THE EFFECT OF GENERATING ERRORS ON SUBSEQUENT LEARNING AND GENERALIZATION

BY

JESSICA SILER

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Professor Aaron S. Benjamin

ABSTRACT

Historically, teaching methods that avoid having students make errors have been favored by educators and learners. This choice was motivated by the belief that errors made during study are likely to persist in memory and consequently interfere with subsequent learning. However, mounting evidence suggests that generating errors during study may actually benefit learning. The *error generation benefit* is the finding that production of errors during study can enhance subsequent learning of the correct study material. It is important to determine if the error generation benefit will support more than rote learning, however. The experiments reported here examine the effect of error generation on generalization and inference in order to better understand the effects of errors on learning more broadly, and to better inform educational applications. Three experiments investigated whether the potential learning benefits that come with this kind of errorful learning would extend to cases that require transfer of knowledge to new situations or problems. Experiments 1 and 2 employed a bird categorization task and Experiment 3 employed an age estimation task. Results from Experiments 1 and 2 do not reveal an error generation benefit, though these results may reflect limitations inherent in the task. Results from Experiment 3 suggest an error generation benefit for recognition memory, but no such benefit for generalization to new stimuli. Although these results do not reveal benefits from error generation, they also provide no evidence that errors are harmful to learning, as is suggested by some theoretical perspectives.

Keywords: errors, generation, memory, learning, generalization

TABLE OF CONTENTS

INTRODUCTION

The benefits to memory of testing and self-generation are critically important for both theoretical and applied research in human learning and memory. The testing effect is the finding that testing enhances memory for material more than simply restudying the material (Benjamin & Pashler, 2015; Carrier & Pashler, 1992; Roediger & Karpicke, 2006a). Similarly, the generation effect is the finding that self-produced material is better remembered than read material (Slamecka & Graf, 1978). As impressive as these effects are in laboratory settings, educators are wary of applying these findings in their curricula because of the potentially harmful effects of errors (Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel, & Metcalfe, 2007). The testing and generation effects are well established for successful cases in which material is correctly retrieved or generated, but what of the unsuccessful cases when errors are made? Do these errors also share in the benefits of the effects and improve learning? Or do these errors impair learning? And ultimately, what (if any) effect do errors made during study have on applying knowledge to novel contexts? There is an increasing body of evidence that suggests that making errors can facilitate learning under certain circumstances. The goal of the current studies is to investigate whether the potential learning benefits that come with making errors extend to cases that require transfer of knowledge to new situations or problems.

ERRORLESS VS. ERRORFUL LEARNING

Historically, psychologists and educators believed errorless learning to be the most effective route to acquiring knowledge and skills. The assumption is that errors committed during study would persist in memory and hinder any subsequent learning (Guthrie, 1952). Allowing commission of errors may encourage rehearsal of those errors which may be extremely difficult to correct in the future. Across many American classrooms, teachers focus on outlining correct approaches to solving problems and discourage any exploratory approaches that may lead to student errors (Metcalfe, 2017).

There is some truth to this view that errors may be harmful. Sometimes errors committed on initial tests may reappear on later tests (Butler & Peterson, 1965; Roediger & Marsh, 2005; Marsh, Roediger, Bjork, & Bjork, 2007). In addition, errorless learning may be the best approach to take for people with memory impairments (Clare & Jones, 2008). Even at the theoretical level, it is easy to see how errors could be detrimental to learning. A reasonable interpretation of the testing effect would predict a stronger association in memory between the study material and the error. Similarly, the generation effect would predict stronger memory for the error because it was self-generated.

Despite evidence that errors can have negative effects on learning, there is also a considerable amount of evidence that demonstrates that errors may not be as harmful as once thought, and may even help learning under certain conditions. Early work demonstrated how repeated errors can lead to an enhancement of subsequent encoding. Izawa (1970) showed that multiple unsuccessful tests before receiving feedback could enhance the encoding of that feedback. Further work demonstrated how unsuccessful generations could benefit learning. Kane and Anderson (1978) demonstrated better memory for sentences that were learned by guessing the final word than sentences that were read in their entirety, despite the former condition having produced many incorrect guesses. Slamecka and Fevreiski (1983) showed a similar advantage for word pairs. Richland, Kornell, and Kao (2009) found that being pretested on to-be-learned material led to better memory for that material, even when those pretests elicited very poor performance. The following section will review one particularly straightforward case of a memory benefit from error commission (Kornell, Hays, & Bjork, 2009).

Error Generation Benefit

The error generation benefit can be thought of as an extension of the testing and generation effects. As long as feedback is provided, the error generation benefit can be likened to the testing effect for failed tests and/or the generation effect for failed generations. Specifically, the error generation benefit is the finding that production of errors can enhance subsequent learning of a correct response to the cue.

Standard Paradigm. Here a direct and well replicated case of error generation is reviewed (Experiment 4 from Kornell et al., 2009). In this experiment, subjects studied weakly associated word pairs under two different study conditions: Read or Guess. In the Read condition, subjects were shown the cue and target words together (ex: olive-branch) for a duration of 13 seconds and were instructed to study them for a future memory test. In the Guess condition, subjects were given the cue and given 8 seconds to guess the target (ex: whale-???) before given the correct target (ex: whale-mammal), which was presented for an additional 5 seconds. Though total study time for a single pair in each condition was equivalent, the Guess condition spent considerably less time with the correct information. After the study phase, subjects completed a 5-minute distraction task and then completed the final cued-recall test. In order to study the causal effect of error generation, materials used in the Guess condition need to elicit high initial error rates to minimize the item-selection artifact that would arise from examining only the subset of items guessed incorrectly. This precondition was met by using word pairs with low associative strengths. Use of these stimuli virtually guarantees that almost all initial responses to the cues were errors. The few items that were initially correct were eliminated from further analyses. Results revealed a significant advantage in recall accuracy for material studied in the Guess condition over the Read condition.

In the years since Kornell, Hays, and Bjork (2009), a number of studies have utilized this paradigm and have replicated and extended this error generation benefit. The benefit seems to disappear when initial responses are strongly constrained (ex: tide-wa__), when corrective feedback is delayed, or when word pairs are completely unrelated (ex: pillow-leaf; Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Vaughn & Rawson, 2012; Hays, Kornell, & Bjork, 2013; Potts & Shanks, 2014). However, the benefit extends to cases when novel stimuli are learned (such as obscure English words and foreign language vocabulary), when final tests are delayed, or when episodic (instead of semantic) retrieval attempts are emphasized (Potts & Shanks, 2014; Yan, Yu, Garcia, & Bjork, 2014; Knight et al., 2012).

Explanations of the Effect. Three main explanations of the error generation benefit have been considered. They are neither exhaustive nor totally independent of one another.

The *semantic activation hypothesis* proposes that when a cue is presented the subject activates a network of related concepts in order to guess the target. Though the initial guess may be wrong, that activation may facilitate subsequent encoding of the correct target. This notion is supported by the finding that the error generation benefit does not extend to unrelated word pairs (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012). Semantic activation cannot provide a complete account of the effect, though, because the benefit does extend to novel stimuli, for which no previous semantic relations are thought to exist (Potts & Shanks, 2014).

The *mediator hypothesis* suggests that the committed errors act as mediating cues which provide an additional route from the cue to the target. For instance, when presented with a cue at final test, the subject first remembers their original error and that may help them remember what the correct answer was. In support of this view, researchers have found that when prompted, subjects were usually able to either recognize or produce their initial guesses as well as the correct answers at test (Vaughn & Rawson, 2012; Potts & Shanks, 2014; Yan et al., 2014). Memory for both their original guesses and the correct answers suggests that errors do not appear to interfere with memory for correct material and may actually aid the retrieval process.

The *attention hypothesis* proposes that error commission motivates subjects to devote more attention and effort to encoding the correct information. This view is supported by findings from a related phenomenon, the hypercorrection effect. The hypercorrection effect is the finding that errors committed with high confidence are more likely to be corrected after feedback than errors committed with low confidence (Butterfield & Metcalfe, 2001). The reasoning may be that when feedback is surprising (i.e., when it deviates most from expectations) it captures attention, and ultimately leads to better memory for that feedback (Butterfield & Metcalfe, 2006). Similarly, curiosity in finding out the correct answer after committing an error improves encoding (Berlyne & Normore, 1972).

TRANSFER OF LEARNING

The error generation benefit and related effects offer insight into how to enhance the retention of information. However, learning is more than just rote memorization. The ultimate goal of learning is to be able to apply knowledge to new contexts. So, for findings like the error generation benefit to be practical and to inform educational outcomes, they need to support generalization and transfer.

There is evidence that testing may enhance the transfer of learning (Rohrer, Taylor, & Sholar, 2010; Carpenter, 2012). In a particularly relevant example, Jacoby, Wahlheim, and Coane (2010) extended the testing effect into the realm of category learning. Throughout their set of experiments, subjects studied bird families by either Repeated Testing or Repeated Study. The study set contained 40 total birds – 5 exemplars from each of 8 bird families. In the Repeated Test condition, subjects were shown the birds individually and were asked to categorize them into their respective families. In the Repeated Study condition, subjects were shown each bird paired with its family name. In both study conditions, subjects were exposed to the entire study set of birds multiple times before the final test. The final test consisted of 80 birds – half were those previously studied (Studied) and half were new birds (Novel) from the same 8 bird families. During the test, subjects were asked to categorize each bird into its family and to report whether or not they had studied each particular bird earlier. The Repeated Testing condition led to higher recognition accuracy for previously studied birds, and, more importantly, higher categorization accuracy for both Novel, as well as previously studied, birds. This result demonstrates that testing benefits both category learning and generalization. Experiment 1 imports the bird-categorization task of Jacoby et al. into an error generation paradigm.

EXPERIMENT 1

The current experiment examines the effect of generating errors on memory for category membership and generalization of category rules.

METHOD

Participants. Fifty-nine undergraduate students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit. The data from two subjects were eliminated from analysis because they failed to follow instructions and to complete all tests within the experimental program.

Design. This experiment employed a 2x2 within-subjects design that manipulated Study Condition (Read or Guess) and Study Status (Studied or Novel). Each experimental session was broken down into two halves. Study Condition was blocked such that Read or Guess were assigned to separate halves of the experiment. Order of the Study Conditions was counterbalanced across subjects.

Materials. A total of 80 bird images were used as stimuli in this experiment. These color images were gathered from www.whatbird.com and feature each bird in a perching position. Ten images were selected from each of the following eight bird families: Finch, Jay, Oriole, Sparrow, Warbler, Flycatcher, Thrush, and Swallow. These families were chosen from the same taxonomic order (Passeriformes) to ensure enough between-family similarity to make the categorization task difficult. The assignment of bird families to study condition, and the order in which the birds were seen was randomized for each subject.

Procedure. After signing the consent form and completing a brief demographics questionnaire, each subject was seated at an individual computer station to complete the experimental program. Within each half of the experiment, subjects studied 4 bird families by way of one of the Study Conditions (Read or Guess). Each half was divided into four phases: training, study, distraction, and test. During training, the subject was introduced to the 4 bird families they would study and was allowed practice with a single exemplar from each family to become familiar with the nature of the task. Any exemplars seen during training did not reappear during the experiment. The study phase promptly began after training was complete. During this phase, the subject was presented with 20 bird images (5 exemplars from each of the 4 families). For the Read condition, each bird image was shown at the center of the screen with its family designation printed under it. Subjects were allowed 13 seconds to study each image. For the Guess condition, each bird image was shown at the center of the screen and the subject was prompted to type in the bird's family designation. They were then told if their guess was correct or incorrect and they were provided with the bird's true designation. Subjects were allowed 8 seconds to submit their guess and were allowed 5 seconds with the corrective feedback. At the end of the study phase subjects completed the distraction task. This task required subjects to answer addition problems for 5 minutes before moving on to the test. The test consisted of 40 bird images (10 exemplars from each of the 4 families, half of which were presented in the study phase). Images were presented individually and subjects were asked to provide the family designation of each bird and report whether or not they had seen that particular bird image earlier in the experiment. No feedback was provided during the test. Upon test completion subjects advanced to the second half of the experimental program. This half followed the same format as the first, the only exception being a change in Study Condition. At the end of the experiment each subject was given a debriefing form and dismissed.

RESULTS

Study Phase. During study, subjects correctly categorized 39.5% (sd=15.12) of the birds. This performance was significantly greater than chance (25%), $t(58) = 7.34$, $p < 0.01$. Note that this value is considerably higher than that seen in a typical error generation experiment, in which study phase performance is intentionally kept low in order to minimize item-selection confounds. The presence of this confound to a greater degree than usually seen should be kept in mind when examining the results from the recognition test of the experiment. However, it is worth noting that, since the critical test involves the classification of novel items, the higher-than-usual correct-response rate during study is not problematic.

Recognition. Hit rates, false alarm rates, and *d'* were calculated for each subject. Average hit and false alarm rates for the Guess condition were 0.65 and 0.25, and for the Read condition were 0.62 and 0.25 (see Figure 1). The analysis revealed no significant differences in *d'* between the Guess (M = 1.21, sd = 0.72) and the Read conditions (M = 1.13, sd = 0.70), $t(56) = 0.76$, $p = 0.22$.

Categorization Accuracy. A paired t-test was conducted to determine if subjects in the Guess condition improved in categorization accuracy from study phase to test. Average categorization accuracy of Studied birds at test (M = 0.54, sd = 0.21) was found to be significantly higher than average categorization accuracy during study (M = 0.40, sd = 0.15), $t(56) = 7.71$, $p < 0.01$. This result indicates that over the course of the experiment subjects got better at categorizing the birds.

When all items from the final test were considered, birds studied in the Read condition were categorized more accurately than birds studied in the Guess condition, $F(1,112) = 5.80$, $p =$

0.02. As would be expected, the Studied birds were more accurately categorized than the Novel birds, $F(1,112) = 62.19$, $p < 0.01$. These data are shown in Figure 2.

To examine the effect of generating errors, any items answered correctly during study were eliminated on a subject-by-subject basis. This analysis revealed a similar pattern for Study Condition and Study Status, also shown in Figure 2. Birds studied in the Read condition were categorized more accurately than birds studied in the Guess condition, $F(1,112) = 13.41$, $p < 0.01$, and Studied birds were categorized more accurately than Novel birds, $F(1,112) = 15.32$, $p < 0.01$. The interaction between Study Condition and Study Status was significant, $F(1,112) = 4.75$, $p = 0.03$. Within the Read condition, the Studied items were categorized more accurately than the Novel items, $t(56) = 3.46$, $p < 0.01$. Within the Guess condition, no significant difference was found between the Studied and Novel items, $t(56) = 0.97$, $p = 0.17$.

Because condition and bird families were nested within the counterbalanced variable of Order, it is possible that the blocked design of the experiment could have created order effects. Even-numbered subjects completed the Read condition during the first half of the experiment; odd-numbered subjects completed the Guess condition during the first half. Table 1 presents the same data as seen in Figure 2, but includes the Order variable. Note that this means that comparison of Read and Guess within the Order variable are now between-subjects. The analysis revealed similar effects of Study Condition and Study Status, but there was no significant difference in categorization performance between Even-numbered and Odd-numbered subjects. There is no evidence that the blocked design of the experiment affected the results.

DISCUSSION

These analyses indicate that Read was the superior study condition in terms of total proportion of birds correctly categorized – whether or not only initial errors were considered. At first glance, these results suggest that studying category exemplars by reading is best for both memory for birds and their families as well as generalization to new birds.

Overall, these results do not align with prior results in the error generation literature, but this failure may be due to a variety of factors. The existence of a larger-than-usual item-selection confound in this experiment makes any conclusions about the effect of error generation on recognition memory quite difficult. This problem may have arisen because subjects learned the categories very quickly during the study phase. In a typical error generation experiment, subjects are only exposed to the items once before being tested. In this experiment subjects were exposed to each category (bird family) five times during study. These multiple presentations may have led to more learning within the study phase than the usual error generation experiment. It is possible that this issue could be addressed by expanding the stimuli set to include more bird families and more exemplars to increase the difficulty of the task.

Another problem may lie in the manner of the response selection during the study phase. Before the onset of the study phase, subjects were told which four bird families they would be studying. Thus, being a categorization task, the subject's response set was constrained to just four options. This kind of limitation may have seriously affected the outcome of the results. It could be argued that this study condition did not afford true error generation. Two error generation studies (both follow-ups of Kornell et al., 2009) have found evidence that the error generation benefit does not extend to cases of constrained guessing. When Grimaldi and Karpicke (2012) constrained guessing by providing the stem of the target word (ex: *tide-wa*__), they no longer found an error generation benefit. In fact, recall performance in this constrained Guess condition was significantly worse than in their Read condition. Potts and Shanks (2014) included a Choice study condition in which subjects were presented a cue and had to choose from a small set of possible targets. The error generation benefit was found for their Guess condition, but the Choice condition performed no better than their Read condition. Taken together it looks as if imposing constraints on response sets could eliminate the advantage for the error generation condition.

EXPERIMENT 2

This experiment makes use of the category learning task of Experiment 1 and attempts to address the issues of task difficulty and constrained guessing. Study Condition was changed to a between-subjects variable; therefore, subjects engaged with the entire stimuli set by way of only one condition (i.e., subjects in the Guess condition would use Guess to learn all 8 categories). This change decreases the potential for carryover effects, and make lower study phase performance and avoid item-selection effects on recognition. In addition, a new study condition was introduced to compare performance in a truly constrained guessing situation to the original Guess condition. The current experiment examines the effect of generating errors on memory for category membership and generalization of category rules.

METHOD

Participants. One-hundred twenty-four undergraduate students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit. The data from one subject were eliminated from analysis because they failed to complete the experimental program.

Design. This experiment employed a 2x3 mixed design. As in Experiment 1, Study Status (Studied or Novel) was manipulated within-subjects, but Study Condition (Read, Choice, or Guess) was manipulated between-subjects.

Materials. The stimuli set was identical to that of Experiment 1. A total of 80 bird images were used – 10 images were selected for each of the 8 bird families.

Procedure. After signing the consent form and completing a brief demographics questionnaire, each subject was seated at an individual computer station to complete the experimental program. One third of the subjects (n=41) were assigned to each of the Read, Choice, and Guess conditions. After a brief set of instructions, subjects completed a study, distraction, and test phase. During the study phase of the experiment, subjects were presented with 40 bird images (5 exemplars from each of the 8 families). Subjects in the Read condition were given 13 seconds to study each bird and its corresponding family designation. Subjects in the Guess condition were shown each bird and given 8 seconds to guess to which family the bird belonged. They were then given corrective feedback which remained displayed for another 5 seconds. The Choice condition was identical to Guess except that for each bird subjects were also provided with a list of the 8 family names at the bottom of the screen from which to choose. After the study phase, all subjects completed a 5-minute addition task before the test. The test consisted of 80 total bird images, half of which were seen before during the study phase. Each image was presented individually and subjects were first asked to categorize the bird into its family, then report whether they had seen that particular image earlier in the experiment. Family names were not displayed on the screen during the test and no feedback was provided. The order in which the birds were seen in both the study and test phases was randomized for each subject. Upon test completion, each subject was given a debriefing form and dismissed.

RESULTS

Study Phase. During the study phase, subjects in the Choice and Guess conditions correctly categorized 21.1% (sd=7.96) and 19.76% (sd=7.07) of the birds respectively. Study phase performance did not differ between the Choice and Guess conditions, $t(78.9) = 0.81$, $p = 0.42$. Collapsed across Study Conditions, study phase performance was significantly greater than chance (12.5%), $t(81) = 9.55$, $p < 0.01$.

Recognition. Hit rates, false alarm rates, and *d'* were calculated for each subject. Average hit and false alarm rates were 0.61 and 0.33 for the Read condition, 0.63 and 0.31 for the Choice condition, and 0.62 and 0.28 for the Guess condition (see Figure 3). The analysis revealed no significant differences in *d'* between the Read (M = 0.80, sd = 0.48), Choice (M = 0.90, sd = 0.36), and Guess (M = 1, sd = 0.47) conditions, *F*(1,121) = 1.05, *p* = 0.19.

Categorization Accuracy. A two-way mixed ANOVA was conducted to determine if subjects in the Choice and Guess conditions improved in categorization accuracy from study phase to test. Performance across Choice and Guess conditions did not significantly differ, *F*(1,158) = 0.001, *p* = 0.979. Average categorization accuracy of Studied birds at test (M = 0.32, sd = 0.14) was significantly higher than average categorization accuracy during the study phase (M = 0.20, sd = 0.08), *F*(1, 158) = 9.37, *p* <0.01. Therefore, over the course of the experiment Choice and Guess subjects improved in categorization accuracy.

As in Experiment 1, any items answered correctly during the study phase were eliminated on a subject-by-subject basis to analyze the effect of generating errors. Average values for performance based on categorization accuracy are summarized in Table 2. The interaction between Study Condition and Study Status was significant, *F*(1, 121) = 5.23, *p* = 0.02 (see Figure 4). For the Studied items, the Read condition outperformed both Choice (*t*(80) = -3.43, *p* < 0.01) and Guess (*t*(80) = 3.57, *p* < 0.01). The Choice and Guess conditions did not significantly differ from one another for Studied items, *t*(80) = 0.09, *p* = 0.93. For the Novel items, performance across all conditions did not significantly differ, *F*(1,120) = 3.13, *p* = 0.05. Within the Read condition, performance dropped from Studied to Novel items, *t*(40) = 8.49, *p* < 0.01. The Choice condition displayed a similar pattern, *t*(40) = -3.02, *p* < 0.01. Yet, within the Guess condition, performance on Studied and Novel items did not differ, *t*(40) = -0.16, *p* = 0.87.

DISCUSSION

The Read condition resulted in the best performance for Studied items, as was found in the preceding experiment. For the Novel items no advantage was found for either condition. Within this category learning task, it appears that reading during study leads to better retention, but when it comes to generalization, neither study condition wins out.

The performance of subjects in the Choice condition closely matched the pattern of performance of those in the Guess condition in both this experiment and in Experiment 1. This finding supports the idea that there was a constrained guessing problem inherent in this category learning task. As noted earlier, limiting subject responses during the study phase may not allow for error generation benefits (Grimaldi & Karpicke, 2012; Potts & Shanks, 2014). Therefore, this failure to replicate the error generation benefit may be due to the limitations of this specific task.

Data from Experiments 1 and 2 were combined by Study Status (Studied or Novel) for Bayesian analysis (see Figure 5). Null hypothesis significance testing (NHST) can only report whether or not data are unlikely under the assumption the null is true, and it cannot quantify amount of evidence for either the null or alternative hypotheses (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Bayes factors (B_{10}) represent the ratio of the marginal likelihoods of the null and alternative hypotheses, thus they can report on which hypothesis is best supported. According to Jeffreys' (1961) guidelines for interpretation, B_{10} greater than 3 indicate some evidence, B_{10} greater than 10 indicate strong evidence, and B_{10} greater than 30 indicate very strong evidence. The following Bayes factors are reported in terms of the odds in favor of the alternative. Bayes factors for the effect of Study Condition on Studied and Novel items were calculated. For the Studied items across both experiments, there was strong evidence for superior memory for the Read (M = 0.52, sd = 0.21) condition over the Guess (M = 0.38, sd = 0.21) condition. (B_{10} = 2334.94). The evidence regarding the Novel items was equivocal (B_{10} = 1.37).

EXPERIMENT 3

Given the complications in the previous experiments, category learning may not be the best task for evaluating transfer and generalization in an error generation experiment. Experiment 3 employed a task involving age estimation that differs from the previous task in many ways. Age estimation is a task in which virtually everyone is experienced, so the task requires minimal explanation and practice. Also, because this task elicits numerical responses, a more precise measure of deviation from the true answer can be examined. Unlike the previous experiment, this dependent variable allows for a new type of accuracy comparison between conditions. Additionally, though people are fairly accurate at guessing the ages of others (usually falling within a range of 7 years), there is both need and room for improvement (Rhodes, 2009).

In terms of the current experiment, this combination of characteristics means that study phase performance should be low enough to avoid item-selection effects on recognition, yet training should improve the accuracy of these age estimates. There is also evidence that age estimation may support error generation benefits. McGillivray and Castel (2010) investigated the effects of study condition (Guess or Read) and age (younger or older adults) on memory for age-face associations. Subjects studied 16 unfamiliar faces either by trying to guess the age (Guess) or by being given the age (Read). On a final cued-recall test, subjects were presented with the same 16 faces and asked to recall the correct age of each. Both younger and older adults benefitted from guessing, despite initial guesses usually being incorrect. Results from this study also revealed an own-age bias for both age groups (i.e., younger adults were best at estimating the ages young faces and older adults were best at estimating the ages of older faces).

The current experiment examines the effect of generating errors on memory for face-age associations and also to generalization of age estimation ability to new faces.

METHOD

Participants. Sixty-seven undergraduate students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit.

Design. This experiment employed a 2x2 mixed design. Study Condition (Read or Guess) was manipulated between-subjects and Study Status (Studied or Novel) was manipulated within-subjects. Study Condition could not be manipulated within-subjects in this experiment because it would be impossible to determine if the Novel items were aided by having been in the Read or the Guess condition.

Materials. A total of 96 face images were used as stimuli in this experiment. These faces all had neutral expressions and were gathered from the Park Aging Mind Laboratory Face Database (https://pal.utdallas.edu/facedb/; Minear & Park, 2004). The faces were presented in color and were edited to control for background and clothing. The ages assigned to each face were the ages of each person at the time the photo was taken. The 96 faces can be broken down into four age groups (18-29, 30-49, 50-69, and 70-92), each consisting of 24 faces. Within these four groups, half were female faces and half were male faces. Furthermore, half of the faces were of people who identified as White and the other half were of people who identified as Black, Asian, or Hispanic.

Procedure. After signing the consent form and completing a brief demographics questionnaire, each subject was seated at an individual computer station to complete the experimental program. Half of the subjects were randomly assigned to the Guess condition (n=34) and half to the Read condition (n=33). Similar to previous experiments, this experiment was divided into a training, study, distraction, and test phase. During training, the subject was shown a sample face to demonstrate the nature of the task. This sample face did not reappear during the experiment. The study phase then began. This phase consisted of 48 faces (8 from each of the four age groups, half female, and half White). For the Read condition, each face was shown at the center of the screen with its corresponding age displayed under it. Subjects were allowed 13 seconds to study each face and its corresponding age. For the Guess condition, each face was shown in the center of the screen and the subject was prompted to guess the person's age. They were then told if their guess was correct or incorrect and were provided with the correct age. Subjects were allowed 8 seconds to submit their guess and were allowed 5 seconds to study the correct face-age pair. At the end of the study phase subjects spent 5 minutes completing an addition distraction task. During the test, subjects were presented with all 96 faces (48 of which had previously been seen in the study phase). Each face was presented individually and subjects were asked to first provide an age estimate and then report whether they had seen that face earlier in the experiment. No feedback was provided during test. Upon test completion, each subject was given a debriefing form and dismissed.

RESULTS

Study Phase. During study, subjects correctly guessed ages of only 4.66% (sd=3.21) of the faces. Unlike Experiments 1 and 2, the study phase performance was very low; therefore, the item-selection problem for assessing recognition is not troublesome. Those faces whose ages were correctly guessed were eliminated from the following analyses on a subject-by-subject basis. The average deviation of subjects' guesses was 8.30 years (sd=1.62).

Study phase performance was further divided by age and race of each face. There were four main age groups (18-29, 30-49, 50-69, and 70-92) and two race groups (White and Other [including Black, Asian, and Hispanic]). Average values for performance based on accuracy and deviation are summarized in Table 3. These initial results show highest age estimation accuracy for the young (18-29) faces, *t*(33) = 2.74, *p* < 0.01. Also, age estimations were most precise for the young faces, *t*(33) = -5.99, *p* < 0.01.

Recognition. Hit rates, false alarm rates, and *d'* were calculated for each subject. Average hit and false alarm rates for the Guess condition were 0.75 and 0.10, and for the Read condition 0.69 and 0.13 (see Figure 6). In terms of discriminability (*d'*), there was a significant advantage in the Guess condition (M = 2.23, sd = 0.61) over the Read condition (M = 1.79, sd = 0.66), *t*(65) = 3.15, *p* < 0.01.

Age Estimation Accuracy. A paired t-test was conducted to determine if subjects in the Guess condition improved in their age estimation accuracy from study phase to test. Average age estimation accuracy of Studied faces at test (M = 0.086, sd = 0.045) was found to be significantly higher than average age estimation accuracy during study (M = 0.047, sd = 0.032), *t*(33) = 3.75, *p* < 0.01. This result indicates that subjects improved in their age estimation accuracy over the course of the experiment.

The following analyses concern accuracy on the final age estimation test. In the first analysis, items were scored as either correct or incorrect, independent of the degree of error. Not surprisingly, the exact age was more likely to be provided for Studied than Novel faces, *F*(1,65) = 44.99, *p* < 0.01. Additionally, the interaction between Study Condition and Study Status was significant, *F*(1,65) = 5.89, *p* = 0.02 (see Figure 7). For the Studied items, the Read condition outperformed the Guess condition, *t*(65) = 2.37, *p* = 0.01. For the Novel items, no significant difference was found between the Read and Guess conditions, though the direction of the effect favored the Guess condition *t*(65) = -0.71, *p* = 0.76.

Another analysis was conducted to include the Age of each face as a variable, in addition to Study Condition and Study Status (see Figure 8). The analysis revealed similar effects of

Study Condition and Study Status, as well as a main effect of Age, $F(3,195) = 33.45$, $p < 0.001$ and a significant interaction between Study Status and Age, $F(3,195) = 24.52$, $p < 0.001$. This analysis reveals markedly higher performance on the Studied faces from the first age group (18-29) above all other groups.

A final analysis was conducted to include the Race of each face as a variable, in addition to Study Condition and Study Status. No effects of Race were found.

Deviation of Estimates. The numerical nature of the responses in this task allows the deviations between responses and the correct answers to be investigated. Absolute deviation from the correct age was calculated for each face on a subject-by-subject basis. To see if subjects in the Guess condition improved in their age estimations from study to test, a paired t-test was conducted. Average deviation of guesses from the actual age during study (M = 8.30, sd = 1.62) was found to be significantly larger than the average deviation on Studied faces during test (M = 7.57, sd = 1.80), $t(33) = 3.24$, $p < 0.01$. This means that over the course of the experiment subjects became more precise in their age estimations.

When only those items were examined for which errors were made during the study phase, no significant differences were found between any of the treatment groups at final test (see Figure 9). This result conflicts somewhat with the prior one, but here it appears as though error generation does not render later age estimations more accurate, though it does make the age feedback more memorable.

Another analysis was conducted to include Age of each face as a variable. This analysis revealed an effect of Age, $F(3,195) = 51.50$, $p < 0.01$. Significant interactions were found between Study Status and Age, $F(3,195) = 3.36$, $p = 0.02$, and Study Condition and Age, $F(3,195) = 2.96$, $p = 0.03$. Figure 10 summarizes these age-related findings. Age estimations tend to be most precise for faces coming from the younger age groups, particularly if those faces were seen before. Additionally, there appears to be greater variability in the magnitude of the deviations across age groups for the Read condition as compared to the Guess condition.

A final analysis was conducted to include the Race of each face as a variable. Average deviation in age estimations for White faces was significantly lower than that of Other (Black, Asian, Hispanic) faces, $F(1,65) = 4.68$, $p = 0.03$. Though statistically significant, this difference was only an average of 0.43 years.

DISCUSSION

These analyses indicate an advantage for the Guess condition with regard to recognition memory. That is, faces studied in the Guess condition were more easily distinguished from new faces than faces studied in the Read condition. The nature of the Guess condition may have encouraged deeper processing of each face that may have led to better memory for the faces themselves. Furthermore, no differences between Read and Guess were found with regard to proportion of correct age estimations at final test. Likewise, no differences were found with regard to average deviations from correct ages. As in Experiments 1 and 2, Bayes factors for the effect of Study Condition on age estimation of Studied and Novel items were calculated. The evidence regarding differences in Study Condition for the Studied items was equivocal (B_{10} = 2.61). For the Novel items, there was some evidence supporting the null hypothesis (B_{10} = 0.31). Though these findings do not conform to the error generation benefit, they do suggest that errors were not harmful to memory for age-face associations and generalization to new faces. The results also support the finding that accuracy in age estimation can improve with training (Rhodes, 2009; McGillivray & Castel, 2010).

When age of each face was considered, the results appear to reflect an own-age bias, like that found in McGillivray and Castel (2010). That is, age estimations were more accurate for the youngest age group, in terms of both absolute accuracy and deviation. Unlike McGillivray and Castel (2010), no comparisons between age groups could be made because subjects in Experiment 3 only ranged from ages 18 to 25.

GENERAL DISCUSSION

Effects of error generation on learning and transfer were examined through use of a bird categorization task (Experiments 1 and 2) and an age estimation task (Experiment 3). Results from Experiments 1 and 2 indicate no differences in recognition memory between the two study conditions. However, results may have been affected by an abnormally large item-selection confound due to the large number of correct responses provided during the study phase. The stimuli set may have been too small and therefore not sufficiently difficult for this type of experiment. Initial analyses showed an improvement in categorization accuracy from study phase to final test along with an overall advantage for the Read condition. The absence of any error generation benefits may be due to limitations in the task itself. Previous studies demonstrated the disappearance of the error generation benefit when guesses were constrained (Grimaldi & Karpicke, 2012; Potts & Shanks, 2014). Bayesian analyses indicated superior performance for the Read condition on the memory test (Studied items), yet no clear difference across conditions for the transfer test (Novel items). Though generating errors may harm performance on a memory test, there seems to be little to no cost when it comes to transfer of knowledge.

Results from Experiment 3 revealed an error generation benefit for recognition memory, which is likely due to deeper processing at time of encoding. For memory for face-age associations, the Read condition seems to result in highest proportion of correct age estimations. However, it should be noted that subjects in the Guess condition did exhibit an improvement in age estimation accuracy from study phase to final test. Therefore, error generation still led to learning. As for generalization of age estimation ability to new faces, no differences were found between the two study conditions. Though error generation did not prevail, this result demonstrates that making errors during study did not impede future learning or transfer of knowledge.

The initial goal of these experiments was to evaluate whether the error generation benefit would extend to cases that require transfer of knowledge to new situations. Overall, the results from these experiments neither support nor condemn the potential learning benefits of error generation. Both errorless and errorful study approaches seem to lead to similar outcomes. These experiments serve as a first step to investigating the effects of error generation on learning and generalization. More research needs to be conducted to investigate when and how error generation affects memory and generalization abilities. Future research should also take into account metacognitive measures. It is important to consider how learners themselves perceive and handle the errors they make in order to alleviate concerns surrounding errors and to better inform educational applications.

TABLES AND FIGURES

Table 1. *Classification of bird stimuli as a function of studied/novel status, study condition, and even- or odd-numbering (Experiment 1)*

EVEN		Study Status			ODD		Study Status		
		Studied	Novel				Studied	Novel	
Study	Guess	0.536	0.448	**0.492**	Study	Guess	0.550	0.407	**0.478**
Condition	Read	0.616	0.521	**0.569**	Condition	Read	0.616	0.522	**0.569**
		0.576	**0.485**	**0.530**			**0.583**	**0.465**	**0.524**

Table 2. *Classification of bird stimuli as a function of studied/novel status and study condition (Experiment 2)*

		Study Status		
		Studied	Novel	
Study Condition	Read	0.384	0.280	**0.332**
	Choice	0.273	0.221	**0.247**
	Guess	0.270	0.268	**0.269**
		0.309	**0.256**	**0.283**

Table 3. *Proportion of exactly correct age estimations and average deviation from correct age as a function of age and race of each face during the study phase (Experiment 3)*

		Accuracy	Deviation
Face Age	18-29	0.081	6.326
	30-49	0.042	8.733
	50-69	0.034	9.076
	70-92	0.029	9.076
Face Race	White	0.051	7.721
	Other	0.042	8.888

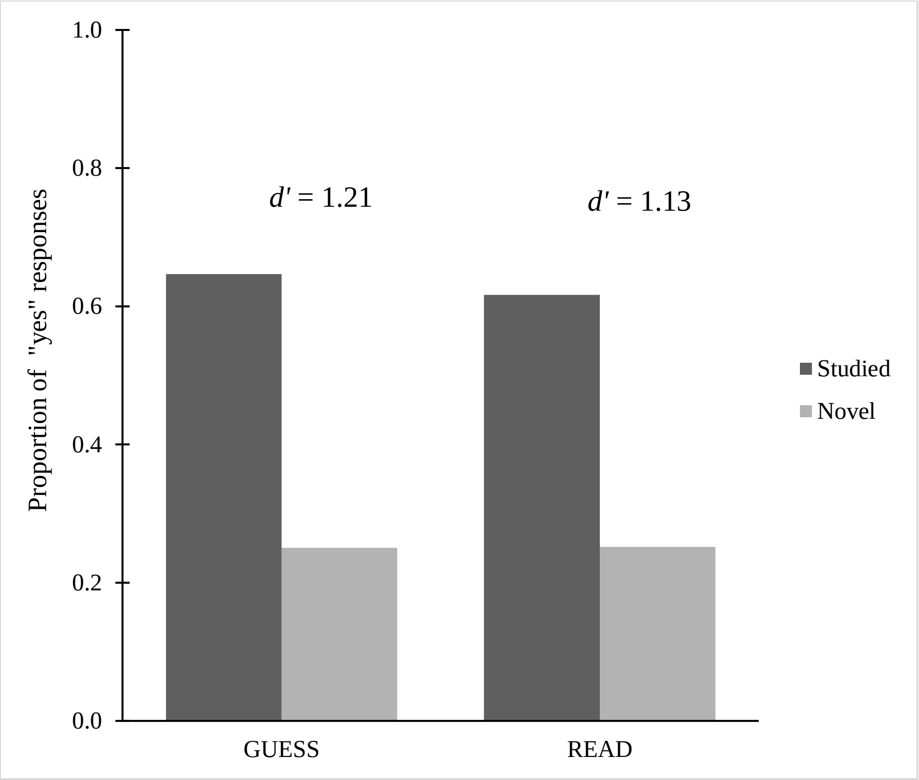

Figure 1. Recognition of bird stimuli as a function of studied/novel status and study condition (Experiment 1)

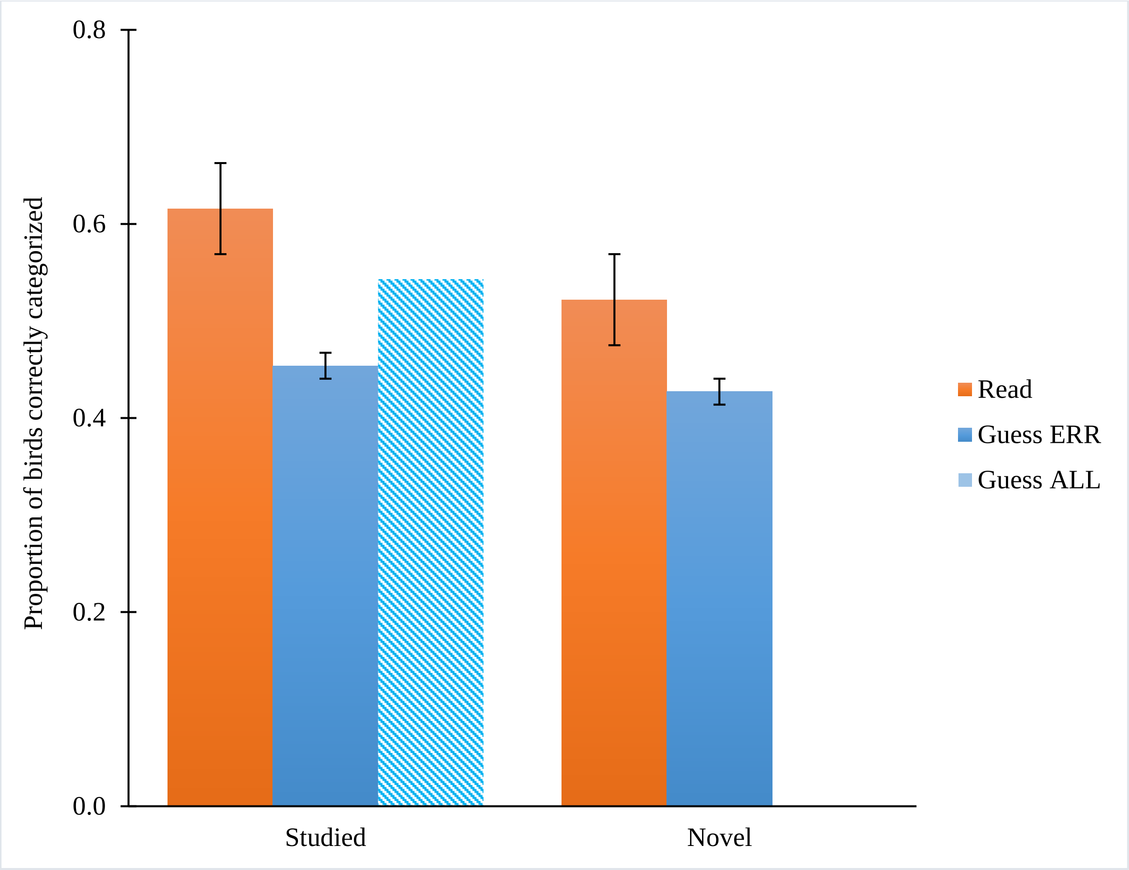

Figure 2. Classification of bird stimuli as a function of studied/novel status and study condition (Experiment 1)

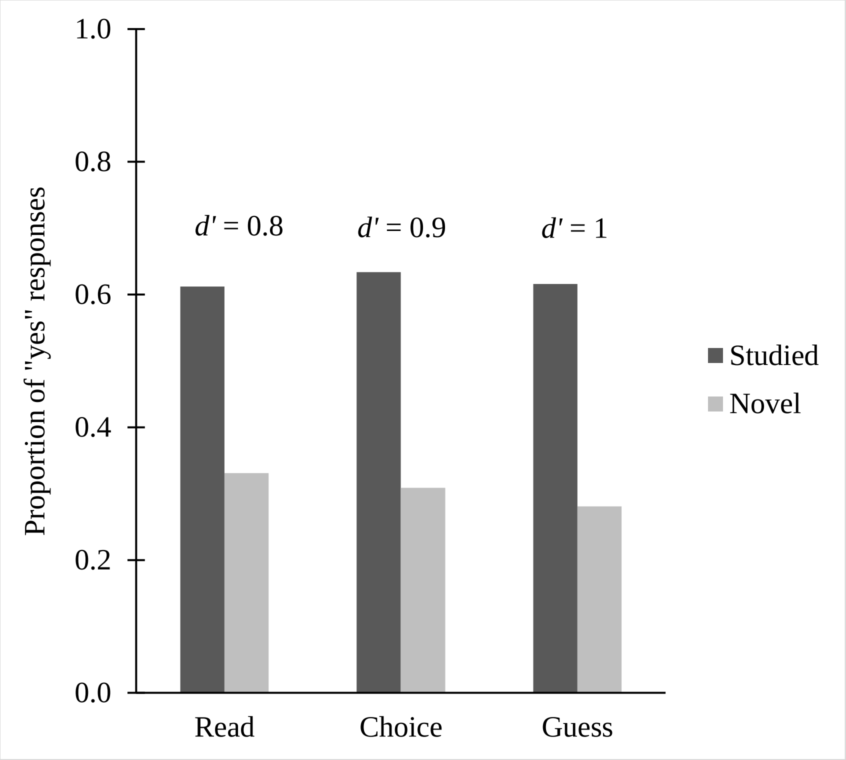

Figure 3. Recognition of bird stimuli as a function of studied/novel status and study condition (Experiment 2)

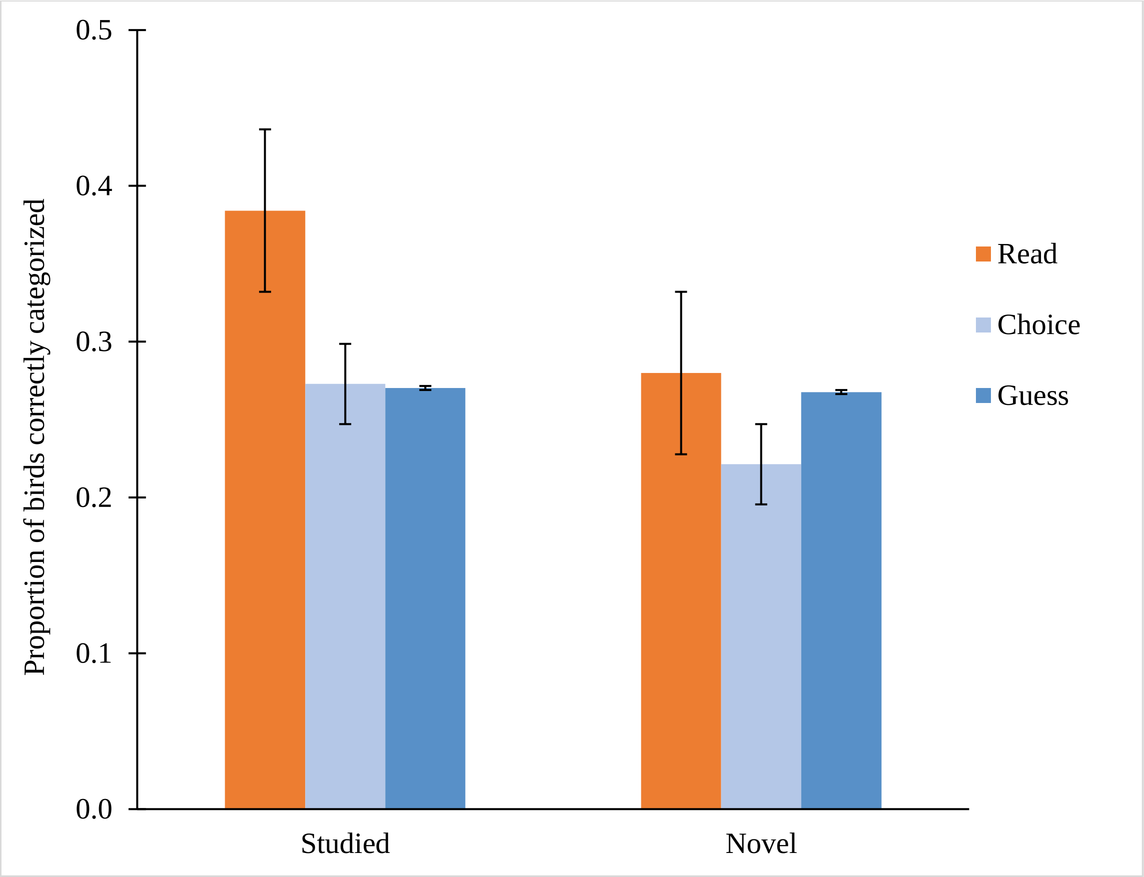

Figure 4. Classification of bird stimuli as a function of studied/novel status and study condition (Experiment 2)

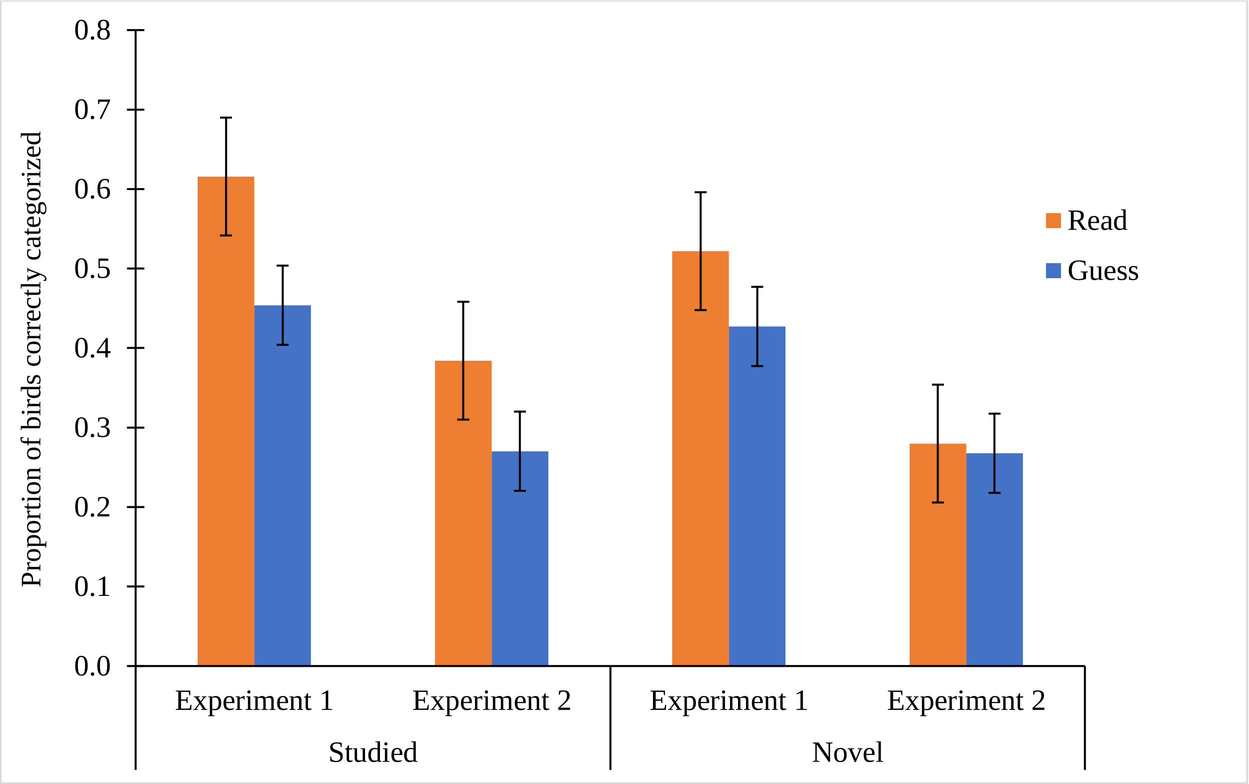

Figure 5. Classification of bird stimuli as a function of studied/novel status and study condition (Experiments 1 and 2)

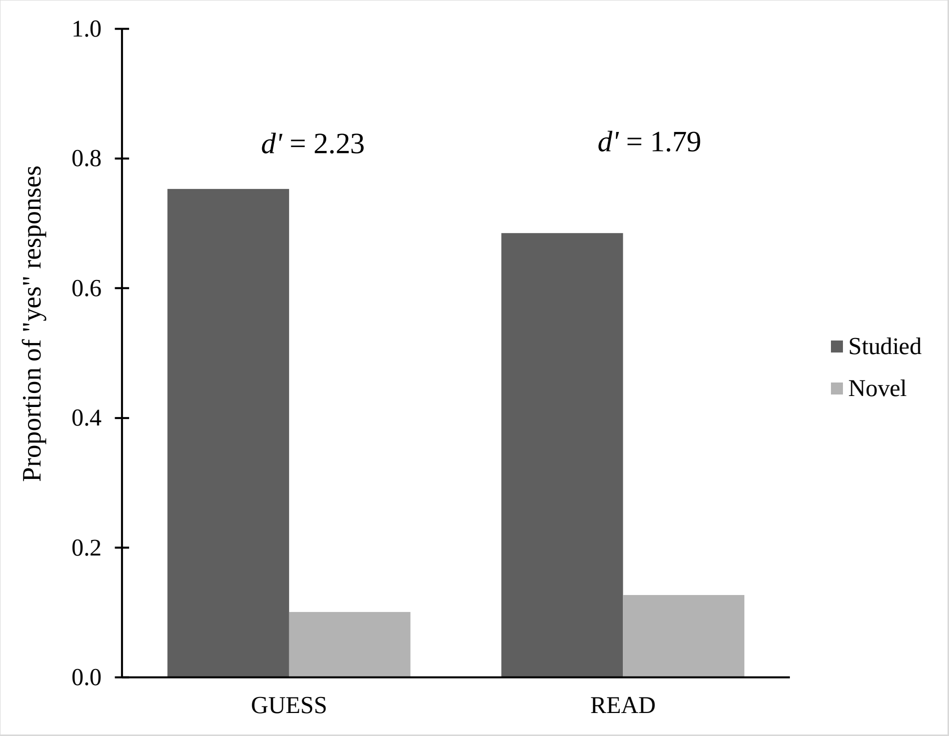

Figure 6. Recognition of face stimuli as a function of studied/novel status and study condition (Experiment 3)

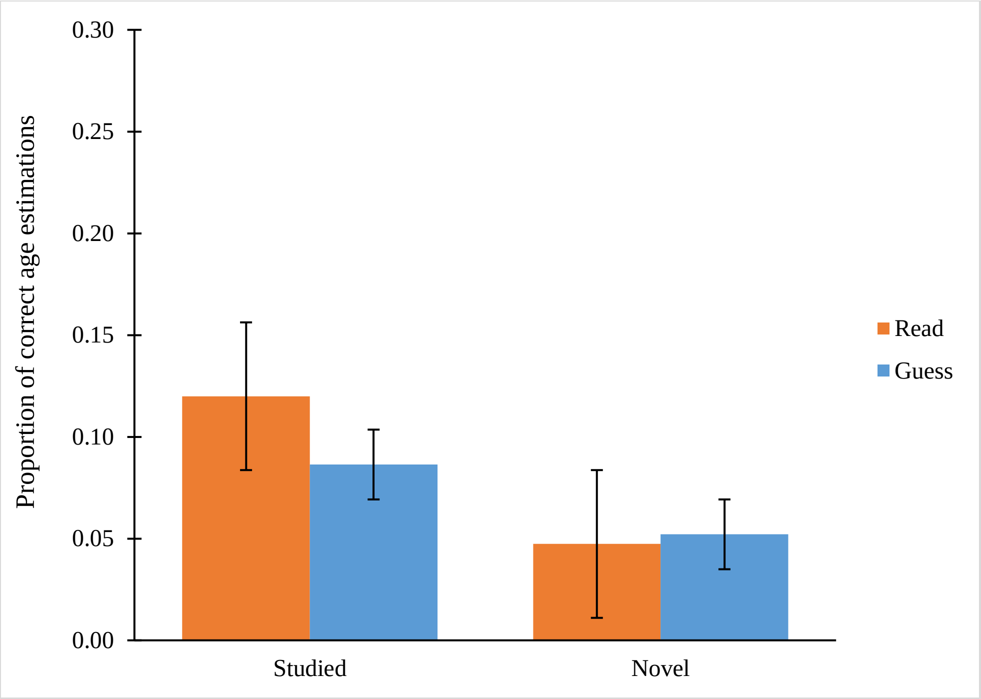

Figure 7. Proportion of exactly correct age estimations as a function of studied/novel status and study condition (Experiment 3)

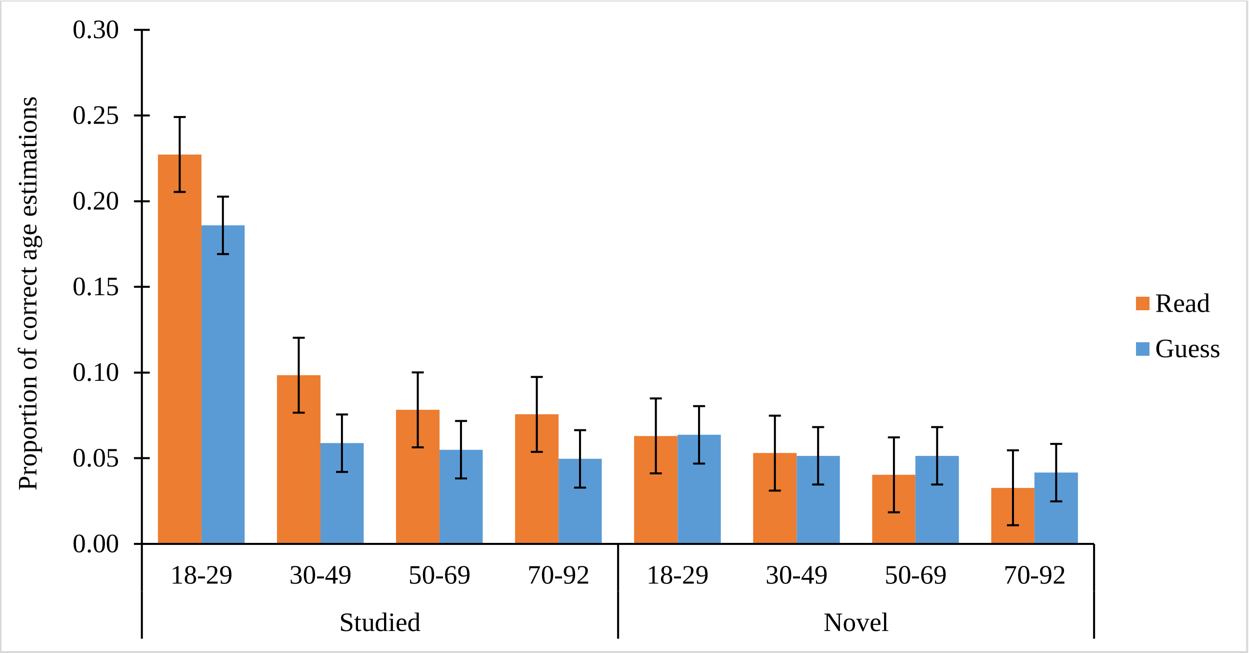

Figure 8. Proportion of exactly correct age estimations as a function of studied/novel status, study condition, and age of each face (Experiment 3)

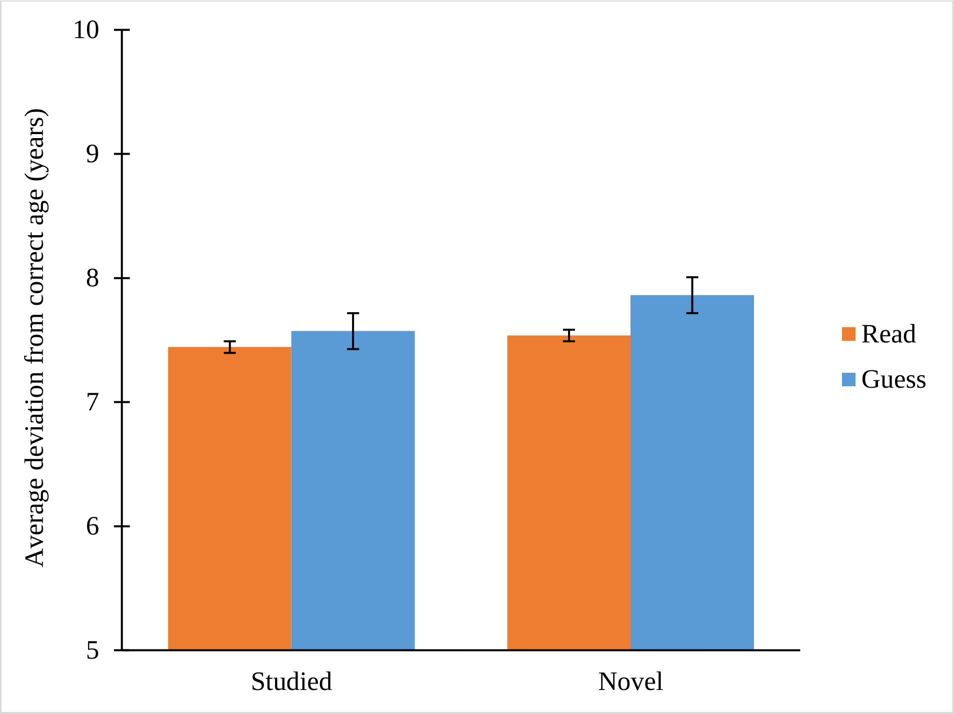

Figure 9. Average deviation from correct age as a function of studied/novel status and study condition (Experiment 3)

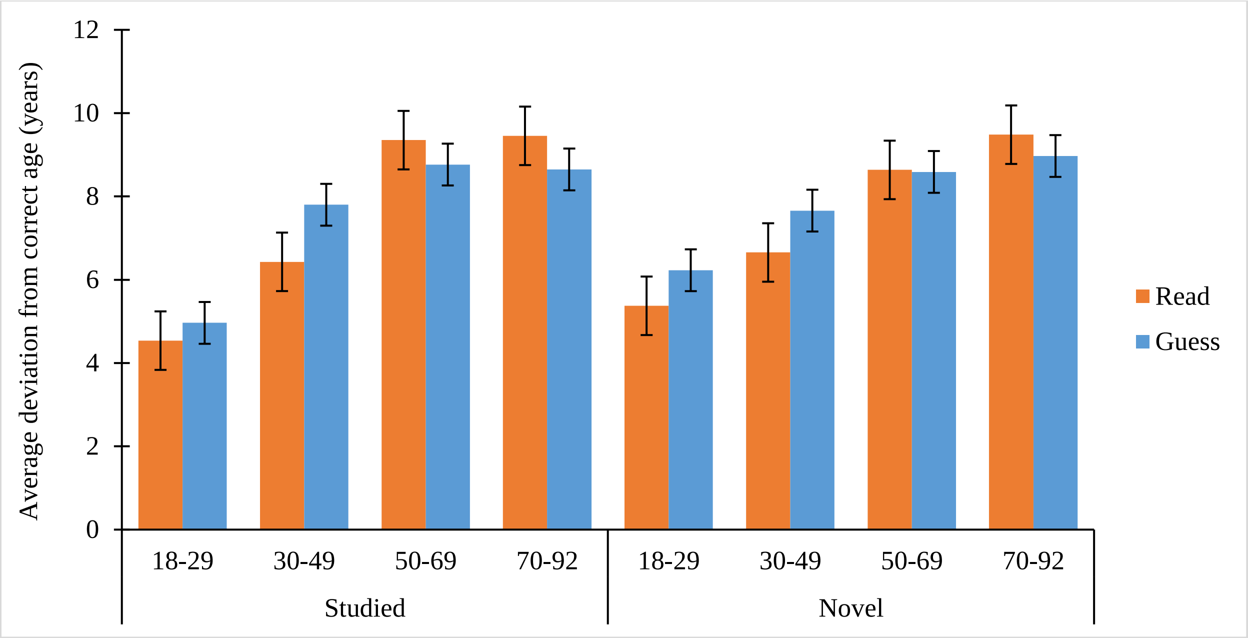

Figure 10. Average deviation from correct age as a function of studied/novel status, study condition, and age of each face (Experiment 3)

REFERENCES

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: A perspective from cognitive psychology. *Policy Insights From the Behavioral and Brain Sciences*, *2*(1), 13-23.

Berlyne, D. E., & Normore, L. F. (1972). Effects of prior uncertainty on incidental free recall. *Journal of Experimental Psychology*, 96(1), 43.

Butler, D. C., & Peterson, D. E. (1965). Learning during "extinction" with paired associates. *Journal of Verbal Learning and Verbal Behavior*, *4*(2), 103-106.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1491.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69-84.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current directions in psychological science*, 21(5), 279-283.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633-642.

Clare, L., & Jones, R. S. (2008). Errorless learning in the rehabilitation of memory impairment: a critical review. *Neuropsychology review*, *18*(1), 1-23.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, 40(4), 505-513.

Guthrie, E. R. (1952). The psychology of learning (rev.)

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(1), 290.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & cognition*, 40(4), 514-527.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, 83(2p1), 340.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1441.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press, Clarendon Press.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology*, 70(4), 626.

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(4), 731-746.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 989.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*,*14*(2), 194-199.

Metcalfe, J. (2017). Learning from errors. *Annual review of psychology*, *68*, 465-489.

McGillivray, S., & Castel, A. D. (2010). Memory for age–face associations in younger and older adults: The role of generation and schematic support. *Psychology and aging*, 25(4), 822.

Minear, M. & Park, D.C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*. 36, 630-633.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. *National Center for Education Research*.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, 143(2), 644.

Rhodes, M. G. (2009). Age estimation of faces: A review. *Applied Cognitive Psychology*, 23(1), 1-12.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243.

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155.

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning taking memory tests improves long-term retention. *Psychological science*, 17(3), 249-255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181-210.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225-237.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 153-163.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, 4(6), 592.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic bulletin & review*, *19*(5), 899-905.

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & cognition*, 42(8), 1373-1383.