Does Incorrect Guessing Impair Fact Learning?

Sean H. K. Kang and Harold Pashler
University of California, San Diego

Nicholas J. Cepeda
York University

Doug Rohrer
University of South Florida

Shana K. Carpenter
Iowa State University

Michael C. Mozer
University of Colorado, Boulder

Taking a test has been shown to produce enhanced retention of the retrieved information. On tests, however, students often encounter questions the answers for which they are unsure. Should they guess anyway, even if they are likely to answer incorrectly? Or are errors engrained, impairing subsequent learning of the correct answer? We sought to answer this question in 3 experiments. In Experiments 1 and 2, subjects read 80 obscure facts (e.g., "Where is Disko Island? Greenland") and then took a cued recall test. When a subject reported being unable to answer a question, on a randomly chosen half of those questions the computer program insisted upon a guess. Corrective feedback was provided either immediately (Experiment 1) or after a delay (Experiment 2). Forced guessing did not affect subjects' performance on a final test given 1 day later. We extended the investigation to more complex material in Experiment 3. Subjects saw a question (e.g., "Why do ice cubes often pop as they melt in your drink?") and its answer, but for half of the questions, subjects did not see the answer until they first provided a plausible explanation. On a test administered either on the same day or 1 week later, recall performance was again unaffected by a prior wrong guess.

Keywords: error correction, hypercorrection, guessing, fact learning

When students take a test, they often encounter questions they cannot answer. This occurs if students did not prepare sufficiently (unfortunately, an all-too-common occurrence) or if the test is used to enhance rather than assess learning (e.g., flashcards). When a question cannot be answered, students typically guess, because

penalties for wrong responses are rarely implemented. While venturing a guess might enhance a student's score on the test, it seems reasonable to wonder whether it might have undesirable effects on the underlying learning when the guess is wrong. In other words, does producing an incorrect guess impair subsequent learning of the correct information?

This issue bears important theoretical and practical implications for learning. The recent advocacy for greater use of testing as a tool to directly enhance learning (e.g., McDaniel, Roediger, & McDermott, 2007) lends special importance to the question of whether wrongful guessing negatively impacts the learning of facts.

In the present study, subjects were tested twice on the to-be-learned material, with corrective feedback provided after the initial test. Our chief focus was on the students' performance on items that they could not answer on the initial test: Does requiring subjects to guess negatively affect accuracy on the final test relative to their not making a guess (i.e., withholding a response)? Our experiments had two critical features that allowed us to address this question more directly than has been the case in prior work: (a) We focused on items that subjects could not answer on the initial test, hence guaranteeing that when guesses were produced they were almost invariably wrong, and (b) these items were randomly assigned to the guess versus response-withheld conditions, thus allowing us to examine the causal impact of erroneous guessing. Although the question on the table has rich connections with a number of studies in different traditions (as we will describe

later), it will be seen that there has been a lack of evidence that could resolve the question posed.

Previous Studies on Guessing

To our knowledge, the earliest formal study on the effects of guessing was conducted by Forlano and Hoffman (1937), who compared the efficacy of two methods of learning foreign language vocabulary. Third- to eighth-grade children were taught the meaning of 40 Hebrew words. For half the words, subjects were cued with the Hebrew word and had to guess the English translation before being told the correct answer. For the remaining words, they were simply told the correct translation without having to first make a guess. Forlano and Hoffman found that on a cued recall test given immediately after the learning phase, performance was significantly worse in the guessing condition. (They repeated the same test 2 days later to assess delayed retention, and the same effect was observed; of course, performance on this second test was contaminated by the earlier test.)

It would be premature to conclude that erroneous guessing impedes subsequent learning on the basis of this study. In their article, Forlano and Hoffman (1937) omitted potentially important details of their classroom experiment—e.g., whether subjects had the same amount of time to process the correct answer across the guessing and no-guessing conditions—and so it is impossible to know whether their finding was the result of a procedure that might have unintentionally placed the guessing condition at a learning disadvantage.

Instead of wrong guesses interfering with the learning of foreign vocabulary, Parlow and Berlyne (1971) showed that the opposite can occur: In situations where the guessing was constrained (i.e., subjects were cued with the foreign word together with its appropriate category), guessing (even though rarely correct) enhanced performance; the authors hypothesized that the guess served as an effective mediator between the stimulus (the foreign word) and the correct response. Although the study is intriguing, it involved a constrained form of guessing that differs from the guessing that would take place in the typical testing situation.

Guessing has also been examined in cases in which guesses are elicited prior to any instruction (*prequestioning*). Berlyne (1954b, 1966) found that forced guesses to "prequestions" improved retention of the information. He argued that guessing enhanced curiosity about the topics and offered some evidence from subjects' self-reports that was consistent with this interpretation (cf. M. J. Kang et al., 2009). Berlyne (1954a) hypothesized that curiosity is a drive that is satiated by correct knowledge, which causes the correct answer to be better learned than it would be if curiosity were not aroused. More recently, similar positive effects of prequestioning were also reported by Kornell, Hays, and Bjork (2009) and by Richland, Kornell, and Kao (2009).

While the findings of these studies on prequestions or pretests seem to suggest that guessing may not harm (and might even enhance) learning, what was compared in the studies was the effect of being asked a (hard-to-answer) question versus not being asked the question at all. Such comparisons do not directly address the question of whether wrongful guessing harms subsequent learning, and extant data cannot provide a clear answer due to factors like the conflation of commission and omission errors (i.e., producing a response that is wrong as opposed to leaving an answer blank) on

the pretests and item-selection artifacts when the analysis is conditionalized on a subset of items. By contrast, our focus in the current article was the causal impact of requiring a learner who has been asked a difficult question to guess the answer when he or she does not know it.

Theoretical Perspectives on the Effects of Errors

The learning theorist Edwin Guthrie (1942) is famous for asserting that people learn only by doing: "A student does not learn what was in a lecture or in a book. He learns only what the lecture or book caused him to do" (p. 55). Consequently, according to Guthrie, making an error stamps in undesirable stimulus–response associations. Guthrie assumed that this would happen even when the learner recognized that he or she had made an error. The idea that producing errors has undesirable consequences was also advocated by researchers in the Skinnerian tradition (e.g., Taber, Glaser & Schaefer, 1965). Skinner (1958) nicely summed up the behaviorists' advice in an article on teaching machines: "It is a salutary thing to try to guarantee a right response at every step in the presentation of a subject matter" (p. 975).

A teaching approach designed to minimize errors has been formalized in a training procedure referred to as *errorless learning*. This technique, which originated in the animal learning domain (Terrace, 1963), has been applied with apparent benefits in the rehabilitation of individuals with memory impairments. A typical errorless learning study involves some form of paired-associate learning. For subjects in the control learning condition, learners might be presented with a cue and asked to guess the identity of the target (which usual results in errors) before the correct target is revealed. In the errorless learning condition, learners do not have to guess (thus avoiding errors); they are simply presented with the cue and the correct target. On a subsequent test, performance is sometimes found to be enhanced by training in the errorless condition (e.g., Baddeley & Wilson, 1994; Squires, Hunkin, & Parkin, 1997).

From the errorless learning literature, one might hypothesize that erroneous guesses on a quiz should lead to poorer learning of the correct answer. It should be noted, however, that the learning that occurs in studies of errorless learning is often thought to be primarily dependent upon implicit memory processes (e.g., Anderson & Craik, 2006; Evans et al., 2000) rather than explicit declarative memory. It is therefore unclear whether the benefit of errorless learning generalizes to the normal population and to tasks in which respondents rely chiefly upon explicit memory (e.g., fact learning).

Contemporary computational models of learning and memory do not offer strong predictions concerning the effect of erroneous guessing on learning. From a modeling perspective, the training paradigm used in experimental studies is best characterized as *supervised learning*, where the teacher provides the correct response to the learner, as opposed to reinforcement learning, where the teacher provides only evaluative feedback, or unsupervised learning, where no feedback whatsoever is provided. The effect that commission errors have on learning depends not so much on the computational model per se but on assumptions concerning how self-generated responses influence learning. From a Guthrian perspective, one might suppose that responses produced by a learner serve as a self-generated supervision signal for training in

much the same manner as do target responses provided by the teacher, resulting in strengthening of both the incorrect and correct responses on a commission trial. According to such a hypothesis, errors of commission should result in poorer memory of the correct answer than errors of omission.

However, errors of commission have value according to some models of learning. In an error-correction framework (e.g., Rescorla & Wagner, 1972; Rumelhart, Hinton, & Williams, 1986), the error signal used for learning is based on the difference between the actual response produced by the model and the target response provided by the teacher. Without the availability of the actual response, error-correction learning is reduced to the less powerful Hebbian associative strengthening. In a theory of learning, it reasonably might be supposed that unless a learner is forced to guess a response explicitly, the error signal could not be computed. Mozer, Howe, and Pashler (2004) used this assumption in their model to account for the benefit of testing over restudying on later retention (which will be elaborated on in the next section). According to this assumption, being forced to guess, even when incorrect, will facilitate learning.

To summarize, modern models of associative learning could predict (a) a cost of commission errors relative to omission errors, if the error response serves as a self-generated teaching signal; (b) a benefit of commission errors, if the process of generating the response serves to obtain a more meaningful corrective training signal; or (c) no effect of commission errors, if the cost and benefit roughly cancel each other out. Alternatively, no effect is predicted if learners discount their low-confidence guesses and thereby prevent either Guthrian strengthening or the use of the response to generate a corrective training signal.

Practical Concerns About Errors and Testing

In addition to the potential for increasing theoretical understanding of the role of errors in learning, our other goal in the present study was to determine what advice should be provided to students and teachers regarding the effects of guessing in practical contexts. A wealth of evidence has demonstrated that taking a test typically produces better retention of the material than restudying or rereading (for a review, see Roediger & Karpicke, 2006). Research on the testing effect has led to strong advocacy for greater reliance upon quizzing to aid student learning (Roediger, Agarwal, Kang, & Marsh, 2009). Moreover, it has been advocated that, in order to maximize learning, a test should require effortful retrieval (e.g., short answer tests tend to be more effective than multiple-choice tests) and be followed by corrective feedback (e.g., S. H. K. Kang, McDermott, & Roediger, 2007).

One likely consequence of using open-ended test questions to enhance learning is, of course, an increase in errors. If a student is unsure of the answer, should he or she guess or simply refrain from answering? A direct investigation of this issue has become timely, given the current advocacy of testing as a tool to enhance learning (e.g., McDaniel et al., 2007; Roediger & Karpicke, 2006). In addition to test-based learning strategies, deciding whether to guess or not commonly arises in daily life. For example, when one is not sure how to spell a word or cannot think of someone's name, one can either guess or withhold a response until reliable information is available. One factor that can be weighed into such

decisions is whether the act of guessing will interfere with later learning of the correct information.

Remmers and Remmers (1926) first raised the possibility that the lures on true/false tests—generally constructed to seem plausibly true—might lead to the learning of false information, referring to such an outcome as the *negative suggestion effect*. Indeed, there is evidence demonstrating that (false) statements presented as lures on a test are later judged as more true, relative to novel statements (Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). More recently, Roediger and Marsh (2005) showed that lures on a multiple-choice test sometimes do intrude on a subsequent cued recall test—especially lures that were erroneously endorsed as correct on the earlier test. It would seem plausible that the same thing could occur for erroneous information that was self-generated by the test-taker (i.e., wrong guesses on a recall test). If this were indeed the case, students would need to be cautioned against being too liberal with guessing in their test-taking behavior.

Summary

The review of the pertinent prior research reveals disparate findings, precluding one from drawing strong conclusions about the effects of erroneous guessing on learning. On the one hand, there is some evidence that wrong guesses interfere with subsequent learning, in line with Skinnerian principles postulating that it is easier to learn a desired response by encouraging responses that are close approximations of it (*shaping*) than by eliciting incorrect responses followed by correction. On the other hand, there is evidence pointing to the benefit of prequestions for learning, even though prequestions are likely to elicit wrong guesses.

In evaluating the prior literature, one should keep in mind that most of these studies were not designed specifically for examination of wrong guesses and their possible impact on subsequent learning; thus, one is left with mostly post hoc analyses from which to derive clues. In order to determine whether incorrect guessing impairs explicit fact learning, additional investigation with designs that avoid some of the ambiguities of past studies is needed.

Present Experiments

Because test takers usually have some exposure to the material before taking the test, our first two experiments did not rely on the prequestioning or pretesting procedure that predominated in the previous research. Instead, subjects in Experiments 1 and 2 first read 80 obscure trivia facts and were then tested on these facts. When the subject indicated that "I have no idea," with a 50% probability, he or she was then required to produce an answer. Regardless of whether a forced guess was required, the subject was always shown the correct response to each item. Subjects returned the next day and were tested on all items without feedback.

This design allowed us to pose a number of unanswered questions about the effect of guessing. First, we could compare performance on (a) trials where the subject spontaneously answered and was wrong; (b) trials where the subject responded correctly; and (c) trials where the subject did not spontaneously answer. Interference theory would seem to suggest that guessing wrongly

(an *error of commission*) would be associated with worse performance on that item the following day, because the erroneous answer would need to be unlearned. Second, and perhaps most pertinent to the primary aim of this study, we could examine whether being forced to guess affected the degree of learning seen the following day.

In Experiment 3, we extended the investigation to the learning of complex phenomena, where the correct answer consisted of a short paragraph rather than a single word or phrase. Subjects learned the explanations for 12 scientific or historical phenomena that were generally unknown to college students. For half of the phenomena, subjects were asked to provide a plausible explanation before they were shown the correct answer (i.e., similar to the pretesting method used in previous research). For the remaining phenomena, subjects saw the correct explanations without having to first provide a guess. A final recall test was administered after a delay of 10 min or 1 week.

The Skinnerian account—according to which the production of an erroneous guess strengthens learning that will compete with correct learning triggered by the corrective feedback—obviously implies that forced guessing will be harmful, assuming that it often elicits an error (as the data will show).

Experiment 1

Method

Subjects.   Sixty-five undergraduates participated in a two-session experiment.

Stimuli.   A list of 80 not-well-known facts was assembled. For each fact, there was a corresponding question and correct answer. For example, one fact was "Richard Nixon set up the only hamburger stand in the South Pacific during World War II," and the corresponding question was "Who set up the only hamburger stand in the South Pacific during World War II?"

Design and procedure.   The procedure is illustrated in Figure 1. Subjects attended two sessions separated by 1 day. The first session was a training session consisting of a presentation of every fact followed by a test on every fact. The 80 facts appeared in a random order (4 s each), and subjects were instructed to study them for a later test. Prior to the test, subjects were told if they could not answer a question, the computer would sometimes insist that they respond. Subjects also were asked not to give a "joke answer" when they were forced to guess (which they sometimes did in our pilot studies). Each of the 80 facts was tested one at a time, in a random order different from the presentation order. For each test question, the subject was given the opportunity to click on the statement "I have no idea." If the subject did respond with an answer, he or she was required to indicate his or her confidence in the accuracy of the response by clicking one of five confidence values ranging from *very low* to *very high*. When the subject clicked on "I have no idea," the computer randomly determined (with probability of .5) whether the item would be assigned to the forced-guess condition (in which case the subject was required to make a guess), or the skip condition (in which case the subject was not required to guess). In every case, the computer then provided both the question and the correct answer for 1 s.
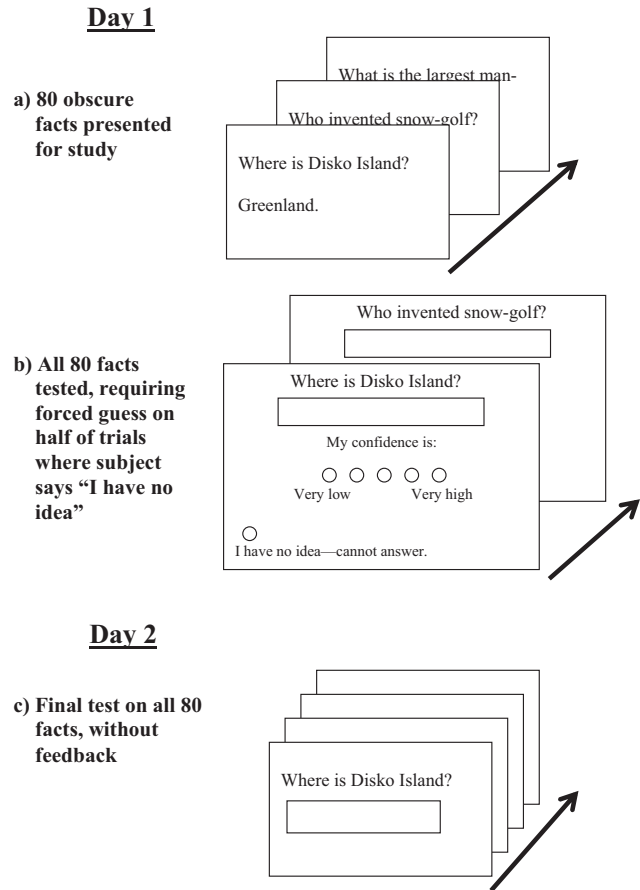

Figure 1.   Basic procedure on Day 1 and Day 2 of Experiments 1 and 2.

On the second day, subjects received a final test on all the items (order randomized). Subjects were required to provide an answer to every question, and no feedback was given.

Results and Discussion

Each of 65 subjects was tested on 80 items, except for one subject who was tested on only 79 items due to computer malfunction, resulting in a total of 5,199 items tested. The alpha level for all analyses was set at .05.

Initial test (Day 1) performance.   On Day 1, subjects voluntarily answered approximately 69% of all trials (3,590 of 5,199). On the remaining 31% of trials where subjects reported having "no idea" of the answer, the computer required subjects to make a guess on 804 trials (15.5%) and did not require a guess on 805 trials (15.5%).

Of the 3,590 responses produced voluntarily, 2,151 (59.9%) were correct. On the other hand, forced guesses were very rarely correct. Of 805 forced guesses, only 11 (1.4%) were correct. In short, when subjects reported that they had "no idea," they were generally correct in this assessment.

Response confidence and initial test (Day 1) performance. When subjects. responded voluntarily, they indicated their confidence, and the distribution of these ratings and their associated mean accuracy are shown in Table 1. A link between confidence

Table 1

Distribution of Confidence Ratings and Associated Mean Accuracy for the Initial Test on Day 1 (Voluntary Answering) in Experiments 1 and 2

Experiment/measure	Confidence rating				
	1 (*very low*)	2	3	4	5 (*very high*)
Experiment 1					
Confidence rating (%)	19.7	7.8	10.4	9.2	52.9
Correct responses (%)	8.4	26.2	42.0	59.4	86.6
Experiment 2					
Confidence rating (%)	19.2	7.6	9.0	12.6	51.5
Correct responses (%)	8.6	21.2	47.9	67.0	82.5

and accuracy is apparent: Responses made with higher confidence tended to be more accurate than those made with lower confidence. The mean within-subject Goodman–Kruskal gamma correlation between response confidence and accuracy (computed across items and then averaged over subjects) was .84 ($p < .001$), confirming a strong positive relationship.

Final test (Day 2) performance.   On the final test, mean accuracy equaled 53.1%. Table 2 shows final test performance as a function of the subject's response on the initial test. For initial test questions answered voluntarily and correctly, final test accuracy averaged 90.7%. For items answered voluntarily but incorrectly, final test accuracy averaged 35.8%. For items eliciting "I have no idea" response on the initial test (regardless of whether subjects were forced to guess), final test accuracy averaged 21.3%. The first of these three means was significantly higher than both the second, $t(64) = 31.54$, $d = 3.91$, and the third, $t(63) = 44.20$, $d = 5.52$.

With regard to the second and third means, final test performance was better for items that elicited spontaneous errors on the initial test than for items that elicited "I have no idea" on the initial test, $t(63) = 7.40$, $d = 0.92$. The size of this effect is moderately large according to the criteria of Cohen (1988). Even when the analysis was restricted to spontaneous errors that were made with the lowest confidence, volunteering a wrong answer was still associated with better final test performance than choosing "I have no idea," $t(63) = 3.32$, $d = 0.41$.

Effects of forced guessing.   When subjects indicated "I have no idea" on a randomly chosen half of the trials, the computer insisted that they produce a guess (as we have seen, these guesses were rarely correct). Subjects in the skip condition exhibited marginally better performance (23.1%) than those in the forced-guess condition (19.0%) on the final test, but this difference was not statistically significant, $t(63) = 1.56$, $p = .12$. The 95% confidence interval for the mean difference in performance ranged from -1.1% to 9.3%.[1] Of the errors made on the final test in the forced-guess condition, 10.7% were intrusions from prior incorrect guesses.

Initial test response confidence and final test performance. Table 3 lists final test performance as a function of accuracy and response confidence on the initial test. Not surprisingly, recall on the final test appeared to be a positive function of confidence in correct answers given on the initial test. This observation was supported by a significant within-subject gamma correlation between confidence (for correct items on the initial test) and final

performance of .27 ($p = .007$). What was less expected was a similar trend observed for items that were wrong on the initial test—confidence was also positively related to final performance, with a gamma correlation of .21 ($p = .001$). That is, greater confidence in spontaneously produced errors on the initial test was associated with greater accuracy on the final test. One might have expected instead that when subjects spontaneously produced the wrong response, the more confidence they had in this response; and therefore, the more difficult it would be to unlearn it and replace it with the correct response. The observed pattern is consistent with the hypercorrection effect reported by Butterfield and Metcalfe (2001), who found that errors made with high confidence were paradoxically better corrected after feedback.

Summary.   This study examined the consequences of being forced to guess in a fact-learning task on trials in which the subject had said that he or she had "no idea" of the answer. Two basic findings emerged. First, being forced to guess and getting it wrong had no significant effect on the probability of recalling the correct response the following day (our reason for not focusing on the trend toward slightly worse performance in the forced-guess condition will become apparent when we get to Experiment 2). The data offer little support for behaviorists' assertion that producing an error automatically results in its being "stamped" into the mind of the respondent. The rate of intrusions of prior errors when subjects were forced to guess was low, albeit not zero. In any case, just because the same error recurs on the final test does not mean that the prior error necessarily caused the subsequent error. Guessing erroneously does not seem to have any notable overall cost for learning, or if it does, the costs are approximately canceled out by some comparable benefit. Of course, this conclusion may be limited to the case in which the immediate corrective feedback is provided, as in the present experiment.

Second, willingly-offered incorrect responses are a positive prognostic sign, in the sense that they indicate a greater likelihood of ultimately achieving correct learning, as compared to a refusal to guess. This, too, would seem surprising if one assumes that incorrect guesses reflect associations that will need to be unlearned

---

[1] Hoenig and Heisey (2001) argue that performing retrospective power analyses to guide the interpretation of statistically nonsignificant results is inappropriate, and recommend instead the reporting of confidence intervals.

Table 2

Mean Accuracy on the Final Test (Day 2) Conditionalized on Subject's Response on the Initial Test (Day 1) in Experiments 1 and 2

| | Initial test (Day 1) outcome | | | | |
| | Voluntary response | | "I don't know" response | | |
Experiment/variable	Correct	Wrong	Overall	Forced guess	Skip
Experiment 1					
Mean final test accuracy (%)	90.7	35.8	21.3	19.0	23.1
Pooled count (no.)	2,151	1,439	1,609	804	805
Day 2 trials[a] (%)	41.4	27.7	30.9	15.5	15.5
Experiment 2					
Mean final test accuracy (%)	92.7	35.7	27.3	28.6	25.7
Pooled count (no.)	1,418	1,004	938	483	455
Day 2 trials[b] (%)	42.2	29.9	27.9	14.3	13.5

[a] $N = 5,199$.   [b] $N = 3,360$.

before the correct linkage can be developed. This issue will be considered further in the general discussion.

Experiment 2

In the first experiment, feedback was provided immediately after subjects responded on each trial of the initial test in Session 1. According to behaviorist theory, the effectiveness of reinforcement depends on its temporal proximity to the behavioral response (e.g., Hull, 1943; see also Pressey, 1950). Therefore, one might expect that the harmlessness of forced guessing could be limited to situations in which corrective information is provided without delay. In other words, it is possible that the harm of forced guessing would emerge when the effectiveness of feedback is reduced (i.e., delayed in time from the moment of the erroneous response). This issue was addressed in Experiment 2.

Method

Experiment 2 was identical to Experiment 1, except that feedback was provided only after the completion of the quizzing on all items in Session 1. After subjects had responded to all 80 questions, the list of 80 facts was shuffled, and each of the facts (i.e., question and correct answer) was displayed for 4 s.

Results and Discussion

Forty-two subjects were tested on 80 items, resulting in a total of 3,360 items.

Initial test (Day 1) performance. On Day 1, subjects volunteered responses on approximately 72% of the test trials (2,422 of 3,360). On the remaining ~28% of trials where subjects did not spontaneously respond, the computer required forced guessing on

Table 3

Final Test (Day 2) Performance as a Function of Response Confidence and Accuracy on the Initial Test (Day 1) in Experiments 1 and 2

| | Confidence (initial test) | | | | |
Variable	1 (*very low*)	2	3	4	5 (*very high*)
Experiment 1					
Correct on initial test					
Wrong on final test (no. of items)	17	16	23	28	93
Correct on final test (no. of items)	43	57	139	192	1,543
Conditionalized performance (%)	71.7	78.1	85.8	87.3	94.3
Wrong on initial test					
Wrong on final test (no. of items)	479	150	137	72	130
Correct on final test (no. of items)	167	58	74	40	132
Conditionalized performance (%)	25.9	27.9	35.1	35.7	50.4
Experiment 2					
Correct on initial test					
Wrong on final test (no. of items)	7	4	10	12	55
Correct on final test (no. of items)	33	35	95	193	974
Conditionalized performance	82.5	89.7	90.5	94.1	94.7
Wrong on initial test					
Wrong on final test (no. of items)	310	102	76	55	137
Correct on final test (no. of items)	116	43	38	46	81
Conditionalized performance	27.2	29.7	33.3	45.5	37.2

Note. No. of items were pooled across subjects.

483 trials (14.4%), and assigned 455 trials (13.5%) to the skip condition (i.e., moved immediately to the next trial).

Of the 2,422 responses produced without any compulsion, 1418 (58.5%) were correct. As in Experiment 1, forced guesses were rarely correct. Of 483 forced guesses, just 14 (2.9%) were correct.

Response confidence and initial test (Day 1) performance. The distribution of confidence ratings and the associated mean accuracy are shown in Table 1. As in Experiment 1, confidence and accuracy appeared to be related: Responses made with higher confidence tended to be more accurate than those given with lower confidence. This positive relationship was evidenced by a mean within-subject Goodman–Kruskal gamma correlation between response confidence and accuracy of .80 ($p < .001$).

Final test (Day 2) performance. Overall accuracy on the final test was 56% (1,882/3,360). Table 2 shows performance on the final test as a function of the subject's response to an item on the initial test. For items on the initial test that subjects answered voluntarily and correctly, final test accuracy averaged across subjects was 92.7%. For items that were answered incorrectly (and voluntarily), final test accuracy averaged 35.7%. For items eliciting "I have no idea" on the initial test, final test accuracy averaged 27.3%. The first of these three means was significantly different from both the second, $t(41) = 26.64$, $d = 4.11$, and the third, $t(41) = 25.97$, $d = 4.01$.

As in Experiment 1, final test performance was better for items on which subjects volunteered an incorrect response on Day 1 as compared with items that elicited "I have no idea" responses, $t(41) = 4.05$, $d = 0.62$. Even when the analysis was restricted to spontaneous errors that were made with the lowest confidence, volunteering a wrong answer was still associated with reliably better final performance than having "no idea," $t(40) = 2.10$, $d = 0.33$.

Effects of forced guessing. In this experiment, for items on which subjects indicated they had "no idea," final test accuracy on skip items (25.7%) was worse than accuracy on forced-guess items (28.6%), but this difference was not statistically significant, $t < 1$. Of the errors made on the final test in the forced-guess condition, 14.4% were intrusions from prior incorrect guesses.

Initial test response confidence and final test performance. Table 3 lists final test performance as a function of accuracy and response confidence on the initial test. As in Experiment 1, the probability of being correct on the final test appeared to be a positive function of confidence for items that were correct on the initial test. But the mean within-subject gamma correlation between confidence (for correct items on the initial test) and final performance was not significantly different from zero, partly due to a restriction of range (i.e., very high chance of success on the final test for items that were correct on the initial test; gamma cannot be calculated for a subject if one of the variables is a constant). A similar trend was observed for items that were wrong on the initial test—greater confidence in spontaneously produced errors on the initial test was generally associated with greater accuracy on the final test. But again, the mean within-subject gamma correlation was not significantly different from zero.

Summary. The results in Experiment 1 were largely replicated in Experiment 2. There was no significant effect of being forced to guess when subjects thought they did not know the correct answer. While it was our suspicion that in the absence of immediate feedback, forced guessing might well have deleterious effects, the results actually showed a trend in the opposite direction, with a slight but unreliable benefit from forced guessing (in Experiment 1, the trend ran in the opposite direction). Putting the results of the two experiments together, it seems reasonable to conclude that forced guessing does not impair fact learning to any notable degree, when feedback is provided fairly soon[2] (but not necessarily immediately). As in the previous experiment, we also found that spontaneous errors of commission are more likely to be corrected, as manifested on a later test, than are errors of omission.

Experiment 3

The previous two experiments showed that producing a wrong guess had no discernable negative impact on subsequent learning of short, discrete facts. Yet information learned in the real world is often more complex. Moreover, guessing in the real world (and especially on classroom tests) may often involve extended construction of an answer, rather than just production of a single isolated fact. It is therefore important to ascertain whether the earlier results generalize to the learning of factual information with richer conceptual structure.

In the third experiment, subjects learned the explanations for various common phenomena, some pertaining to science and others to history or culture. Although these were common phenomena that are familiar to most people, the causal explanations are unknown to the college student sample used (as confirmed by pilot work and the results of Experiment 3). Each item consisted of a question and an answer; the correct answer was a paragraph, rather than just a single word or phrase. Given the increased length of each answer, far fewer items were used than in Experiments 1 and 2. Also, there was no study phase prior to the initial test. Instead, for half of the items, subjects were asked to construct a plausible account before they were shown the correct answer; for the remaining items, the phenomenon and its explanation were presented concurrently. In other words, we used the prequestioning or pretesting method described in the introduction of this article, and this was done for reasons of feasibility—to obtain sufficient numbers of forced-guess versus no-guess items per subject in order to provide a meaningful comparison.

Indeed, we believed that this experiment, perhaps more so than Experiments 1 and 2, should reveal harmful effects of guessing, if any exist. This is because subjects would be composing explanations for familiar phenomena that they presumably found plausible. Since in some sense, the guesses produced would be relatively believable (even if objectively wrong), such a situation would seem most likely to cause interference with the learning of the correct information. Moreover, past research has demonstrated that memory for prose passages (which have richer conceptual structure than isolated facts or word lists) can be very resistant to updating—subjects' errors across multiple recall trials tended to persist despite opportunities to reread the passage (Fritz, Morris, Bjork, Gelman, & Wickens, 2000; Kay, 1955).

---

[2] We also ran another experiment like the two presented here but in which no feedback was provided on the initial test. The results showed final test performance very near zero for both the continue and forced-guess conditions, which is not surprising in light of prior findings on the necessity of feedback after errors (Pashler, Cepeda, Wixted, & Rohrer, 2005).

Method

Subjects. Fifty-four undergraduates participated in partial fulfillment of course requirements.

Stimuli. A list of 12 common phenomena and their corresponding explanations was compiled: half were scientific (e.g., "Why does the moon influence the Earth's tides more than the sun, even though the sun has the greater gravitational pull?"), and half related to history or culture (e.g., "Why do dimes, quarters, and half-dollars have notched edges?"). The items were divided into two equal sets (with three scientific phenomena in each set); one for the pretest condition and one for the study-only condition. The assignment of sets to conditions was fully counterbalanced across subjects.

Design and procedure. Subjects were informed that their task was to learn the explanations for particular phenomena for an upcoming test. They were told that the information would be presented in the form of questions and corresponding answers and that for half of the items, they would be presented first with a question only, and their job was to construct an answer to the question, after which they would be shown the correct answer (pretest condition). It was emphasized to the subjects that we did not expect them to have prior knowledge of the correct explanations but that they should nonetheless strive to compose plausible answers, each with at least two to three sentences. Subjects could take as much time as they wanted to make their guess. To help motivate subjects to comply with these instructions, we offered a $15 cash bonus for each of the six subjects with the overall most plausible guesses. Following the composition of each guess response, subjects rated their confidence in the correctness of the response (5-point scale), after which the question was presented with the correct answer for 90 s. For the other half of the items, the guessing phase was bypassed, and the questions were presented immediately with their answers for 90 s each (study-only condition). Subjects were reminded that for the entire duration of the presentation of the question with its correct answer, they should study and think about the information with a goal of maximizing performance on a later test. Items in the pretest and study-only conditions were combined in a single list and randomly ordered. At the end of the training phase, subjects watched a 10-min unrelated video (i.e., distractor task). Subjects who were assigned to the immediate-test condition ($n = 30$) then completed a cued recall test on all items (self-paced). Those who were assigned to the delayed-test condition ($n = 24$) returned to the lab 1 week later for the test.

Results and Discussion

The correct answer to each test question was divided into three critical idea units, so that each response received a score between 0 and 3, inclusively. All responses were scored by two independent raters who were blind to experimental condition, and interrater reliability was high for the final test scoring (.84 at the level of each item; .97 at the level of subject totals). The results reported are based on the average of the two raters' scoring.

Pretest (guess) performance. Although accuracy of the responses on the pretest was not the focus of this experiment, we scored the guesses that subjects produced on the pretest for accuracy to check that subjects indeed lacked prior knowledge of the correct information. Mean proportion correct on the pretest was .02, confirming that subjects knew very little about the true causal explanations for the phenomena in question. It was not the case, however, that subjects simply did not bother to make reasonable guesses—the responses were also rated for plausibility on a 4-point scale ($1 = $ *not at all plausible*; $4 = $ *very plausible*), and 86.7% of responses received a mean rating at the midpoint (i.e., 2.5 when averaged across two raters) or higher, suggesting that subjects by and large did make an effort to generate answers that could conceivably be true.[3]

Table 4 lists the distribution of confidence ratings given by subjects for their pretest responses. Although the most common rating was of the lowest confidence level (41%), an absolute majority of responses (59%) were made with a relatively higher level of confidence that the response could be correct, suggesting subjects often had some belief in the possible accuracy of the explanations they generated.

Final test performance. Mean final test performance is shown in Table 5. Retention interval had the expected impact on recall, with subjects in the immediate test condition outperforming those in the delayed-test condition (75.4% vs. 57.4%). Forcing subjects to guess, however, seemed to have a negligible effect on learning (pretest: 67.0%; study-only: 67.9%). These observations were supported by a 2 (test delay) $\times$ 2 (pretest vs. study-only) mixed analysis of variance, which revealed only a significant main effect of test delay, $F(1, 52) = 10.99$, partial $\eta^2 = .174$. There was neither a main effect of pretesting nor an interaction with test delay, Fs < 1. The results did not change when we excluded from the analysis the handful of items that subjects got partially correct on the pretest.

We assessed whether confidence on the pretest (where the guess responses were almost always wrong) was associated with differences in learning of the correct answer. Due to the small number of pretest questions per subject, it was not viable to break the data down into five confidence levels (i.e., there would be too many missing cells to perform a proper analysis). Instead, the items were divided into two sets: (a) those for which subjects gave a confidence rating of 1 (i.e., the lowest level) and (b) everything else (confidence ratings of 2–5). In other words, we compared the impact of guesses that subjects believed were wrong against guesses that subjects believed (with relatively higher confidence) could potentially be correct. Subjects' confidence in their pretest response was not associated with any differences in learning on the final test, $t < 1$, in contrast to the evidence of greater accuracy on the final test for higher confidence errors found in Experiment 1 (cf. Butterfield & Metcalfe, 2001).

Summary. Producing wrong guesses again did not impair subsequent learning of the correct information, even when the guesses were generally plausible answers. In addition, interfering effects of wrong guesses did not appear even after a 1-week delay, thus reinforcing the main conclusions from the earlier two experiments. Confidence for the guess responses did not demonstrate a relationship with later learning, unlike in the previous experiments,

---

[3] Two research assistants who were unaware of the correct answers rated the pretest responses for plausibility. They were instructed to consider whether each response could conceivably be a true explanation for the phenomenon in question.

Table 4

Distribution of Confidence Ratings for the Pretest (Guess) Responses in Experiment 3

	Confidence rating (%)			
1 (*very low*)	2	3	4	5 (*very high*)
41.0	26.2	19.8	8.0	4.9

possibly because there were too few items per subject and hence insufficient power to detect a subtle relationship or perhaps because the relationship may not generalize to more complex materials.

General Discussion

The primary purpose of the present study was to answer the unresolved question of whether learning of facts is affected by guesses, especially incorrect guesses. Influential behaviorists have long asserted that learning is facilitated by the avoidance of errors, for the occurrence of an error does more harm than punishment can offset (e.g., Thorndike, 1932). The claim is that the commission of an error causes it to be ingrained in memory, hence interfering with correct learning (e.g., Guthrie, 1942). Although there were previous studies that attempted to address this issue (e.g., Forlano & Hoffman, 1937), ours is the first to use a procedure that, in our opinion, provides a fairly straightforward model for examining consequences of incorrect guessing for fact learning.

In Experiments 1 and 2, when a subject indicated that he or she was unable to offer an answer for a particular question, the computer randomly decided whether or not to insist that the subject make a guess. Independently assigning items to either the guess or no-guess condition from among those that subjects decide they cannot answer offers two advantages over the prequestioning or pretesting method that was customary in previous research. First, our procedure guarantees that the guess response will almost always be incorrect. In previous studies in which the pretesting method was used, pretest performance was sometimes not reported or on occasion was unexpectedly high (see, e.g., Richland et al., 2009). While the items that subjects answer correctly on the pretest are typically excluded from the analysis, what remains is a comparison of guess and no-guess conditions that is based on different sets of items. The situation is made even worse when pretest performance is not reported or taken into account—one has no idea whether subjects had prior knowledge about some of the items (in which case they were not really guessing the answer). Second, our random assignment technique allows subjects an opportunity to study the information before being tested (unlike the pretesting

method) and hence is arguably more similar to typical testing situations in the real world.

Of course, the pretesting method is not without its own merit. When more complex materials were used in Experiment 3, the number of items had to be reduced, which meant that pretesting was the only feasible way to ensure that we obtained a sufficient number of guess responses from each subject. Regardless of the method used to elicit erroneous guesses from subjects, the results from the present experiments are consistent in suggesting that producing erroneous guesses does not impair learning by any significant degree.

As to why our results diverged from previous studies involving guessing, we point to differences in methodology as the likely factor. Forlano and Hoffman (1937), for example, found that guessing impairs learning of foreign vocabulary, relative to a read-only condition. In that study, the overall time spent in the guessing (pretesting) and read-only conditions was reported to be equal, meaning that the subjects probably had less time to study the correct answer in the guessing than read-only condition. There were insufficient procedural details to assess whether the subjects had enough time to study the feedback; it is possible that their classroom experiment involved a guessing condition that was too rushed, leading to a decrement in learning.

In Experiment 3, even when the to-be-learned material consisted of short paragraphs with more elaborate conceptual structure than discrete facts, generating an erroneous (but plausible) answer did not affect subsequent learning from corrective feedback. At first blush, this finding might appear inconsistent with prior research showing that memory for text passages can be highly resistant to change, such that errors tend to recur across repeated test trials even after rereading (e.g., Fritz et al., 2000). Procedural differences can account for the apparent divergence: In those earlier studies, repeated reading probably had minimal impact on the underlying memory representation of the text because repeated exposure increased subjective familiarity of the text, which may have led to less thorough processing during each rereading opportunity. In contrast, in our study, subjects received only a single presentation of the correct answer (whether preceded by a guess or not), so high levels of familiarity would not likely have been an issue.

Other previous studies found a benefit of guessing, in the context of a pretesting manipulation, in which subjects had to guess the answers to questions before even studying the material. The advantage of pretesting has been attributed to an increase in curiosity about the answer (Berlyne, 1954a, 1954b) and deeper processing of the question and the subsequent correct answer (Kornell et al., 2009). In Experiment 3, we used a pretesting procedure, yet no advantage was found for the pretest condition.

Table 5

Final Test Performance as a Function of Test Delay and Training Condition in Experiment 3

Test delay	Study-only condition	Pretest		
		Overall	Confidence = 1	Confidence ≥ 2
Immediate (10-min)	75.8	75.0	73.7	75.7
Delayed (1-week)	57.9	56.9	54.5	59.1

Perhaps the study material, which consisted of explanations for various everyday phenomena, inherently induced a high level of curiosity (people tend to be relatively more curious about things that they have some familiarity with; see Berlyne, 1966), thus wiping out the positive effect of pretesting. Alternatively, it could be that the pretesting advantage is not reliable for more complex factual material.

Before we discuss the practical implications of our main findings, we will first discuss an aspect of the present data that we think might shed some light on the role of metacognitive errors in learning.

Hypercorrection Effect

In Experiment 1, response confidence for wrong responses on the initial test was positively associated with correct recall on the final test—that is, higher confidence errors were better corrected (this trend was observed in Experiment 2 too, but it was not statistically reliable). This unintuitive relationship between response confidence in an error and subsequent correction has been termed *hypercorrection* (Butterfield & Metcalfe, 2001; see also Kulhavy, Yekovich, & Dyer, 1976). One explanation for this effect is that when an individual responds to a question with high confidence and finds that the response is an error, the feedback is unexpected or surprising, leading him or her to encode the feedback more deeply or pay more attention to it (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009).

Although the positive association between confidence in errors given during the initial test and final recall accuracy found in the first experiment appears consistent with the surprise hypothesis, further consideration of the data suggests some discrepancies. First, if surprise at the feedback causes better learning, then the same hypercorrection should occur for low-confidence *correct* responses—that is, subjects did not expect these responses to be correct, so the feedback should also be surprising and hence better retained on a later test (Butterfield & Metcalfe, 2006; Fazio & Marsh, 2009). In our data, however, we found that confidence for correct responses was positively related to final accuracy, with the correct responses given with the lowest confidence associated with the lowest accuracy on the final test. Second, the lowest confidence wrong responses (i.e., subjects should not have been surprised by the feedback) in Experiments 1 and 2 were still associated with significantly higher final accuracy than items for which subjects refrained from venturing a response, and this difference is not predicted by the surprise hypothesis.

A possible alternative account is that an individual's knowledge of or familiarity with the domain of a test question influences the type of error made and the probability of subsequent error correction (Butterfield & Metcalfe, 2001). In particular, the willingness to volunteer a low-confidence guess (even if it is wrong), relative to withholding a response, may reflect a higher degree of domain knowledge or familiarity, which also facilitates learning of the corrective feedback (i.e., it is easier to integrate new information into existing related knowledge structures). In other words, an individual's decision to either volunteer a guess or omit a response may be a predictor of later successful learning.

It should be pointed out that the two accounts of the hypercorrection effect—surprise and domain knowledge—are by no means mutually exclusive. There is empirical support for the surprise

hypothesis, but some aspects of our data are incongruent with that account. Future research is necessary to examine whether our speculations about how domain knowledge influences error correction and the decision to volunteer or withhold a response are supported empirically.

Practical Implications

More than a century of research on learning and memory has yielded great insights into the nature of our memory systems and the factors that influence learning, with the potential for direct pedagogical application in the classroom or in computer-aided instruction systems. However, practical translation from memory research has often been lacking in educational practice, and one reason may be that contemporary researchers have not focused enough on the sorts of concrete procedural choices that arise in real learning situations (Pashler, Rohrer, Cepeda, & Carpenter, 2007).

The main findings provide some reassurance that requiring a learner to produce factual information even when the learner is highly unsure does not seem to produce a detrimental effect on the ability to profit from feedback. This is important, because a growing body of evidence indicates that retrieval promotes learning and retention better than does rereading (for a review, see Roediger & Karpicke, 2006), and this "retrieval practice" will, of course, occasionally produce incorrect answers. Aside from the use of practice tests in the classroom to promote learning, many computer-based tutoring programs (e.g., Smart.fm, SpacedEd, SuperMemo) incorporate retrieval practice as a core component. Our results suggest that these programs need not discourage students from guessing when they are unsure of the answer. In addition, spontaneous guessing (even if incorrect) may indicate a higher state of learning than the withholding of a response, which may have implications for tutoring systems that feature adaptive sequencing of items based on a user's responses.

It should be pointed out, however, that the lack of any harmful consequences of erroneous guessing may well have important boundary conditions that remain to be determined. In Experiments 1 and 2, the conclusion of a null effect of erroneous guessing was based on items for which subjects chose not to volunteer a response (i.e., they were forced to guess when they had no idea of the answer). In other words, our subjects were clearly aware that the responses they were producing were almost certainly errors. Making errors in situations where errors are easily recognized as such by the subject might not be as harmful as guessing in more ambiguous situations. What about a situation in which the subject responds spontaneously with an erroneous response? Certainly, that too could be considered guessing, especially if the response is not made with high certainty. We did find that responses on the initial test that were offered spontaneously and were wrong (even those given with the lowest confidence) ended up being better corrected than items that subjects chose not to answer. Of course, it would be impossible to randomly assign items to a "spontaneous" guessing condition, and hence any analysis of these spontaneous guesses would necessarily be correlational.

Also, it may be that there are forms of nondeclarative memory for which producing an error does indeed "stamped-in" responses (e.g., poorly planned motor actions like bad golf swings or tennis serves). Even for verbal materials, there may be certain types of learning tasks in which familiarity may play an especially critical role, for which

incorrect guessing might be harmful (e.g., spelling; see Jacoby & Hollingshead, 1990, for evidence that being exposed to other people's spelling errors impairs subsequent spelling performance).

For the explicit learning of facts, at least, the present results make a reasonably strong case that with individuals not suffering from memory disorders, wrong guesses do not hurt acquisition of the correct information as long as feedback is provided soon after the errors—and thus that it makes sense to try to obtain the benefits of testing without being overly concerned about the possibility that erroneous responses will result in deleterious learning.

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

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia, 32,* 53–68. doi:10.1016/0028-3932(94)90068-X

Berlyne, D. E. (1954a). A theory of human curiosity. *British Journal of Psychology, 45,* 180–191.

Berlyne, D. E. (1954b). An experimental study of human curiosity. *British Journal of Psychology, 45,* 256–265.

Berlyne, D. E. (1966). Conditions of prequestioning and retention of meaningful material. *Journal of Educational Psychology, 57,* 128–132. doi:10.1037/h0023346

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

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition & Learning, 1,* 69–84. doi:10.1007/s11409-006-6894-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Evans, J. J., Wilson, B. A., Schuri, U., Andrade, J., Baddeley, A., Bruna, O., . . . Taussik, I. (2000). A comparison of "errorless" and "trial-and-error" learning methods for teaching individuals with acquired memory deficits. *Neuropsychological Rehabilitation, 10,* 67–101. doi:10.1080/096020100389309

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review, 16,* 88–92.

Forlano, G., & Hoffman, M. N. H. (1937). Guessing and telling methods in learning words of a foreign language. *Journal of Educational Psychology, 28,* 632–636. doi:10.1037/h0056518

Fritz, C. O., Morris, P. E., Bjork, R. A., Gelman, R., & Wickens, T. D. (2000). When further learning fails: Stability and change following repeated presentation of text. *British Journal of Psychology, 91,* 493–511. doi:10.1348/000712600161952

Guthrie, E. R. (1942). Conditioning: A theory of learning in terms of stimulus, response, and association. *Yearbook of the National Society for the Study of Education, 41*(2), 17–60.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Hull, C. L. (1943). *Principles of behavior.* New York, NY: Appleton-Century-Crofts.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology, 44,* 345–358.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20,* 963–973. doi:10.1111/j.1467-9280.2009.02402.x

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558. doi:10.1080/09541440601056620

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology, 46,* 81–100.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 989–998. doi:10.1037/a0015729

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68,* 522–528. doi:10.1037/0022-0663.68.5.522

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.

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the Cognitive Science Society* (pp. 975–980). Mahwah, NJ: Erlbaum.

Parlow, J., & Berlyne, D. E. (1971). The effect of prior guessing on incidental learning of verbal associations. *Journal of Structural Learning, 2,* 55–65.

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

Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14,* 187–193.

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology, 29,* 417–447. doi:10.1080/00223980.1950.9916043

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect on true–false examination questions. *Journal of Educational Psychology, 17,* 52–56. doi:10.1037/h0070067

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

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. doi:10.1037/a0016496

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2009). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies and D. B. Wright (Eds.), *Current issues in applied memory research* (pp. 13–49). New York, NY: Psychology Press.

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. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159. doi:10.1037/0278-7393.31.5.1155

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986, October 9). Learning representations by back-propagating errors. *Nature, 323,* 533–536. doi:10.1038/323533a0

Skinner, B. F. (1958, October 24). Teaching machines. *Science, 128,* 969–977. doi:10.1126/science.128.3330.969

Squires, E. J., Hunkin, N. M., & Parkin, A. J. (1997). Errorless learning of novel associations in amnesia. *Neuropsychologia, 35,* 1103–1111. doi:10.1016/S0028-3932(97)00039-0

Taber, J. I., Glaser, R., & Schaefer, H. H. (1965). *Learning and programmed instruction.* Reading, MA: Addison–Wesley.

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

Thorndike, E. L. (1932). *Fundamentals of learning.* New York, NY: Teachers College Bureau of Publications. doi:10.1037/10976-000

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true–false examinations. *Journal of Educational Research, 83,* 119–124.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Research, 86,* 357–362. doi:10.1080/00220671.1993.9941229