DOI 10.3758/s13421-014-0454-6

Why does guessing incorrectly enhance, rather than impair, retention?

Veronica X. Yan · Yue Yu · Michael A. Garcia · Robert A. Bjork

Published online: 14 August 2014

Abstract The finding that trying, and failing, to predict the upcoming to-be-remembered response to a given cue can enhance later recall of that response, relative to studying the intact cue–response pair, is surprising, especially given that the standard paradigm (e.g., Kornell, Hays, & Bjork, 2009) involves allocating what would otherwise be study time to generating an error. In three experiments, we sought to eliminate two potential heuristics that participants might use to aid recall of correct responses on the final test and to explore the effects of interference both at an immediate and at a delayed test. In Experiment 1, by intermixing strongly associated to-be-remembered pairs with weakly associated pairs, we eliminated a potential heuristic participants can use on the final test in the standard version of the paradigm—namely, that really strong associates are incorrect responses. In Experiment 2, by rigging half of the participants' responses to be correct, we eliminated another potential heuristic—namely, that one's initial guesses are virtually always wrong. In Experiment 3, we examined whether participants' ability to remember—and discriminate between—their incorrect guesses and correct responses would be lost after a 48-h delay, when source memory should be reduced. Across all experiments, we continued to find a robust benefit of trying to guess to-be-learned responses, even when incorrect, versus studying intact cue–response pairs. The benefits of making incorrect guesses are not an artifact of the paradigm, nor are they limited to short retention intervals.

Keywords Testing · Forgetting · Generation · Errors

V. X. Yan (✉) · Y. Yu · M. A. Garcia · R. A. Bjork
Department of Psychology, University of California, Los Angeles, CA 90095, USA
e-mail: veronicayan@ucla.edu

An abundance of research on testing and generation effects has shown that the act of retrieval is a learning event—and often a powerful learning event—in the sense that the retrieved information becomes more retrievable in the future than it would have been otherwise (see, e.g., Roediger & Karpicke, 2006). The retrieval processes triggered by testing are, therefore, opportunities for learning—a basic fact about human learning that is often not appreciated or, at least, is underappreciated, by students (see, e.g., Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2007).

Testing effects and generation effects, however, typically refer to the consequences of successful retrieval or generation. One justifiable concern about testing or generation is that what is retrieved, whether correct or incorrect, will be learned: That is, by virtue of the very power of retrieval as a learning event, it seems likely that any errors that are produced will persist. One influential school of thought, for example, inspired by Skinnerian principles of learning, has emphasized "errorless learning" procedures (Skinner, 1958; Terrace, 1963), and a number of studies have, in fact, shown that initially incorrect responses often persist on subsequent tests (e.g., Cunningham & Anderson, 1968; Elley, 1966; Kaess & Zeaman, 1960; Marsh, Roediger, Bjork, & Bjork, 2007). Additionally, generating errors before being given feedback mirrors a classic A-B/A-D interference paradigm (e.g., Briggs, 1954), in which researchers have found that participants do, indeed, become more likely to output the initial "B" response as the retention interval increases.

The picture, though, is not so clear. Other studies investigating the effects of errors on multiple-choice tests (e.g., Butler, Marsh, Goode, & Roediger, 2006), for example, have shown no effect of generating errors, and other recent—and not so recent—findings suggest that there might, in fact, be *benefits* of trying to generate a correct response, even when the effort fails.

That even failed efforts to generate a to-be-remembered response might have benefits is suggested by the results of early research by Slamecka and Fevreiski (1983). Participants were presented with a list of related cue–target word pairs and were asked to say the target word aloud. In a *study-only* condition, the participants were shown the intact pair (e.g., *pursue–avoid)*; in a *generate* condition, participants were shown the full cue word together with a fragment of the target word (e.g., *pursue–av–d)*. If they failed to generate the target word within a 4-s interval, they were provided with corrective feedback immediately for 3 s. On a subsequent free recall test of the targets, there was a benefit of *generate* over *study-only*, even when only those items for which participants failed to generate the correct response were examined. The authors argued that failed generations were, in fact, incomplete generations, where semantic features, but not surface features, were processed.

In Slamecka and Fevreiski's (1983) study, however, 93 % of the errors were errors of omission, not errors of commission, so their findings leave open the possibility that generating overt errors has negative, not positive, effects. Recently, though, Kornell, Hays, and Bjork (2009), using a procedure in which participants' guesses of to-be-learned responses are wrong with high probability (thus, eliminating differences between items in the errorful and errorless conditions—a confound in some previous studies), extended the finding to cases where participants do not simply omit responses, but produce errors. Their results suggest that producing errors, at least under some circumstances, enhances subsequent learning.

Kornell et al.'s (2009) findings have stirred considerable interest, not only because producing incorrect guesses does not seem, intuitively, to be a good learning technique, but also because their specific procedure involved taking what would otherwise be study time to predict (erroneously) an upcoming to-be-learned response. In the *guess-first* condition of their Experiment 4, for example, participants were shown cues such as *Whale:* ______ for 8 s and were asked to predict the upcoming to-be-learned associate of that cue. Immediately after, they were then shown the cue together with the to-be-learned response (*Whale: Mammal*) for 5 s (97 % of the guesses were incorrect, and the trials on which guesses matched the target were removed from analyses). In their study-only condition, on the other hand, pairs such as *Whale: Mammal* were shown for the full 13 s. The guess-first condition produced better later recall of the correct target than did the study-only condition, despite the shorter study time and the reasonable expectation that generating a competing associate would create proactive interference. Kornell et al.'s basic finding has now been replicated by a number of other investigators (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Vaughn & Rawson, 2012), as well as with foreign language learning (Potts & Shanks, 2014) and more semantically rich text passages (Richland, Kornell, & Kao, 2009) and trivia facts (Kornell, 2014).

Questions and issues motivating the present research

Why do we not find interference in these experimental paradigms? In the present series of experiments, we seek to address two issues: (1) that the experimental paradigm design allows participants to distinguish between their guess and the correct answer at the time of the final test, and (2) whether the guess-first benefit will be maintained or whether the generated guesses will interfere with target recall at a longer retention interval.

One explanation of the benefits of guessing incorrectly is that a participant's incorrect guess acts as a mediator between the cue and the correct response. An assumption that underlies this explanation is that learners have a means of knowing, at the time of the final test, which response—the one they generated or the one they then studied—is the correct response.

In Experiment 1, we set out to examine whether a feature intrinsic to Kornell et al.'s (2009) paradigm might play a key role in learners being able to make that judgment. Because Kornell et al. wanted to examine whether making incorrect guesses would help or hinder learning, they chose weak associates of the cue word as to-be-learned response targets—that is, words that were unlikely to come to mind and be guessed in advance by participants. In Experiment 1, we explored whether participants in prior experiments may have been able to use the fact that generated errors tended to be strong associates of the cue words, whereas target responses were always weak associates of a given cue. Could participants have mitigated interference at the final test between competing responses, generated errors and targets, by learning that targets are weak associates? We nullified that possible heuristic in Experiment 1 by designing the materials so that the correct answer for half the pairs was a strong associate of the cue word.

In Experiment 2, we sought to nullify another possible heuristic that participants could be using in this paradigm: that their guesses are always wrong. The errorful generation paradigm—as used by Kornell et al. and in subsequent follow-up studies (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012; Potts & Shanks, 2014; Vaughn & Rawson, 2012)—ensures that the guess is almost always wrong, leaving open the possibility that when presented with the cue at final test, participants are able to simply select whatever response they did not generate for themselves.

Therefore, we rigged Experiment 2 so that, in one condition, half of participants' guesses were always deemed to be correct, and compared the benefit of making errors in this condition with the original condition where just about all the guesses were incorrect.

Variations of the original paradigm have been investigated to test different theories as to why there is a benefit of generating incorrect guesses, and these theories are further discussed in the General Discussion section. Despite variations on this original design, however, whether participants could use a heuristic remains an open question. One variant (e.g., Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013) found that delaying feedback of the correct answer eliminates the benefit of making incorrect guesses. While one explanation is that delaying feedback means that the correct target is not encoded into an activated semantic network, it could also be that having first generated guesses to all the guess-first word pairs before receiving the correct answer makes it more difficult for learners to recognize that all the correct responses are less obvious associates of the cue or even that all their initial guesses are wrong. Another variant on the original design showed that the benefit of generating responses was eliminated when participants' guesses were constrained to a particular word (through the provision of a two-letter stem—e.g., *tide–wa*____; Grimaldi & Karpicke, 2012). By constraining the guess to one obvious target response, the experimenters created a very different task than is experienced by participants making unconstrained anticipatory "guesses." Instead of interpreting the constrained generations as "wrong answers," participants may simply interpret them as other correct answers that are simply not required on the later test.

Experiment 3 was designed to examine whether the ability of participants to discriminate at the time of test between the response they guessed and the actual correct responses depends on the retention interval to the final testing being relatively short. Prior studies have used very short retention intervals in which participants are readily able to retrieve their initial guesses and to distinguish their guesses from the correct targets (e.g., Knight et al., 2012; Vaughn & Rawson, 2012). A question that remains, however, is whether incorrect guesses might become interfering at a long delay. At a delay, we expect that participants will display weaker episodic discrimination and a relatively stronger memory trace for generated guesses, as compared with studied targets. The combination of these two factors could create a case where generated guesses proactively interfere with access to the correct targets. If there is indeed no benefit (or even a detriment) to making guesses at long delays, this finding would have implications for applications of generating errors in education. In Experiment 3, therefore, we investigate whether making erroneous guesses starts to interfere after a longer retention interval (48 h).

Experiment 1

In Experiment 1, we replicated Kornell et al.'s (2009) Experiment 4, but with two changes. First, as was mentioned above, to nullify participants being able to use relative associative strength as a discriminative cue at the time of the final test, we made half of the to-be-learned responses strong, rather than weak, associates of the cue words. If the advantage of guessing before study is due to the use of a "the-answer-is-always-weakly-associated" heuristic at the final test and mixing high associates with the low associates prevents the usage of this strategy, the benefit of guessing-first over only studying on the final test should be eliminated. Second, in addition to asking participants to recall the correct targets on the final cued-recall test, we also asked them to recall their initial guesses. We reasoned that if guesses competed and interfered with the ability to retrieve the targets, we should see better recall of targets when participants are unable to recall their incorrect guesses.

Method

Participants and design

Thirty-four undergraduates from the University of California, Los Angeles (UCLA) participated in Experiment 1. The participants received partial course credit as compensation. We manipulated study condition (*guess-first* vs. *study-only*) and word-pair association strength (*strong* vs. *weak*) within subjects. In the cued-recall test phase, participants were asked—in response to each cue word—to recall the correct target and then the target they had guessed during the study phase prior to seeing the correct response.

Materials and apparatus

Sixty paired associates were used. Half were weakly associated word pairs with forward association strength between 0.05 and 0.054 (e.g., *Olive: Branch*); half were strongly associated word pairs with a forward association strength between 0.3 and 0.4 (e.g., *Table: Chair*). The weak associates were a randomly selected set of 30 pairs, taken from the materials of Kornell et al. (2009). All the words were, at minimum, four letters long. Half of the word associates were randomly assigned to the guess-first condition, which comprised 15 strong associates and 15 weak associates, and the remaining 30 were assigned to the study-only condition. Assignment of these two sets of 30 word pairs was counterbalanced across participants. The order in which the four within-subjects conditions (strong vs. weak; guess-first vs. study-only) appeared was block randomized; the list was divided into 15 blocks of four trials, where each block consisted of one pair from each within-subjects condition (therefore controlling for serial position effects between the conditions). From the participants' point of view, however, they saw only one long list of 60 word pairs. Finally, the order of the word pairs was fully randomized.

The experiment was created using Collector (https://github.com/gikeymarcia/Collector), an open-source PHP-based program designed to run psychology experiments and conducted via an Internet browser. Participants came into the laboratory and were administered the study on 21.5-in. Apple iMac desktop computers. The web browser was opened full-screen, and instructions and word pairs were all presented in the center of the screen.

Procedure

The study was composed of two phases: a study phase and a final cued-recall test phase. For the study phase, participants were told that they would study pairs of related words. Sometimes they would see complete pairs, whereas other times the second word would be missing. When pairs were shown incomplete, participants were told that they should try to guess the upcoming to-be-learned response, after which they would be shown the correct answer. Participants were shown the 60 word pairs one at a time. In the guess-first condition, they were presented with a cue and a blank (e.g., *Olive: _____)* and were given 8 s to make a guess (e.g., they might guess "Martini"). Participants were instructed to always make a guess, rather than to leave the space blank. The full cue–target pair (e.g. *Olive: Branch*) was then shown for 5 s immediately after making their guess. In the study-only condition, participants were presented with the full cue–target pair twice consecutively, for 8 and 5 s, respectively.

After a 5-min retention interval, participants were then given a final cued-recall test on all 60 word pairs. During the final cued-recall test, participants were shown a given cue twice followed by a blank line each time. Participants were informed that every cue word would be presented twice consecutively and were instructed to fill in the first blank with the correct target. For example, if they were presented with the cue: "*Olive*:_____," they should type "*Branch*" (the correct target) the first time they see "*Olive:* _____" in the final test. They were instructed that for the second blank, they should type in their original guesses, if the pair was in the guess-first condition. In the example given then, that means that they should type in "*Martini*" for the immediately subsequent, second presentation of "*Olive:* _____." If they had not been asked to make a guess for the cue word in the study phase (i.e., the pair had been in the study-only condition), participants were told to type in "*Read*," instead of an initial guess. It was not indicated during the final test whether the pair was in the guess-first or the study-only condition, and the second blank appeared regardless of whether participants were able to fill in the first blank (i.e., recall the correct target). Participants were not given any explicit instruction about whether they should always fill in the blank, and many left the space blank if they could not recall the answers. The pairs were presented in a randomized order, and the test was self-paced.

Results and discussion

Although comparison of the weak and strong associates is not of primary concern—the strong associates were included simply to reduce the possible use of a "the answer-is-always-weakly-associated" heuristic—we analyzed the strong and weak pairs separately. Successful guess rates were 4 % for the weak associates and 9 % for the strong associates for the guess-first pairs during the study phase. The rates for the weak associates were about as expected, but the rates for the strong associates were lower than expected. This lower rate may reflect that the pairs were intermixed, meaning that participants could learn that the most obvious associates were only infrequently the correct responses, leading to a reduced success rate for the strong associates. All analyses reported in this article are restricted to the items where the guess was incorrect. Additionally, responses were counted as correct only if they were typed into the appropriate spaces; in other words, recalled targets were counted as correct if entered into the first blank, but not if entered into the second.

Recall of correct targets

As is shown in Fig. 1, we replicated the basic finding—namely that the guess-first condition produced better later recall of the target response than did the study-only condition, despite the presence of strongly associated to-be-learned pairs.

Furthermore, the benefit of making incorrect guesses was present for both strongly associated to-be-learned pairs and weakly associated pairs. A two-way (study condition × association strength) within-subjects ANOVA showed that there

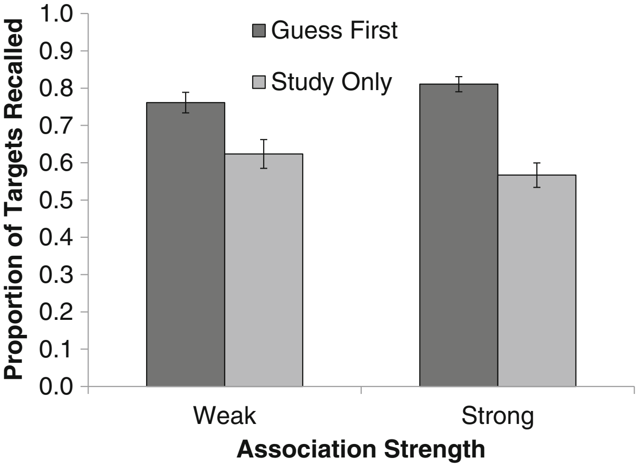

Fig. 1 Recall of the targets by study condition and association strength of the pairs in Experiment 1. Error bars represent standard errors of the means

was a main effect of study condition, $F(1, 33) = 45.06$, $MSE = .03$, $p < .1$, $\eta^2_p = .58$: Pairs in the guess-first condition ($M = .79$, $SD = .11$) were recalled significantly better than the pairs in the study-only condition ($M = .59$, $SD = .18$), $t(33) = 6.73$, $p < .001$, Cohen's $d = 1.15$. There was no significant effect of association strength, $F(1, 33) =0\ .04$, $MSE = .01$, $p > .05$, $\eta^2_p = .001$.

The study condition × association strength interaction was marginally significant, $F(1, 33) = 4.04$, $MSE = .02$, $p = .053$, $\eta^2_p = .11$. The benefit of making an incorrect guess appears to have been marginally larger for the strong associates (.81 vs. .57 for guess-first and study-only conditions, respectively) than for the weak associates (.76 vs. .62), although the benefit was significant for both weak and strong associates.

Whatever the reason for the strongly associated pairs showing at least as large a benefit of error generation, the key point is that the benefits of the guess-first condition found by Kornell et al. (2009) and subsequent research findings appear not to be a consequence of participants being able to adopt a heuristic at the time of the final test—namely, that the correct response is the weaker of the two remembered associates to a given cue word.

Participants' ability to recall their initial guesses

Participants ability to recall their initial guesses ($M = .79$, $SD = .13$) and the correct answers in the guess-first condition ($M = .79$, $SD = .11$) did not differ, $t(33) = 0.15$, $p > .05$, Cohen's $d = 0.026$; neither was there a difference in their recall of their guesses to the strong-associate cues ($M = .81$, $SD = .14$) and to the weak-associate cues ($M = .77$, $SD = .17$), $t(33) = 1.30$, p > .05, Cohen's $d = 0.22$. Intrusion rates of guesses into the blank space provided for targets and vice versa were very low: Guesses intruded into recall of targets only 1.4 % of the time ($SD = 3.1$ %), and targets intruded into recall of guesses only 2.9 % of the time ($SD = 4.8$ %). Thus, there was no evidence that initial guesses were suppressed, replicating the findings of Vaughn and Rawson (2012) and Knight et al. (2012).

For the study-only trials, participants correctly typed "*Read*" in the second blank provided for each given cue 78.2 % ($SD = 30.4$ %) of the time. The large standard deviation simply represents the 6 participants who may have misunderstood the instructions (3 of whom mostly left the space blank, and 3 of whom either provided the target a second tim or entered in completely new cue-related words).

Target recall conditional on guess recall

When we examine target recall, conditional upon ability to recall initial guesses, we see interesting patterns. A 2 (strength: weak, strong) vs. 3 (study condition: guess recalled, guess unrecalled, study only) within-subjects ANOVA revealed a main effect of study condition, $F(2, 58) = 31.6$, $MSE = .04$, $p < .001$, $\eta^2_p = .52$, but no main effect of strength and no interaction, Fs < 1. Data from 4 participants were not included in this ANOVA analysis because they had perfectly recalled their initial guesses to either all the weak associate or strong associate pairs. In Fig. 2, we collapse across strength and compare correct target recall performance of all of the study-only items with that of the guess-first items for which guesses were also recalled and with the guess-first items for which the guesses were not recalled. As is shown in Fig. 2, there is a benefit of generating guesses, but only when those initially incorrect guesses are later recallable.

Post hoc t-tests showed that while there was a benefit of making incorrect guesses ($M = .85$, $SD = .11$) over pure study ($M = .59$, $SD = .19$) when guesses were retrieved, $t(33) = 8.38$, $p < .001$, Cohen's $d = 1.43$, there was no difference between recall of targets of guess-first items when participants could not recall their guesses ($M = .59$, $SD = .22$) and recall of targets of study-only items, $t(33) = 0.01$, $p > .05$, Cohen's $d < 0.01$. Additionally, there was a significant difference between recall of the guess-first targets when guesses were recalled and when they were not, $t(33) = 7.29$, $p < .001$, Cohen's $d = 1.25$. These analyses suggest that participants' accessibility to their guesses also allows for greater accessibility of the targets and replicate the patterns found in prior studies (Butler, Fazio, & Marsh, 2011; Knight et al., 2012; Vaughn & Rawson, 2012).

Experiment 2

In Experiment 1, despite our expectations, successful guess rates between high and low associate words pairs were not dramatically different. Prior research by Koriat, Fiedler, and Bjork (2006) also suggests that hindsight bias can make it difficult for participants to accurately judge the likelihood of generating a target given a cue, particularly when the cue–

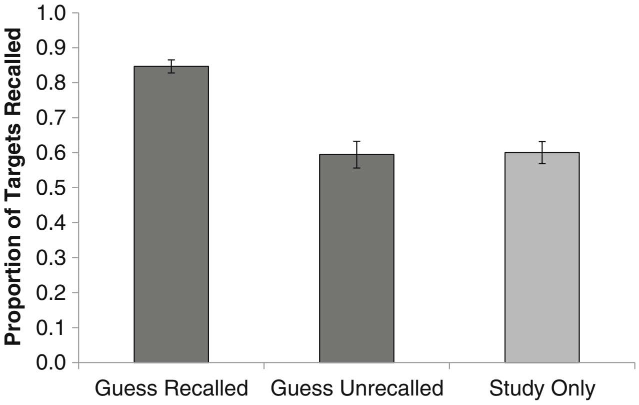

Fig. 2 Recall of targets, with guess-first trial broken down by whether initial guesses were recalled or not on the final test in Experiment 1. Error bars represent standard errors of the means

target pair is related: When shown a list of word pairs of zero, low, or high association, participants grossly overestimated the percentage of people who would generate the target given the cue and showed a remarkable underappreciation of the difference between high- and low-associate pairs. If, in hindsight, participants are unable to judge which word is a stronger associate to the cue word, then Experiment 1 might not have worked to fully address the heuristic that the correct target is always a weak associate.

In an attempt to address these concerns, we conducted an experiment similar to Koriat et al.'s (2006) study. We presented 33 participants with the word pairs and asked them, first, to judge the number of people out of 100 who would generate the target given the cue and then to categorize half of the pairs as "strong" and the other half as "weak" associates. As with Koriat et al., participants greatly overestimated the likelihood of generating the target given the cue for both strong (M = 60 %, SD = 12 %) and weak (M = 47 %, SD = 13 %) associate pairs, and they miscategorized 39 % (SD = 5 %) of the pairs as strong or weak. These findings suggest, therefore, that the subjective experience of participants in Experiment 1 did not differ as markedly as we expected for strongly and weakly associated pairs.

Another heuristic that would be easy for participants to use at the time of test in the original error generation paradigm is that almost every response they generate is incorrect. That is, if it is easy to distinguish between their generated response and the correct response at the time of test, then the one should not interfere with the other. We attempted to eliminate the use of this heuristic in Experiment 2 by rigging half of participants' responses to be correct. If the benefit of guessing first is a result of participants using a "my-guess-is-always-wrong" heuristic, then mixing correct guesses with the incorrect guesses should eliminate this strategy and eliminate the benefit or, at least, reduce the size of the benefit, as compared with when guesses are always incorrect.

Method

Participants and design

Fifty-nine participants were recruited from Amazon Mechanical Turk and were paid $1.50 for their participation. As in Experiment 1, study condition (read vs. guess-first) was manipulated within subjects. In Experiment 2, however, we also manipulated the presence of correct guesses (all-incorrect, n = 27, vs. half-incorrect n = 32) between subjects. For those in the half-incorrect condition, whatever guess the participant generated was deemed to be the "correct answer" for half of the guess-first word pairs.

Materials and procedure

The procedure of Experiment 2 was the same as that in Experiment 1, with three exceptions: First, instead of the combination of strong and weak associate pairs used in Experiment 1, we used the original 60 weak associate pairs used in Kornell et al. (2009). Second, for each individual, the word pairs were randomly assigned into one of three groups of 20 word pairs: study-only, guess-first, or filler. The two former conditions matched the study-only and guess-first conditions in Experiment 1. Filler words were presented in the same manner as guess-first words (i.e., participants spent 8 s generating a guess for the target word, given the cue, and 5 s studying the "correct" word that goes with the cue). The only difference was that in the half-incorrect condition, the "correct" word shown was whatever guess the participant had generated (i.e., Cue:Guess), while in the all-incorrect condition, the "correct" word shown was the weakly associated target from the Kornell et al. stimuli (i.e., Cue: Target). The study phase presentation order of the word pairs was block randomized into 10 blocks of six word pairs. Each block of six pairs consisted of two study-only pairs, two guess-first pairs, and two filler pairs. Following the study phase, participants were tested on all 60 presented pairs in random order. Finally, to reduce the complexity of instructions at the test phase, participants were tested only on their recall of the correct responses; that is, they were not asked to recall their original guesses or to identify whether the pair had been in the guess-first or study-only condition.

Results and discussion

Successful guess rate in the guess-incorrect condition was 6.5 %. Those pairs in which the guesses matched the intended target in the guess-incorrect condition were eliminated from the analyses.

If the benefit of guessing over study-only in the original paradigm was a result of using a guesses-are-always-incorrect heuristic, the benefit should be eliminated in the half-incorrect condition. A 2 (study-only vs. guess-first) × 2 (all-incorrect vs. half-incorrect) mixed ANOVA revealed, however, only a main effect of study condition, $F(1, 57) = 31.39$, $MSE = .02$, $p < .001$, $\eta_p^2 = .36$. In other words, there was a benefit of making incorrect guesses ($M = .63$, $SD = .24$) over study-only ($M = .50$, $SD = .25$). There was no main effect of the presence of correct guesses, $F(1, 57) = 1.63$, $MSE = .10$, $p > .05$, $\eta_p^2 = .03$, although performance in the half-incorrect condition ($M = .60$, $SD = .22$) was numerically higher than that of the all-incorrect condition ($M = .53$, $SD = .23$). Critically, however, there was no interaction between study condition and the presence or absence of correct guesses, $F(1, 57) < 1$. In other words, guess-first was significantly better than the read condition when all guesses were incorrect (M = .15, SD = .19) and when half of the guesses were rigged to be correct (M =.12, SD = .16), and the magnitude of the benefit did not change depending on the presence or absence of correct guesses.

Within the all-incorrect condition, performance on the 20 "filler" pairs (M = .61, SD = .23) was, as expected, not significantly different from performance on the guess-first pairs, $t(26) = 0.61$, $p > .05$, Cohen's $d = 0.12$. Within the half-incorrect condition, guesses rigged to be correct ($M = .80$, $SD = .15$) was significantly higher than recall of targets in the guess-incorrect and study-only conditions, $p < .01$. This benefit of correct guesses is to be expected on the basis of what we know about the generation effect (Slamecka & Graf, 1978). Finally, although we did not ask participants in Experiment 2 to distinguish between their correct and incorrect guesses, or the guess-correct trials from the guess-incorrect trials, it is interesting to note that on the guess-incorrect trials, initial guesses intruded in on the recall of correct answers only 6 % of the time in the all-incorrect condition and 8 % in the half-incorrect condition. Coupled with the fact that participants were able to correctly recall correct guesses 80 % of the time, it appears that they are able to distinguish between their correct and incorrect guesses, ruling out the "my-guess-is-always-wrong" heuristic.

Experiment 3

Experiments 1 and 2 eliminated two potential heuristics that might help explain why making an incorrect guess can enhance later recall. We also replicated the finding that, at a short delay, participants not only are able to recall their original guesses very well, but also are able to discriminate between the incorrect guesses they had generated and the correct target. The conditional analyses in Experiment 1 also suggested that participants' ability to recall their original guesses—rather than interfering—was related to their ability to recall the correct answer. In other words, source memory is very accurate with only a 5-min delay between the study and test phases. Source memory after a longer retention interval, however, may not be as accurate, and an inability to distinguish between generated responses and correct responses may lead to an overall benefit of study-only over guess-first.

Method

Participants

Twenty-nine undergraduates from UCLA participated in Experiment 3 for course credit.

Design, materials, and procedure

Experiment 3 was the same as Experiment 1, with the exception of two changes: First, instead of a 5-min delay, there was a 48-h interval between the study phase and the test phase. Second, the study was conducted entirely online, instead of in the laboratory. Participants were first given a link to complete the study phase, which was identical to the study phase in Experiment 1. Approximately 48 h later, participants were asked, via email, to finish the test phase online, recalling both the correct targets and their initial guesses, as in Experiment 1. Of the participants, 100 % completed the test phase. On average, participants took the delayed test 61 h after initial study, and there was no significant correlation between the time of delay and final recall performance or initial responses recall performance.

Results and discussion

Overall, correct guess rates were 4 % for the weak associates and 8 % for the strong associates for the guess-first pairs during the study phase. These figures are comparable with those found in Experiments 1 and 2. Again, all analyses are restricted to those items for which initial guesses were incorrect.

Recall of correct targets

As was expected, overall recall performance was lower after this longer delay than it was after the 5-min delay in Experiment 1, but the pattern, as shown in Fig. 3, is otherwise remarkably similar to the findings in Experiment 1. A two-way (study condition × association strength) within-subjects ANOVA revealed a main effect of study condition, $F(1, 28) = 12.22$, $MSE = .02$, $p < .01$, $\eta^2_p = .30$, no main effect of association strength, $F(1, 28) = 2.04$, $MSE = .01$, $p > .05$, $\eta^2_p = .07$, and no interaction, $F(1, 28) = 0.29$, $MSE = .01$, $p > .05$, $\eta^2_p = .01$. On average, targets in the guess-first condition were recalled 32 % (SD = 16 %) of the time, while targets in the study-only condition were recalled 23 % (SD = 14 %) of the time, a difference which was significant, $t(28) = 3.46$, $p < .05$, Cohen's $d = 0.64$.

Participants' recall of their initial guesses

Guesses were recalled at a significantly higher rate ($M = .44$, $SD = .19$) than the targets ($M = .32$, $SD = .16$), $t(28) = 3.72$, $p < .05$, $g = .69$. Participants were less able also to correctly type "*Read*" into the second prompt for those items that had been in the study-only condition ($M = .32$, $SD = .16$), with many of the spaces left blank ($M = .34$, $SD = .28$) or with new cue-related words entered ($M = .25$, $SD = .32$). It is unclear, however, whether these responses for the study-only items reflect

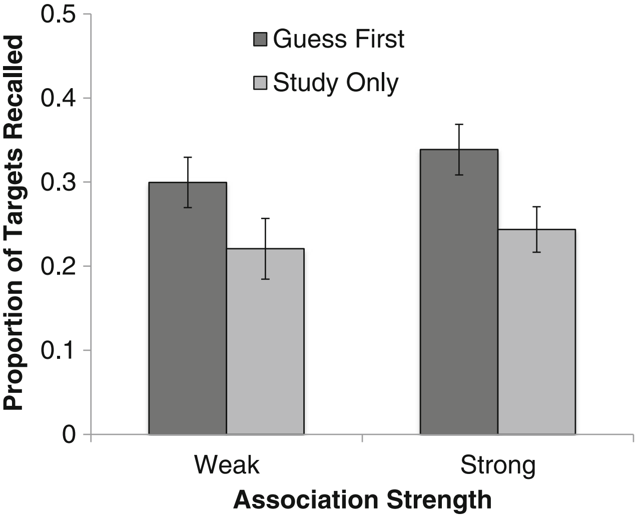

Fig. 3 Proportion of targets correctly recalled, by study condition and association strength, in Experiment 3. Error bars represent standard errors of the means

blurred source memory (i.e., participants could not remember whether they had initially generated a guess for those words), confusion with respect to the instructions, or, perhaps, a combination.

As one would expect with an increased retention interval and, hence, decreased episodic discrimination, intrusions rates for the guess-first items were also increased, as compared with those of Experiment 1: Initial guesses intruded into recall of the targets 12.4 % (*SD* = 14.7 %) of the time, and targets intruded into the recall of the initial guesses 7.5 % (*SD* = 12.7 %) of the time. Finally, there was also a marginally significant difference in the recall of guesses for the strong-associate cue–target pairs (*M* = .39, *SD* = .21), as compared with the recall of guesses for the weak-associate cue–target pairs (*M* = .48, *SD* = .23), *t*(28) = 2.01, *p* = .054, Cohen's *d* = 0.37. This pattern of results may be a result of greater interference from strong-associate targets, or because the normative association strength of the guesses to the cues was higher for the weak-associate pairs than for the strong-associate pairs. In support of this speculation, the pattern was reversed for the recall of targets: Target recall was higher for the strong-associates (*M* = .34, *SD* = .16) than for the weak-associates (*M* = .30, *SD* = .19). Although the difference in target recall was not significant between the strong and weak associates, *t*(28) = 1.28, *p* > .05, Cohen's *d* = 0.24, there was a significant association strength × response (target vs. guess) interaction, *F*(1, 28) = 11.27, *MSE* = .012, *p* < .01, η^2_p = .29.

In sum, with a longer retention interval, guesses—having been generated—were better recalled than targets. One possible outcome of guesses being stronger is that they then, with a delay, become more interfering. Yet, despite this greater potential for interference, we still found a significant benefit of making guesses.

Target recall conditional upon recall of the guesses

As in Experiment 1, we examined the likelihood of target recall conditional upon guess recall, the results of which are represented in Fig. 4. A one-way within-subjects ANOVA revealed that there were significant differences between the recall of the targets of guess-first items when guesses were retrieved and when they were not retrieved and the study-only items, *F*(2, 56) = 4.05, *MSE* =.03, *p* < .05, η^2_p = .13. Post hoc *t*-tests showed that while there was a benefit of making incorrect guesses over pure study (*M* = .23, *SD* = .14) when guesses were retrieved (*M* = .36, *SD* = .24) *t*(28) = 3.46, *p* < .01, Cohen's *d* = 0.64, there was no difference between recall of targets of *guess-first* items when guesses could not be retrieved (*M* = .27, *SD* = .19) and the targets of the study-only items, *t*(28) = 1.01, *p* > .05, Cohen's *d* = 0.19. That is, even after a 48-h delay, ability to recall guesses is positively associated with recall of the correct targets—one possible interpretation is through a "mediator" lens: that recalling the guess enhanced recall of the target; failure to retrieve one's initial guess led to no benefit over pure study. Unlike in Experiment 1, however, there was not a significant difference between recall of the guess-first targets, when guesses were recalled (*M* = .36, *SD* = .24) and when guesses were not recalled (*M* = .27, *SD* = .19), *t*(28) = 1.54, *p* > .05, Cohen's *d* = 0.29.

In sum, Experiment 3 showed that even after a long delay (of at least 48 h), participants' incorrect guesses did not interfere with the recall of the correct targets.

General discussion

In Experiments 1 and 2, we ruled out two factors that might contribute in an artificial way to the observed benefits of

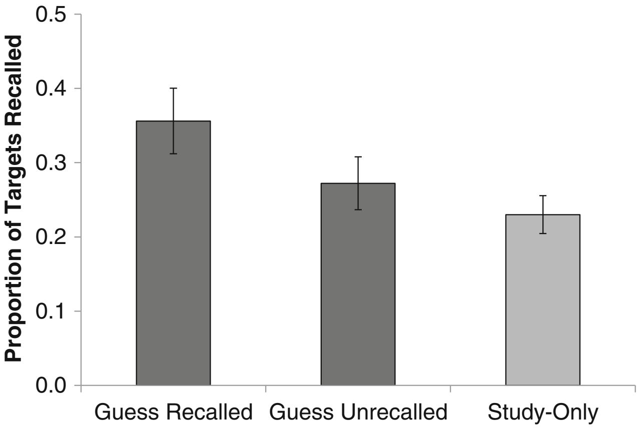

Fig. 4 Recall of targets, with guess-first trial broken down by whether initial guesses were recalled or not on the final test in Experiment 3. Error bars represent standard errors of the means

making an incorrect guess, versus studying the correct pair. When participants could not, at the time of the final test, rely on "pick the weaker associate" or "pick the one that I did not generate," we still found benefits of making incorrect guesses over a study-only condition. Finally, Experiment 3 showed, remarkably, that errors did not interfere with the ability to recall the correct answers but (counterintuitively) were related to greater recall of the correct answer.

Why guessing incorrectly might enhance later recall: possible mechanisms

How do conditions that should, logically, create proactive interference and also reduce the participants' time to study the target response actually lead to better recall of a to-be-learned response? Several possible mechanisms have been proposed.

Suppression

One mechanism that might explain, at least partially, the benefits of making incorrect guesses is that when corrective feedback is provided, the incorrect guesses become inhibited or suppressed and, therefore, do not interfere. However, consistent with prior research (Knight et al., 2012; Vaughn & Rawson, 2012), the results of Experiments 1 and 3 should that participants are readily able to retrieve their initial guesses, suggesting that the guesses were not suppressed.

We might have predicted that making incorrect guesses—while beneficial for short-term learning—would proactively interfere with recall of correct responses at a longer delay, where episodic discrimination between initial guesses and correct responses should be degraded. The results of Experiment 3, however, show that even after an average of 61 h, we still find a benefit of making incorrect guesses.

Mediation

Another mechanism that has been proposed is that making incorrect guesses can, in fact, function as an additional cue, aiding the recall of the correct target. Pyc and Rawson (2010) demonstrated that when participants are instructed to generate mediators, mediator effectiveness (as measured by ability to both retrieve and decode mediators during the criterion test) is enhanced through testing. Findings by Carpenter (2011) suggest that explicit instructions to use mediators may not be necessary. Rather, semantic mediators may be covertly generated during initial study and, more so, when initial study involves testing rather than purely studying. Carpenter found that never-presented strong associates to cue words in a *study-test* condition were more likely to be falsely recognized on a later recognition test than never-presented strong associates to cue words in a *study-restudy* condition. Furthermore, when cued with these strong associates, participants were more likely to recall the correct targets in the *study–test* condition than in the *study–restudy* condition. As they apply to Kornell et al.'s (2009) paradigm, these mediation ideas suggest that the cue for a given pair, the erroneous response that is generated, and the target response are integrated into a kind of triplet that then aids recall of the target response at the time of the final test.

Three prior studies—Butler et al. (2011), Knight et al. (2012), and Vaughn and Rawson (2012)—have demonstrated that target responses are better recalled when the initial guesses are also recalled. In the latter two studies, the same cue–target paradigm was used (Butler et al. used general knowledge questions), and participants were asked to recall their initial guesses first before recalling the targets. In our present study, we reversed this order, asking participants to provide the targets first before recalling their initial guesses. Despite this reversal of output order, our results are similar to their findings: Experiments 1 and 3 found that when participants were able to recall their initial guesses, they were more likely to also recall the correct targets.

It is not clear, however, whether these studies constitute evidence for the mediator hypothesis. While this pattern of results would be consistent with the mediator hypothesis, it does not necessitate it; this pattern could be the result of item selection effects. An alternative account of the results is that those trials for which guesses are recalled have simply been encoded more deeply and, therefore, both guessed and actual targets are more easily recalled. This account would, therefore, not posit that retrieval of the guesses must precede the targets (as would be predicted in the strict sense of 'mediation') but, rather, allow for the two to be simply correlated.

Semantic activation

Finally, another mechanism proposed by Kornell et al. (2009) has gained considerable support—namely, that trying to predict an upcoming to-be-learned response requires activating the semantic network associated with the cue. The basic idea is that this activation then affords a richer encoding of the subsequently presented target. That is, the to-be-learned response is then encoded in a richer, more elaborated way, in relation to the cue, than would have been had the intact pair been shown for study only.

In support of the semantic activation hypothesis, researchers have found that the benefit of making incorrect guesses is eliminated in cases where semantic activation is misguided (e.g., in the case of unrelated word pairs; Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012), when feedback in the guess-first condition is delayed (Grimaldi & Karpicke, 2012; Hays et al. 2013; Vaughn & Rawson, 2012; but see Kornell, 2014), and when the activation is constrained during guess generation (Grimaldi & Karpicke, 2012).

Many other converging studies—using different types of learning materials, both in the laboratory and in the classroom—also show benefits of incorrect guesses, given that the process of generating errors activates semantic networks. In the lab, McGillivray and Castel (2010) demonstrated that in learning face–age associations, both older and younger adults benefit from making a guess as to a face's age before being given the answer, even though these guesses were almost always incorrect. Importantly, McGillivray and Castel found that guessing benefited learning of face–age associations only when there was schematic support (i.e., when the to-be-learned ages made sense given the cues from the face).

For Singapore math classrooms, Kapur and Bielaczyc (2012) demonstrated the benefit of what they called "productive failure." In their study, half of the classes spent six class periods trying and failing to solve math problems before receiving one period of instruction being given the correct answer ("productive failure" condition). Critically, in this one period of instruction, teachers not only explained what the correct answer was, but also compared and contrasted the correct solution to the incorrect solutions. The other half of the classes spent all seven periods being taught the correct method, practicing questions, doing homework and getting feedback ("directed instruction" condition). On a final test, those in the productive failure condition performed better than those in the directed instruction condition, particularly on complex problems and a test of representational flexibility.

The semantic activation hypothesis cannot be the whole story, however, since Potts and Shanks (2014) recently showed benefits of anticipating upcoming to-be-learned responses in a series of experiments where (relevant) semantic activation is impossible: Participants had to guess and learn the definitions of rare or obscure English words and Euskara words; words for which they had no way of activating relevant semantic concepts. Despite the lack of semantic relationship between participants' guesses, the cues, and the targets, Potts and Shanks found a robust benefit of generating guesses first over simply studying the cue–target pairs.

Concluding comments

Our results, together with those obtained by other researchers, show that activating knowledge before study and testing, even when responses are incorrect, can benefit learning. Understanding fully the dynamics that offset what would seem major costs of generating incorrect guesses, including introducing proactive interference and reducing study time, awaits further research, but one implication is that instructors need not consider difficult tests as inherently "risky" to administer.

Author note This research was supported by Grant No. 29192G from the McDonnell Foundation. We thank the members of Minifog and CogFog for their contributions to this article. Portions of this research were presented at the 53rd annual meeting of the Psychonomic Society, Minneapolis, MN.

References

Briggs, G. E. (1954). Acquisition, extinction, and recovery functions in retroactive inhibition. *Journal of Experimental Psychology, 47,* 285–293.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review, 18,* 1238–1244.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20,* 941–956.

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.

Cunningham, D. J., & Anderson, R. C. (1968). Effect of practice time within prompting and confirmation presentation procedures on paired associate learning. *Journal of Verbal Learning and Verbal Behavior, 7,* 613.

Elley, W. B. (1966-66). The role of errors in learning with feedback. *British Journal of Educational Psychology, 1966-66, 35–36*, 296–300.

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

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 290–296. doi:10.1037/a0028468

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40,* 514–527.

Kaess, W., & Zeaman, D. (1960). Positive and negative knowledge of results on a pressey-type punchboard. *Journal of Experimental Psychology, 1,* 12.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21,* 45–83.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., 3rd. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory, 17,* 471–479.

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.

Koriat, A., Fiedler, K., & Bjork, R. A. (2006). Inflation of conditional prediction. *Journal of Experimental Psychology: General, 135,* 429–447.

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., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin and Review, 14,* 219–224.

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.

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14,* 194–199. doi:10.3758/bf03194051

McGillivray, S., & Castel, A. D. (2010). Memory for age-face associations: The role of generation and schematic support. *Psychology and Aging, 25,* 822–832.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143,* 644–667. doi:10.1037/a0033194

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 333,* 335.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15,* 243–257.

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

Skinner, B. F. (1958). Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction. *Science, 128,* 969–977. doi:10.1126/science.128.3330.969

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior, 22,* 153–163.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology, 4,* 592–604.

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

Vaughn, K., & Rawson, K. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19,* 1–7. doi:10.3758/s13423-012-0276-0