed qualitatively from Experiment 1, the overall numbers of errors at test remained broadly comparable across the two experiments, and final cued recall for the errorful condition did not differ across Experiments 1 and 2 ($p > .31$).

Experiment 3

Experiment 2 revealed that the errorful learning advantage relative to items in the errorless condition was present for low-constraint stimuli even when participants could no longer ignore their own answers. The pattern of errors made at test for errorful items, however, did reveal a change in the kinds of errors that participants made in Experiment 2, relative to Experiment 1: Participants were more likely to make self-commissions and less likely to make omissions, in line with the notion that they were less inclined to withhold their own responses in Experiment 2. This observation may present a challenge to the use of designs in which all responses lead to an error because they may encourage participants to ignore their own incorrect responses. The strength of this challenge on the basis of these data alone is limited, however, because these contrasts were made across experiments, which additionally differed in terms of the number of items that participants had to learn (60 in Exp. 1, 90 in Exp. 2). One indication that this may have played a role comes from a comparison of the kinds of errors made to errorless items, which also differed across experiments: Participants generally made fewer

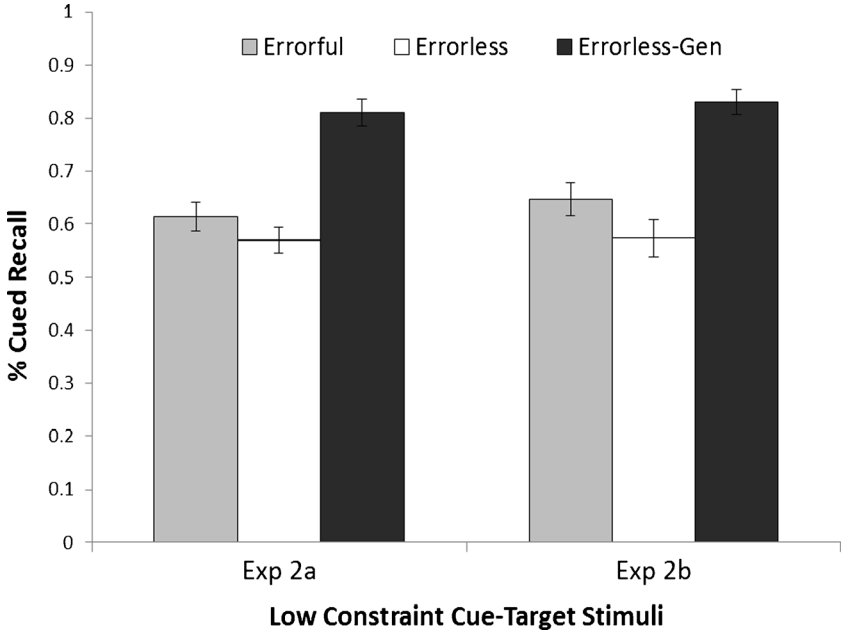

Fig. 2 Mean proportions correct in the final cued-recall test for the three learning conditions in Experiments 2a and 2b. Error bars reflect ±1 standard error of the mean.

omissions in Experiment 2 than in Experiment 1 ($p < .002$). In order to provide a more robust test of the impact of making an error, Experiment 3 was conducted, in which the likelihood of making an error during testing (50% or 100%) was explicitly manipulated whilst leaving all other factors comparable. A final motivation for Experiment 3 was to determine whether the likelihood of making an error would interact with cue–target constraints. In Experiment 1, high-constraint cue–targets elicited significantly more self-commissions than low-constraint cue–targets. Increasing the number of instances at study in which generated responses are correct may in turn further increase the number of self-commissions given for these items in the final test phase. In this case, high-constraint cue–targets might be more sensitive to an increase in the number of correct responses given at study than low-constraint stimuli, and errorful performance for these items would be even worse in the 50% error-likelihood condition.

Table 2 Proportions of omissions and total commissions (also separated into self- and other commissions) made at test in Experiment 2

			Exp. 2a	Exp. 2b
Errorless	Omissions		.13 (.11)	.09 (.10)
	Commissions		.30 (.14)	.34 (.16)
Errorless–generate	Omissions		.07 (.09)	.05 (.06)
	Commissions		.12 (.10)	.10 (.08)
Errorful	Omissions		.08 (.08)	.06 (.07)
	Commissions		.28 (.14)	.29 (.14)
		Self	.13 (.09)	.15 (.09)
		Other	.14 (.10)	.13 (.09)

Standard deviations are shown in parentheses.

Method

Participants and design A group of 32 native German speakers (25 female, seven male) successfully completed Experiment 3 for course credit or monetary payment (€8 per hour). A 2 × 2 × 2 mixed design was employed, with error-likelihood as a between-subjects variable and cue–target constraint (high, low) and learning condition (errorless, errorful learning) as within-subjects variables. Participants were randomly allocated to one of the two error-likelihood conditions (100% Errorful, 50% Errorful).

Stimuli The stimuli comprised the 90 weakly semantically associated word triplets used in Experiment 2b as well as 90 word-stem stimuli taken from the pilot rating study described in Experiment 1. For each word stem, the probabilities of generating the first and second most common response were compared, and only those stimuli for which the difference between these probabilities was no greater than .12 were selected. From these, the 90 word stems whose modal responses had the highest probability were selected for the experiment. All participants were required to learn 90 items in the word-stem and 90 items in the word-pair task. In the 100%-errorful condition, participants encountered 45 of each stimulus type in the errorful condition and 45 in the errorless condition. In the 50%-errorful condition, participants encountered 30 of each item in the errorful, 30 in the errorless condition and 30 items in the errorless-generate condition.

Separate counterbalanced versions of lists were made for each error-likelihood condition, in which lists were matched for word length, frequency, and strength of association.

Procedure The trial parameters and task instructions were identical to those used in Experiments 1 and 2. Participants completed a practice block at the beginning of the experiment in which they completed an equal number of practice study trials for the two stimulus types. These practice trials were blocked according to stimulus, and participants always completed three errorful and one errorless trial for each stimulus type. Participants in the 50%-errorful condition also completed one additional errorless-generate trial per stimulus type. During the study phase proper, stimulus type was blocked and the order of stimulus blocks was counterbalanced across participants. Participants completed blocks for both sets of stimulus types before moving onto the distractor tasks. These were a digit symbol task and the digit span task between study and test phase, which led to an average study–test interval of approximately 5 min. The test phase was also blocked according to stimulus type and kept in the same order as blocks presented at study. The experiment lasted on average 80 min.

Results

Figure 3 shows the final cued-recall patterns in Experiment 3. A 2 × 2 × 2 ANOVA with the factors Error Likelihood, Cue–Target Constraint, and Learning Condition (errorless, errorful) revealed a main effect of cue–target constraint [$F(1, 30) = 19.80$, $MSE = .028$, $p < .001$, $\eta_p^2 = .40$] and an interaction between learning condition and cue–target constraint [$F(1, 30) = 29.13$, $MSE = .012$, $p < .001$, $\eta_p^2 = .493$]. No significant interactions included the Error Likelihood factor (all ps > .654). As in Experiment 1, the reason for the Learning Condition × Cue–Target Constraint interaction is clear: Errorful learning led to better performance than did errorless learning for low-constraint stimuli [$t(31) = 4.06$, $p < .001$, $d = 0.56$], but to poorer performance for high-constraint stimuli [$t(31) = 4.30$, $p < .05$, $d = 0.70$], when collapsed across error-likelihood conditions. We found no significant difference in accuracy for the errorful condition across cue–target constraint ($p = .357$), whereas accuracy was significantly higher in the errorless high-constraint than in the errorless low-constraint condition [$t(31) = 5.86$, $p < .001$, $d = 1.25$]. No difference in final cued-recall performance emerged for the errorless-generate condition across cue–target constraint ($p = .274$), but performance in this condition was always better than in the errorful [$t(15) > 6.58$, $p < .001$, $d > 1.26$] and errorless [$t(15) > 2.78$, $p < .05$, $d > 0.60$] conditions.

Table 3 shows the errors made at test for all conditions in Experiment 3. A first 2 × 2 × 3 ANOVA contained the between-group factor Error Likelihood and the within-subjects factors Cue–Target Constraint and Error Type (omissions, self-commissions, other-commissions). This revealed a main effect of error type [$F(2, 60) = 28.39$, $MSE = .013$, $p < .001$, $\eta_p^2 = .486$] and an interaction between cue–target constraint and error type [$F(2, 60) = 12.59$, $MSE = .008$, $p < .001$, $\eta_p^2 = .296$]. This reflects the fact that participants were significantly more likely to make a self-commission [$t(31) = 2.76$, $p = .020$, $d = 0.56$] and significantly less likely to make an other-commission [$t(31) = –3.99$, $p < .001$, $d = 0.82$] for the high- than for the low-constraint stimuli at test. We found no significant difference in the likelihoods of making an omission across cue–target constraint ($p = .153$).

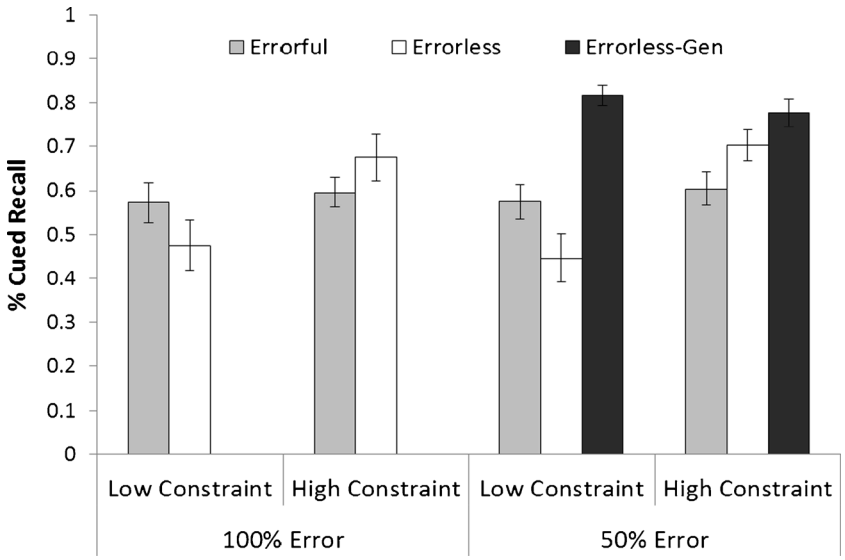

Fig. 3 Mean proportions correct in the final cued-recall test for the low and high cue–target constraint stimuli for the two different error-likelihood conditions in Experiment 3. Error bars reflect ±1 standard error of the mean

Table 3 Proportions of omissions and total commissions (also separated into self- and other commissions) made at test in the errorless and errorful conditions of Experiment 3

				Low-Constraint	High-Constraint
100% Error	Errorless	Omissions		.10 (.14)	.07 (.09)
		Commissions		.39 (.25)	.22 (.12)
	Errorful	Omissions		.09 (.09)	.05 (.06)
		Commissions		.32 (.22)	.34 (.11)
			Self	.16 (.13)	.23 (.12)
			Other	.16 (.10)	.11 (.07)
50% Error	Errorless	Omissions		.08 (.13)	.08 (.13)
		Commissions		.48 (.24)	.22 (.13)
	Errorful	Omissions		.05 (.05)	.05 (.06)
		Commissions		.38 (.16)	.34 (.16)
			Self	.19 (.11)	.26 (.13)
			Other	.19 (.13)	.09 (.06)

Standard deviations are shown in parentheses.

Discussion

The results of Experiment 3 reveal two important outcomes. First, the data represent a direct within-subjects replication of the findings from Experiment 1: The extent to which errorful learning is beneficial depends upon the constraints of the cue–target association. A moderate increase in final cued-recall performance emerged for low-constraint cue–target associations learned under errorful conditions, whereas these same conditions led to worse performance for highly constrained cue–target associations. The second insight drawn from the Experiment 3 data is that changes in the likelihood of making an error at study appeared to have no impact on later cued-recall performance. This null effect held across different cue–target constraints, and we found no evidence that error likelihood had an impact on the kind of errors made at test. This pattern provides evidence that previous observations of an errorful learning advantage for semantically related stimuli are unlikely to be a consequence of the fact that participants could ignore all of their own self-generated responses. This pattern is clear, but it doesn't explain the change in the pattern of errors at test (i.e., ratio of omissions to self-commissions) for low-constraint stimuli from Experiment 1 to 2. The principal remaining difference between Experiments 1, 2, and 3 is the overall number of items that participants had to learn, which increased in each successive experiment (Exp. 1 = 60 items, Exp. 2 = 90, Exp. 3 = 180). One possibility is that the smaller set of items to be learned in Experiment 1 led participants to be more conservative overall when responding in situations in which they were uncertain. Given a possible low-confidence response, one might be more certain about whether or not that item was present within a list of studied items if that list is relatively short. When the list goes beyond this size, it may be less easy to determine the likelihood with which a low-confidence response was within the list of studied items, making participants relatively more liberal with uncertain responses.

General discussion

The importance of determining the consequences of making an error during learning is reflected in the recent spike of studies in which the conditions under which errorful learning is beneficial for learning have been investigated (Grimaldi & Karpicke, 2012; Knight et al., 2012; Vaughn & Rawson, 2012). The data from Experiments 1 and 3 add to these reports by demonstrating that a central factor determining whether errorful learning conditions will be beneficial or detrimental to performance is how constrained targets are by a given retrieval cue, and that guessing may in fact be detrimental when the target information is highly constrained by a cue. The findings from Experiments 2 and 3 indicate that although the errorful learning advantage for low cue–target constraints may be smaller when test conditions do not always lead to an error (see Exp. 2a), this advantage is still observed when participants are required to determine between self-generated correct and incorrect responses.

Experiments 1 and 3 comprise an explicit endeavor to connect the distinct patterns observed in one set of recent reports (Kang et al., 2011; Kornell et al., 2009) with other work based on the errorless-learning paradigm (Hammer et al., 2011; Rodriguez-Fornells et al., 2004), by focusing upon the stimulus characteristics employed in the respective literatures. Figure 4 depicts two components of the relationship between the two types of cues and target answers, which are presumed to contribute to the impact that errorful learning has on final recall: specifically, the number of associates and the strength of association between cue and target, characterized in the figure by the thickness of connecting lines. Word-stem items were specifically constructed to lead to the generation of two words with relatively high and equal probabilities, and thus the strength of activation between the cue and target for these items is necessarily greater than that between weakly associated words, for which multiple associates exist (upper panel). After errorless learning in which the cue and target are presented together to be studied, the to-be-learned item should have a relative increase in the strength of its association (see the shaded target nodes in Fig. 4). This increase in activation is a function of both the preexisting association strengths and the number of

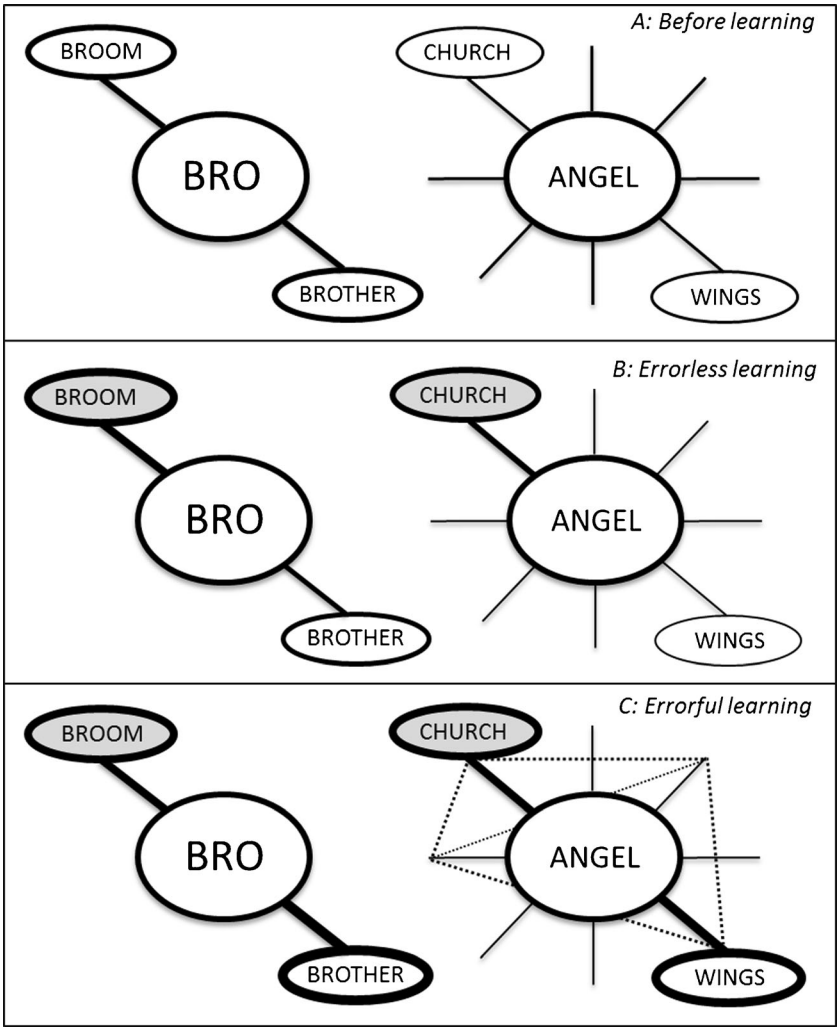

Fig. 4 Schemata representing the strength of activation between cues and targets for high and low cue–target constraint stimuli before learning (upper panel), after errorless learning (middle panel), and after errorful learning (lower panel) when the presence of preexisting semantic relationships is taken into account. Shading denotes the correct to-be-learned target, and dotted lines represent hypothetical associations between members of a semantic network. The strength of associations between the cues and targets is represented by the thickness of the adjoining lines.

potential items associated with each cue. The latter stipulation is comparable to the architecture of spreading activation in Reder et al.'s source-of-activation confusion model (Reder et al., 2000; Reder, Paynter, Diana, Ngiam, & Dickison, 2007), which states that the more links are associated with a particular cue, the less activation will spread along each particular link. It is also consistent with data indicating that recall is lower for items with relatively more associates than for items with fewer associates (see Nelson, Schreiber, & Xu, 1999). Words derived from word stems thus benefit from a single study presentation to a greater extent than do weakly associated words. This is in line with the marked difference in cued recall following errorless learning for the two different materials reported in Experiment 1. Strength of activation is assumed to increase to a greater extent following retrieval than following studying alone, so that incorrect answers given in the errorful learning condition have the highest overall strength at the end of a study trial (see the unshaded nodes in the lower panel of Fig. 4). For high-constraint cues, this marked increase in activation leads to considerable interference from the self-generated item at test. This is represented in the present data by the relative decrease in accuracy for errorful learning of high-constraint stimuli, alongside the high proportion of self-commissions for these items observed in Experiments 1 and 3. Such interference from self-generated errors is also likely to operate during the errorful learning of word pairs, although it should not lead to such a high level of self-commissions, because the number of potential responses that can be given at test is not so highly constrained. Further research could test these predictions using word-

stem stimuli that differed in their numbers of associates only (i.e., the word-stem cue "BAL" can generate fewer candidate words than can "BA"), in order to determine whether errorful learning is significantly worse for more- than for less-constrained cues.[1]

A simple increase in the number of preexperimental associates cannot account for a benefit in learning during errorful responding, however, if simply being shown the correct answer leads to an increase in activation strength equivalent to that established during errorless learning. If this is the case, then errorful learning of word pairs should lead either to equal performance or to a decrease in accuracy relative to errorless learning, which the present data show not to be the case. The advantage in correctly recalling a to-be-learned word could be explained if the presence of a preexisting semantic network between word pairs were taken into account (Grimaldi & Karpicke, 2012; Knight et al., 2012), and this is represented by the additional dotted lines in the lower panel of Figure 4. Such a network might increase the likelihood that the correct answer is remembered at test after an errorful response via two complementary routes: a relatively automatic buildup of activation within an associative network (Collins & Loftus, 1975) leading to an increase in activation for the correct answer (Grimaldi & Karpicke, 2012), as well as the use of explicit retrieval cues to help the recovery of the correct item, as specified by the mediator hypothesis of the testing effect (Carpenter, 2011; Pyc & Rawson, 2010). This observation means that where some form of relationship does not already exist, errorful learning conditions will not necessarily be advantageous to learning, and this in turn is consistent with data showing an errorful advantage for cued recall when the stimuli comprise semantically related but not unrelated word pairs (Grimaldi & Karpicke, 2012; Knight et al., 2012). A final point to note, however, is that despite the assumed differences concerning the underlying cue–target networks in the cases represented here, when self-generated responses were judged correct in the errorless-generate condition, performance was comparable across cue-constraint conditions. Thus, both cue-constraint types showed equivalent generation effects, and differences appeared only when self-generated error conditions were contrasted with the errorless learning condition (see Fig. 3). The term *errorless learning*, as it has been employed here, borrows heavily from Baddeley and Wilson's (1994) definition, in order to investigate how these established learning conditions interact with stimulus characteristics, but it may be more appropriate to consider the errorless condition here as being comparable to a read/restudy condition. Another way of considering errorless learning is as a learning environment that requires active participation from the learner, but for which sufficient support is available to ensure that responses are never incorrect. Insofar as the errorless-generate condition here meets this criterion, it would appear that active participation that is always correct is similarly beneficial for stimuli, regardless of levels of cue–target constraint. Given this definition, errorless learning is always better than errorful learning, regardless of stimulus type.

From an educational perspective, determining that an errorful learning advantage, as it has been principally defined here, depends on the extent to which to-be-learned stimuli support meaningful semantic elaboration (see Knight et al. 2012, and Kornell, 2014, for direct examples of this) may help set important boundary conditions for situations in which errorful guessing can be recommended: Rather than being conducive to the learning of new information, it may be limited to the strengthening of connections within an already-existing network. The present data indicate that in cases in which the to-be-learned information comprises one of a very small number of possible options, errorful conditions may especially hamper learning. One example of this is encountered when learning a second language with grammatical gender, for which the correct answer would be one of two or three possibilities. Grammatical gender is a widespread linguistic phenomenon common to most Indo-European languages (Grüter, Lew-Williams, & Fernald, 2012) that is particularly difficult for second language learners (Sabourin, Stowe, & De Haan, 2006). An example would be deciding whether the German word for "traffic" (*Verkehr*) takes a neutral, feminine, or masculine gender. The present high-constraint cue–target data indicate that incorrect guessing within this small pool of items might in fact impair the ability to learn the correct response, and that this might be one factor that makes learning of this kind so difficult.

One possible point of concern that the present data address directly is that the frequency with which errors are made during test does not appear to influence the errorful learning advantage for those stimuli for which it is observed. This provides important validation for those reports in which errorful learning has been investigated with paradigms wherein self-generated responses are deemed to be always incorrect. Nonetheless, in line with findings that have indicated that surprising incorrect feedback decreases the likelihood that an error will be repeated (Butterfield & Metcalfe, 2001), further research may seek to employ parametric manipulations of error likelihood in order to determine the point at which errors are sufficiently rare such that they always boost performance. A final point concerns the general increase in the number of commission errors in the final cued-recall task in Experiments 2 and 3, as compared to Experiment 1. Above we considered the possibility that this is a consequence of the total number of items that participants had to learn. It is known that participants can employ metamemorial monitoring processes to maximize memory accuracy by choosing whether or not to withhold responses (Koriat & Goldsmith, 1996). It may be

[1] We thank an anonymous reviewer for this suggestion

that when participants are required to learn very large lists, they become more liberal with low-confidence responses. Although further research will be required to determine whether list length can indeed drive response bias in cued-recall tasks, this nonetheless highlights the importance of determining factors that influence the likelihood of guessing at test, because incorrect guesses in turn constitute errorful learning. The present data suggest that whether or not these incorrect guesses will hinder later learning depends upon the kind of stimuli that one is learning.

Author note This research was supported by the German Research Foundation under Grant No. DFG-IRTG-1457, and was conducted in the International Research Training Group "Adaptive Minds," hosted by Saarland University, Saarbrücken (Germany). We thank Hubert Zimmer for valuable discussion on this topic, as well as Marie Schwartz, Leon Markelis, Katharina Jung, and Stefanie Kolb for assistance with stimulus preparation and data collection.

References

Anderson, N. D., & Craik, F. I. M. (2006). The mnemonic mechanisms of errorless learning. *Neuropsychologia, 44*, 2806–2813. doi:10.1016/j.neuropsychologia.2006.05.026

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396–401). New York: Wiley.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1491–1494. doi:10.1037/0278-7393.27.6.1491

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

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

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

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 77*, 760–771. doi:10.1002/acp.1507

Carpenter, S. K., Sachs, R. E., Martin, B., Schmidt, K., & Looft, R. (2012). Learning new vocabulary in German: the effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin & Review, 19*, 81–86. doi:10.3758/s13423-011-0185-7

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

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431–437. doi:10.1037/0278-7393.33.2.431

Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review, 18*, 1–23. doi:10.1007/s11065-008-9051-4

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428. doi:10.1037/0033-295X.82.6.407

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*, 505–513. doi:10.3758/s13421-011-0174-0

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28*, 191–215.

Hammer, A., Mohammadi, B., Schmicker, M., Saliger, S., & Münte, T. F. (2011). Errorless and errorful learning modulated by transcranial direct current stimulation. *BMC Neuroscience, 12*, 72–79. doi:10.1186/1471-2202-12-72

Heldmann, M., Markgraf, U., Rodríguez-Fornells, A., & Münte, T. F. (2008). Brain potentials reveal the role of conflict in human errorful and errorless learning. *Neuroscience Letters, 444*, 64–68.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*, 679–709. doi:10.1037/0033-295X.109.4.679

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40*, 514–527. doi:10.3758/s13421-011-0167-z

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. doi:10.1037/a0021977

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

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh, UK: Edinburgh University Press.

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66*, 731–746. doi:10.1016/j.jml.2011.12.008

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517. doi:10.1037/0033-295X.103.3.490

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. doi:10.1037/a0033699

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. doi:10.1037/a0015729

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

Luque, D., López, F. J., Marco-Pallares, J., Càmara, E., & Rodríguez-Fornells, A. (2012). Feedback-related brain potential activity complies with basic assumptions of associative learning theory. *Journal of Cognitive Neuroscience, 24*, 794–808. doi:10.1162/jocn_a_00145

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory & Cognition, 17*, 423–434. doi:10.3758/BF03202614

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200–206. doi:10.3758/BF03194052

Melinger, A., & Weber, A. (2006). Database of Noun Associations for German. Retrieved from www.coli.uni-saarland.de/projects/nag/

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14*, 225–229. doi:10.3758/BF03194056

Nelson, D. L., Schreiber, T. A., & Xu, J. (1999). Cue set size effects: Sampling activated associates or cross-target interference? *Memory & Cognition, 27*, 465–477. doi:10.3758/BF03211541

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1051–1057. doi:10.1037/0278-7393.29.6.1051

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335. doi:10.1126/science.1191465

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 294–320. doi:10.1037/0278-7393.26.2.294

Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., & Dickison, D. (2007). Experience is a double-edged sword: A computational model of the encoding/retrieval trade-off with familiarity. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use (The Psychology of Learning and Motivation* (Vol. 48, pp. 271–312). London, UK: Academic Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Rodriguez-Fornells, A., Kofidis, C., & Münte, T. F. (2004). An electrophysiological study of errorless learning. *Cognitive Brain Research, 19*, 160–173. doi:10.1016/j.cogbrainres.2003.11.009

Sabourin, L., Stowe, L. A., & De Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research, 22*, 1–29.

Terrace, H. S. (1963). Discrimination learning with and without "errors. *Journal of the Experimental Analysis of Behavior, 6*, 1–27. doi:10.1901/jeab.1963.6-1

Unsworth, N. (2009). Examining variation in working memory capacity and retrieval in cued recall. *Memory, 17*, 386–396. doi:10.1080/09658210902802959

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505. doi:10.3758/BF03192720

van der Helden, J., Boksem, M. A. S., & Blom, J. H. G. (2010). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex, 20*, 1596–1603. doi:10.1093/cercor/bhp224

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19*, 899–905. doi:10.3758/s13423-012-0276-0