Does Testing with Feedback Improve Adult Spelling Skills Relative to Copying and Reading?

Steven C. Pan

Benjamin R. Rubin

Timothy C. Rickard

University of California, San Diego

This manuscript was accepted for publication in the *Journal of Experimental Psychology: Applied* on August 12, 2015. This document may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final version is available at: http://dx.doi.org/10.1037/xap0000062

Word Count: 12,211 (main text and references)

Author Note

Steven C. Pan, Department of Psychology, University of California, San Diego; Benjamin R. Rubin, Department of Psychology, University of California, San Diego; Timothy C. Rickard, Department of Psychology, University of California, San Diego.

S. C. Pan is supported by a National Science Foundation (NSF) Graduate Research Fellowship.

The authors thank Angela Jones for sharing insights on spelling and reviewing an earlier version of this manuscript, Kylie Hsu for insights on foreign language instruction, Larry Jacoby and Henry Roediger for suggesting the comparison of testing to reading in Experiment 4, and to many others, including anonymous reviewers, who commented on this work. Thanks also to Christine Diao, Crystal Pang, Drew Walker, Megan Iida, Milo Chan, and Tania Romero for their assistance with setting up and running the experiments, as well as Arpita Gopal, Bijan Malaklou, Darian Parsey, Jason Chang, Johnny Barry, Jonathan Mejia, and Michelle Hickman for assistance with data coding.

Please address correspondence to: Timothy C. Rickard, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109. Email: trickard@ucsd.edu

Abstract

We examined testing's ability to enhance adult spelling acquisition, relative to copying and reading. Across three experiments in which testing with feedback was compared with copying, the spelling improvement after testing matched that following the same amount of time spent copying. A potent testing advantage, however, was observed for spelling words free-recalled. In the fourth experiment, a large testing advantage for both word free recall and spelling was observed, versus reading. Subjects also generally preferred testing and rated it as more effective than copying or reading. The equivalent performance of testing and copying for spelling contrasts with prior work involving children and suggests that retrieval practice may not be the only effective mechanism for spelling skill acquisition. Rather, we suggest that the critical learning event for spelling is focused study on phoneme-to-grapheme mappings for previously unlearned letter sequences. For adults with extensive spelling expertise, focused study is more automatic during both copying and testing with feedback than for individuals with beginning spelling skills. Reading, however, would not be expected to produce efficient focused study of phoneme-to-grapheme mappings, regardless of expertise level. Overall, adult spelling skill acquisition benefits both from testing and copying, and substantially less from reading.

Keywords: spelling, orthography, testing effect, retrieval practice, reading, copying

Does Testing with Feedback Improve Adult Spelling Skills Relative to Copying and Reading?

The effectiveness of testing in promoting learning, relative to control tasks such as restudy, has been established across a wide variety of task domains including paired-associate word learning, fact learning, and passage comprehension. Such *testing effects* or *retrieval practice effects* are not just confined to the laboratory; in recent years, test-enhanced learning has been demonstrated in real-world classroom settings and with grade school and college students alike (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel, Anderson, Derbish, & Morisette, 2007; McDaniel, Roediger, & McDermott, 2007; Roediger, Agarwal, McDaniel, & McDermott, 2011). Accordingly, the use of testing as an instructional tool ranks prominently among research-backed recommendations to improve learning in classrooms and other educational settings (e.g., Pashler et al., 2007; Roediger & Pyc, 2012).

Important directions for ongoing research on the testing effect include exploration of its efficacy across the full range of educationally relevant materials and the identification of principles that describe the boundary conditions of test-enhanced learning, where such boundaries may exist (Rowland, 2014). One particular task domain that remains largely unexplored in the testing literature is that of spelling, and especially among adults. Given numerous successful demonstrations of testing's benefits for different types of verbal materials (for reviews see Roediger & Butler, 2011; Roediger & Karpicke, 2006a), the majority involving adult subjects, it would be reasonable to expect that similar benefits will accrue for spelling skills in the same population.

Spelling Research and Adults

The search for optimal techniques to teach spelling has persisted for centuries and has

attracted interest from prominent individuals ranging from Noah Webster to Theodore Roosevelt (Venetsky, 1980). However, aside from broad guidelines about which instructional approaches (e.g., weekly training schedules) and general techniques (e.g., whole word presentation) are preferable (e.g., Horn, 1967), there still remains little consensus about the relative effectiveness of widely-used strategies such as testing and studying (Cronnell & Humes, 1980; Treiman & Cassar, 1997), and a dearth of empirical research with adults (Ormrod, 1986). Moreover, rarely have two or more spelling instructional methods been directly compared under experimental conditions that precisely control for time on task, and no adult studies to date have compared the effectiveness of the three techniques explored here: testing with feedback, copying, and reading.

In this manuscript, *testing with feedback* is operationally defined as attempting to write out a spelling word after aural presentation of that word, followed by visual presentation of the correctly spelled answer. *Copying* is defined as written transcription of a visually and aurally presented spelling word. *Reading* is defined as simultaneous viewing and vocal pronunciation of a visually and aurally presented spelling word. These techniques (including variants thereof) are among the most commonly used for spelling acquisition across a wide range of age levels (for the cases of copying and reading, see Cronnell & Humes, 1980; Ormrod & Jenkins, 1989).

Studies of adult spelling are an important complement to those involving children, for multiple reasons. First, despite the popular belief that adult spelling instruction is unnecessary, among college students (and hence the more general adult population) spelling performance is often sub-par. In fact, the deficient spelling skills of undergraduates have long been recognized as a persistent problem (Alper, 1942; Guiler, 1931). Second, adults can benefit from specialized spelling instruction, such as through remedial programs and deliberate spelling instruction (Ormrod, 1986); without intervention, poor adult spelling skills persist (Hartmann, 1931). Third,

new words and their spellings must often be learned in technical and specialty domains throughout college, graduate training and careers. Fourth, adult foreign language learning involves spelling acquisition for new words that may or may not have similarities to learners' native languages in their phoneme-to-grapheme mappings. Finally, the study of adult spelling acquisition under controlled laboratory conditions may facilitate new insights into the underlying learning processes and provide important reference data facilitating the development of a theory of spelling acquisition across the lifespan.

Spelling Research and Children

Given the paucity of spelling studies with adults, our understanding of effective spelling instructional techniques relies on studies of developing populations, namely children. Accumulating evidence from research with young children indicates that testing with correct answer feedback, also known in the spelling literature as self-correction, can be highly effective at improving spelling proficiency. In such studies, testing with feedback appears to outperform restudy (Murphy, Hern, Williams, & McLaughlin, 1990) and syllable-by-syllable word analysis (McNeish, Heron, & Okyere, 1992). Testing may also promote superior spelling acquisition than repeated copying of spelling words (Grskovic & Belfiore, 1996; McGuffin, Martz, & Heron, 1997; Wirtz, Gardner, Weber, & Bullara, 1996). However, these studies have almost exclusively involved special populations (e.g., learning disabled students) and single-digit sample sizes, and thus have limited generalizability.

To address these limitations, we recently demonstrated (Jones et al., 2015) that testing with feedback produced substantially more learning among normally developing first- and second-grade students than did a copying technique known as rainbow writing, in which children repeatedly write spelling words in different colors in an effort to maintain engagement. In

Experiment 3 of that work, which was the most sensitive to learning, testing yielded an estimated 377% more learning than did copying. Our results provide strong evidence that the benefits of retrieval practice extend into the domain of spelling. Given this and earlier findings, it might be reasonable to expect that a similarly large advantage for testing, relative to copying, will exist for the case of spelling acquisition in adults.

Reading, which ranks among the most commonly used forms of spelling practice, has mixed support for its effectiveness as a spelling instructional technique. With children, reading of both real words and pseudowords has been found to produce spelling improvements (Ehri, 1997; Share, 2004). Conversely, reading has also been shown to be less effective for spelling acquisition in children than testing without feedback (Shahar-Yames & Share, 2008), testing with correct answer feedback (Conrad, 2008), and repeated copying (Bosman & de Groot, 1992; Bosman & Van Orden, 1997). Overall, it appears that reading may benefit spelling skills in children under some circumstances, but it remains unclear whether reading will be competitive with testing for spelling more generally, and in particular with adults.

Test-Enhanced Learning and Reading

Besides its frequent use for spelling instruction, reading also ranks as one of the most common control conditions in the testing effect literature (Roediger & Karpicke, 2006a). In most demonstrations of test-enhanced learning, reading produces markedly less retention of to-be-learned materials. One prominent theoretical explanation for this difference is the *retrieval practice* account of the testing effect, whereby the act of recalling information during a test modifies and strengthens its memory representation (Bjork, 1975; Roediger & Karpicke, 2006a). This phenomenon is assumed to be absent during reading of the same materials. The prominence of reading in testing effect studies provides additional motivation for including it as a training

task, as a comparison of testing and reading provides a natural connection point between the spelling and testing literatures.

The Current Study

In the present study, we conducted four experiments to assess testing's effectiveness for adult spelling acquisition, relative to copying and reading. Experiments 1 through 3 explore the effectiveness of testing with feedback compared to copying, while Experiment 4 evaluates testing with feedback compared to reading. In all experiments, subjects trained on 40 difficult spelling words while alternating between techniques. After a one-week delay, subjects returned for a test session which involved a free recall test (in which subjects were to recall as many spelling words as possible and attempt to spell them correctly) followed by a cued recall test (in which words were presented aurally and subjects attempted to spell them correctly). The primary dependent measure for spelling performance was the cued recall test, which allows spelling accuracy to be assessed on all 40 words for all subjects. The free recall test supported exploration of two ancillary questions. First, will any testing effect that may be observed for spelling on the cued recall test also be observed on the free recall test? Second, does testing enhance phonetically identifiable (regardless of spelling accuracy) free recall of the spelling words?

Experiment 1

Subjects attempted to learn 40 spelling words during a classroom training session, half through copying and the other half through spelling tests with correct answer feedback (T+FB). To maximize the applied implications of the work, the amount of time on task in the two training conditions was equated. After a one-week delay, subjects returned to the classroom for a second session in which a free recall and cued recall test was administered.

Method

Subjects. Thirty-eight University of California San Diego undergraduate students participated for course credit. All subjects completed both sessions of the experiment.

Materials. From lists of frequently used spelling bee words (Scripps National Spelling Bee, 2004) and commonly misspelled words (Heckendorn, 2014), we obtained 110 candidate words. A pilot spelling test involving seven volunteer students identified highly recognizable yet frequently misspelled words. Using this data, we selected 40 nouns of between 7 and 13 letters in length. Those 40 words were randomly divided (subject to the constraints here) into two word lists (Lists A and B; see Appendix A), each containing 20 words with an average length of 10 letters and an average frequency of 4-5 per million (Wilson, 1988). Both lists had a range of 12 different first letters. Each list was further halved (creating sub-lists A1, A2, B1, and B2) to be used in 10-word training blocks during the training session.

The fact that the words used in this study are frequently misspelled, yet highly recognizable was confirmed by training and exit survey data from the 78 subjects that participated in Experiments 1 and 2. Among all 40 words, the average misspelling rate on an initial test was 67% and the average recognition rate was 74% (i.e., “mostly know” or “fully know” responses to the question, “how well do you know how to use the following word in a sentence?”). Because subjects rated individual words following visual and not aural presentation, those results likely underestimate the degree to which the words were recognizable (as subjects may be more familiar with the phonology than orthography of difficult spelling words). Thus, most subjects had working knowledge of the words used in this study, but in most cases not their exact spellings, as was intended.

Design and procedure. Training condition was manipulated within-subjects. Subjects

were run sequentially in groups of four, with each group assigned to one of four counterbalance groups formed by orthogonal manipulation of two factors: (1) training condition presentation order (copying, T+FB, copying, etc. or T+FB, copying, T+FB, etc.), and (2) list assignment (List A trained under T+FB and List B under copying vs. the reversal of that assignment). In each session, subjects were seated in a classroom facing a projector screen on which a slideshow created using PowerPoint (Microsoft, Redmond, WA) was used to present instructions and words.

Training session. Subjects were provided with a ballpoint pen and two identical double-sided worksheets which were evenly divided into four quadrants of 10 blank lines each. An introductory slide indicated that they were participating in an assessment of university-level spelling skills for which "very difficult" spelling words would be shown. Subjects were specifically made aware that their attention should be focused on the spelling properties of each word that they encountered, in line with prior work which indicates that such instructions facilitate better learning (Ormrod, 1986). For coding purposes, subjects were instructed to print neatly; if errors were made while copying, entire words were to be crossed out (rather than individual letters) and the intended answer rewritten in adjacent space on the same line. Moreover, throughout the experiment subjects were instructed to print their responses in manuscript handwriting and not use cursive.

The training block design for an example counterbalance group is detailed on the left side of Figure 1. Subjects practiced all 40 words in eight four-minute blocks of 10 words each. In each block, words were presented one at a time for 12 seconds each (for a total of two min) using consistent audio and mixed visual presentation. All 10 words were then cycled through again for 12 s each (either for additional copying or feedback, depending on training condition), bringing

the total trial time for each block to four min. Word order was randomized within each block. Audio presentation for both conditions consisted of an .mp3 clip of the word being played through the classroom speakers once every four s (for a total of three repetitions per 12 s). Visual presentation (copying condition or feedback sections only) consisted of each word appearing in large, bold-face serif font (Times New Roman, size 60) at the center of the screen.

Copying. During each trial of a copying block, the presented word was shown and repeatedly copied, and then all 10 words were presented a second time (and copied again). Each word was presented both aurally through the classroom speakers and visually on the slideshow screen, as detailed above. Subjects were instructed to copy the correct spelling of each word as many times as possible using one line per copy. If subjects were midway through copying a word when the next word was shown, they were instructed to immediately move to the next line and begin copying the next word. Three written repetitions of the presented word were typically accomplished within the 12 s timeframe of each trial, with subjects actively engaged in copying throughout.

Testing with feedback. Each T+FB block involved two min of testing and two min of feedback (block time was evenly divided by necessity to accommodate the two-step nature of the T+FB task). During testing, individual words were presented aurally through the classroom speakers, just as for the copying condition, but without any visual presentation, at a rate of 12 s per word. Subjects were instructed to attempt to spell each word once, using one line per word, and to be as accurate as possible. After all 10 words were tested, FB involved the visual presentation of individual correctly-spelled words on the screen, accompanied by audio presentation, at a rate of 12 s per word. Subjects were instructed to check the spelling of each word that they had written, letter by letter, and to mark each correctly spelled word with a

checkmark and each incorrectly spelled word with an "X" mark. The provision of feedback only after a series of test trials aligns with prior research showing that delayed feedback leads to the most robust testing effects (e.g., Roediger & Butler, 2011).

Eight training blocks alternated between copying and T+FB (see Figure 1). All 40 words from Lists A and B were shown during the first four blocks, and then shown again during the remaining four blocks. Thus, there was not just an initial testing and feedback opportunity for all tested words, but also a second testing and feedback opportunity for all tested words in the latter half of the training session. To eliminate visual cues of previously written words, subjects were instructed to turn their worksheets over and use the reverse side after every two blocks. Worksheets were collected after every four blocks. At the end of training, subjects were told to return one week later at the same time of day for the test session. To minimize practicing between sessions, subjects were falsely informed that the second session would entail learning a new set of vocabulary words.

Test session. Subjects were provided with a ballpoint pen and a single double-sided worksheet with 40 blank lines on each side, numbered 1 through 40. An introductory slide reminded subjects that they had learned to spell 40 words in the previous session, and that the actual purpose of the second session was to assess their memory of those words. This was followed by the free recall test, in which subjects had eight min to write as many of the words that they had learned as possible (from either condition and in any order) on the worksheet using one line per word. Subjects were told to spell as accurately as possible. Because no feedback of any type was provided for the free recall task, performance of that task should have negligible influence on subsequent cued recall performance (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). Afterwards, subjects turned their worksheets over for the cued recall test, during which

all 40 words were presented aurally in random order one at a time through the classroom speakers. Subjects were given 12 s to spell each word, during which its respective audio clip was played three times.

Following the cued recall test, subjects completed a questionnaire in which they answered demographic and opinion questions. This exit survey included a section on past educational experiences with learning spelling, a section on familiarity ratings for each of the spelling words used, and a metacognitive section assessing preference for the techniques used in the experiment (forced-choice of either T+FB or copying). The metacognitive questionnaire for this entire study (which includes questions added after Experiment 1) is detailed in Appendix B.

Coding. The worksheets from the training and test sessions were transcribed into Excel (Microsoft, Redmond, WA) spreadsheets. Two independent coders transcribed each worksheet. A computer algorithm was used to identify any discrepancies between the two transcriptions per worksheet, and all discrepancies were adjudicated against the original worksheets by a third independent coder. Intercoder disagreement invariably involved differing interpretations of individual written letters that did not change incorrectly spelled responses to correct responses. To determine spelling accuracy, a matching algorithm was run on finalized transcripts to compare coded spellings against a master list of correctly spelled words.

Transcribed data from the training session included subject responses to the testing and feedback phases of each T+FB block, and a count of the number of repetitions per word in each copying block (where words were rarely misspelled due to being copied directly from the screen). For the free recall portion of the test session, each written word was transcribed exactly as written and then separately coded as being a phonetically identifiable match or not to one of the 40 words. All 40 words from the cued recall portion of the test session were also transcribed

exactly as written.

Results and Discussion

Training. For tested items, the mean proportion of words spelled correctly increased significantly from 0.33 on the first training block to 0.63 on the second (see Figure 2, left-side bar chart), $t(37) = 14.5$, $p < 0.0001$, $d = 2.36$. Subjects correctly processed the feedback on nearly all T+FB trials; the proportion of spelling attempts accurately scored during feedback was 0.97. Subjects completed an average of 3.21 correctly spelled repetitions per word on all copying trials.

Delayed test. Performance on the final tests administered one week after training is shown for both the testing and copying conditions on the left side of Figures 3, 4, and 5, respectively, for the measures of phonetically identifiable words free-recalled (henceforth, *words free-recalled*), proportion of free-recalled words spelled correctly, and proportion of words spelled correctly on the subsequent cued recall test.

For proportion of words free-recalled, there was a highly significant 72% performance advantage for the testing condition over copying, $t(37) = 7.75$, $p < 0.0001$, $d = 1.26$. That result is in agreement with the testing advantage that is typically observed in other domains of verbal learning. In this case, however, the testing effect is incidental; in the training phase, words were presented aurally (and in the copying condition also visually) as spelling cues and thus no long-term memory recall for words (i.e., no testing of word recall) was required.

Remarkably, however, there was no evidence for a testing effect for the target skill of spelling. Among the 34 subjects who free-recalled at least one word in both the testing and copying conditions, spelling accuracy did not depend on training condition (see left side of Figure 4). Similarly, for cued recall, which is likely the more sensitive measure given that all 40

words were assessed for each subject, the mean proportion correct was identical (0.58) for the two conditions (Figure 5). As is evident through comparison of Figures 2 and 5, about 80% of the learning that occurred during training was retained on the one-week delayed cued recall test. Given the equivalence of copying and T+FB on that test, we can infer that a similar percentage of the learning in the copying condition was retained after the one-week delay.

Questionnaire results. The exit survey confirmed that copying and T+FB are commonly used to learn spelling. Among respondents, 89% reported previously having used copying to learn spelling and 74% reported previously having used testing with feedback to learn spelling. For the forced-choice preference question on the two techniques, 68% of respondents selected copying. We will further discuss the questionnaire results of this and subsequent experiments in the General Discussion.

Experiment 2

The lack of a testing effect for spelling in Experiment 1 contrasts with both the robust testing effect for words and other linguistic materials in the literature and the clear advantage for testing over copying for children's spelling (Jones et al., 2015). One possible explanation for this result is that testing in Experiment 1 was implemented inefficiently relative to copying. We observed that, whereas the copying task kept most subjects engaged throughout each 12 s copying trial, subjects typically completed their spelling attempt for each testing trial within about half of that time. Further, it was apparent through observation that most subjects did not require 12 s to process feedback for each word after testing. Thus, in Experiment 2 we sought to improve the efficiency of the testing condition by reducing the duration of each T+FB trial from 24 s to 12 s (thus, 6 s for testing and 6 s for feedback per trial) and by having two rather than one T+FB trial(s) for each word within each block. The copying task appears to have been

implemented efficiently (i.e., subjects were fully engaged in copying throughout each trial) and thus was left unchanged.

Method

Subjects. Forty-six University of California San Diego undergraduate students participated for course credit. All but six subjects completed both sessions of the experiment. Data from two additional subjects that did not follow instructions was not analyzed.

Materials, design, procedure, and coding. The materials, design, and procedures of the training session were identical to those in Experiment 1, except for modifications to the T+FB training condition, where we doubled the number of T+FB training trials per word while retaining an equal amount of time on task for both training conditions. Thus, subjects were tested and received feedback for each word in the T+FB condition four times over the entire training session, each spaced over four-minute intervals during which the copying task was performed. The resulting training block design is detailed in center of Figure 1. In the T+FB condition, subjects turned their worksheets over after each T+FB cycle, eliminating visual cues from prior cycles through the same 10 words.

The only other design modification in the training session was a reduction in the number of audio repetitions per presentation in the T+FB training condition, from once every four s to once every six s, which was necessary to equate the copying and testing conditions with respect to the number of aural word presentations per trial. The copying training condition and the free and cued recall tasks of the test session were identical to those of Experiment 1.

The only change to the test session was to the metacognitive portion of the exit survey, in which the assessment of subjects' preference for the techniques used was changed from a forced-choice question to a seven-point numerical scale of relative preference between T+FB and

copying (shown in Appendix B). Subjects were also asked to rate the effectiveness of T+FB and copying for learning spelling on a five-point scale (from “not effective” to “highly effective”).

Results and Discussion

Training. The mean proportion of words spelled correctly increased with each block (from first to last: 0.34, 0.71, 0.76, 0.86). As expected, doubling the number of T+FB cycles resulted in better spelling improvement over blocks than was observed in Experiment 1 (see Figure 2). A two-sample *t*-test on the difference between first and last block performance in the two experiments was highly significant, $t(74) = 5.76$, $p < 0.0001$, $d = 1.32$. As in Experiment 1, subjects correctly processed the feedback on nearly all T+FB trials; the proportion of spelling attempts accurately scored during feedback was 0.96. Subjects completed an average of 2.90 correctly spelled repetitions per word on all copying trials.

Delayed test. The proportion of words free-recalled exhibited the same large incidental testing effect that was observed in Experiment 1, $t(37) = 6.79$, $p < 0.0001$, $d = 1.10$. With respect to spelling, however, there was again no positive testing effect**.** For proportion of free-recalled words spelled correctly (limited to the 36 subjects who recalled at least one word in both conditions), there was instead a significant advantage for the copying condition (Figure 4), $t(35) = 3.47$, $p = 0.0014$, $d = 0.58$. However, that result was not observed in Experiment 1 and does not replicate in Experiment 3. Further, the *p-value* of the test should be interpreted with caution. In this experiment in particular, there were multiple subjects who free recalled only one word in the copying condition. Proportion correctly spelled for those subjects can only take the extreme values of 0 or 1.0 (and mostly 1.0 in this experiment), exerting disproportional influence on the mean proportion correct over subjects.

Of most importance, there was again no testing effect on the cued recall spelling test,

(Figure 5), $t(37) = 0.35$, *ns*. That result indicates that the improved spelling performance in the testing condition by the end of training (proportion correct of 0.86 vs. 0.63 for Experiment 1), achieved by increasing the number of training test trials, did not translate into superior spelling performance in the T+FB condition on the delayed test. In other words, testing remained equally effective as copying, even with double the amount of testing during training.

What accounts for this result being obtained again in Experiment 2? One possibility is that the words that were spelled correctly for the first time on the third and fourth training blocks in Experiment 2 tended to be more difficult and more easily forgotten over the one-week delay. This may have largely negated any advantage conferred by doubling the amount of training tests.

A second possibility is that the actual difference in achieved spelling skill at the end of training in Experiments 1 vs. 2 is smaller than indicated by differences in final training block performance. That possibility is based on the joint assumptions that (1) true achieved spelling skill at the end of training in Experiment 1 was higher than training data indicate, as additional learning which occurred on the last testing with feedback cycle was not immediately assessed, and that (2) the benefit of additional training test trials in Experiment 2 may have been relatively small, given that rates of accuracy improvement are generally known to decrease over successive training blocks. Thus, if doubling the amount of testing produced only slightly better performance by the end of training in Experiment 2, it follows that delayed test performance would not differ much from the results of Experiment 1.

A third possibility is that doubling the amount of test trials in Experiment 2 increased the learning rate not only in the T+FB condition but also in the copying condition. Specifically, more frequent testing could in principle have caused subjects to be more motivated to learn or more attentive in not only the T+FB condition but also the copying condition. It should be

noted, however, that motivational carry-over is a consideration not just for the experiments in this study, but also for all within-subjects testing effect experiments in general. We will return to this possibility in the General Discussion.

A final possibility is that differences in the amount of motor learning may be responsible for the lack of a testing effect relative to copying. Specifically, there was more practice at handwriting of words in the copying condition. We address that possibility in the next experiment, which is described below.

Questionnaire results. Similar to the results of Experiment 1, 80% of respondents reported previously having used copying to learn spelling and 78% reported previously having used testing with feedback to learn spelling. When asked to rate their relative preference for either of the two techniques used in the experiment, 61% of respondents gave ratings that were in the direction of testing. The relative preference for testing was statistically significant, $t(37) = 2.41$, $p = 0.021$, $d = 0.39$. Respondents did not significantly differ in their effectiveness ratings for either technique, $t(37) = 1.86$, $p = 0.070$.

Experiment 3

As mentioned above, one candidate account of the absent testing effect for spelling in the first two experiments is that motor learning yielded enhanced final test spelling performance in the copying condition relative to the T+FB condition. The mode of responding during both training and the final test involved handwriting, and there was more handwriting in the copying condition than in the T+FB condition. It may be that the copying task results in enhanced handwriting motor memory for the correct letter sequence that promotes correct spelling on the final test, potentially masking a testing advantage for spelling that may be present at non-motor levels of representation (for prior work that is consistent with an influence of handwriting motor

memory on spelling see Cunningham & Stanovich, 1990; Longcamp et al., 2008; Longcamp, Zerbato-Poudou, & Velay, 2005). We explored that possibility in this experiment by shifting to computer-based stimulus presentation and typed responses on the final test, while leaving the training phase identical to that of Experiment 2.

Method

Subjects. A priori power analyses were performed to select a target sample size for this experiment (and for Experiment 4), based on the standard deviations of the condition differences scores for the cued recall tests in Experiments 1 and 2. For a one-tailed matched *t*-test (testing the hypothesis of a positive testing effect) a target sample size of 60 was selected, yielding an 82% chance of detecting a testing effect of at least 0.05 (in units of proportion correct) and a better than 99% chance of detecting an effect of at least 0.10. Accordingly, sixty-three University of California San Diego undergraduate students participated for course credit. All but two subjects completed both sessions of the experiment.

Materials, design, procedure, and coding. Materials, design, and procedures were identical to those in Experiment 2, except that the test session was typed rather than handwritten, and took place in our laboratory rather than in a classroom. This involved programming the free recall test, cued recall test, and exit survey using E-Prime software (Psychology Software Tools, Pittsburgh, PA). The free recall and cued recall responses were no longer handwritten on a double-sided worksheet; instead, subjects typed their answers on laboratory computers while viewing a screen which featured an indicator number, a blinking cursor, and a blank space. During the free recall test, subjects pressed the Enter key to save and clear each typed response and increment the indicator number by one. During the cued recall test, each typed response would automatically save and clear after 12 s had elapsed.

Due to test session data existing in typed form, no transcription was necessary; however, as before, each word typed in the free recall portion had to be separately coded as being a phonetically identifiable match or not to one of the 40 studied words. Unlike the prior experiments, on the free recall test the computer screen showed only the word the subject had recalled and attempted to spell on each trial, resulting in an increased rate of duplicate words and intrusions; each instance was separately coded as such, and only the first instance of a phonetically identifiable word that was previously trained on was analyzed.

Results and Discussion

Training. The training results are in line with the prior experiments (Figure 2); the mean proportion of words spelled correctly increased with each block (from first to last: 0.33, 0.69, 0.73, 0.84). The proportion of spelling attempts accurately scored during feedback was 0.96. Subjects completed an average of 3.08 correctly spelled repetitions per word on all copying trials.

Delayed test. On the delayed test, there was again a robust incidental testing advantage for words free-recalled (Figure 3), $t(60) = 10.53$, $p < 0.0001$, $d = 1.35$. There was no trend toward a training condition effect on spelling accuracy among words free-recalled (Figure 4). On the cued recall test there was a non-significant trend toward better performance in the T+FB training condition (Figure 5), $t(60) = 1.95$, $p = 0.06$, $d = 0.25$.

Although the marginally significant cued recall results of this experiment raise the possibility that the matched motor processing during training and test in the prior experiments may have slightly benefited the copying condition, the broader implication of the first three experiments is that, for adults at least, copying is highly competitive with T+FB as a strategy for learning to spell. Copying is emphatically not competitive with testing, however, with respect to

free recall of spelling words.

Questionnaire results. Consistent with the results of the prior experiments, 93% of survey respondents reported previously having used copying to learn spelling and 95% reported previously having used testing with feedback to learn spelling. When asked to rate their relative preference for either technique, 56% of respondents gave ratings that were in the direction of testing. The relative preference for testing was statistically significant, $t(55) = 3.19$, $p = 0.0023$, $d = 0.43$. Respondents also endorsed testing as more effective than copying, $t(60) = 4.26$, $p < 0.0001$, $d = 0.54$.

Experiment 4

The final experiment compared T+FB to a more traditional control condition in the testing effect literature, reading. A testing effect relative to reading has been established in multiple task domains outside of spelling (e.g., McDaniel, Howard, & Einstein, 2009; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006b). The evidence for reading's effectiveness for spelling acquisition, however, is mixed (e.g., Bosman & de Groot, 1991; Conrad, 2008; Ehri, 1997).

Method

Subjects. Fifty-nine University of California San Diego undergraduate students participated for course credit. All but one subject completed both sessions of the experiment.

Materials, design, procedure, and coding. Materials, design, procedures, and coding were identical to those of Experiment 3, with the exception of the following modifications. Because the reading condition involved vocalization, training in a classroom setting was not desirable. Instead, each subject completed the training session in one of four isolated laboratory rooms with no other persons present except for the experimenter. Both sessions of the

experiment were programmed using E-Prime software and conducted using our laboratory computers. Materials and procedures from Experiment 3 were duplicated, except that reading blocks were used in place of copying blocks, and that within each reading block, each word was presented for 6 s at a time. This duration of word presentation matched that used in the T+FB blocks, and was more than sufficient for subjects to read aloud each word three times (in contrast, for Experiments 1-3, copying blocks with a duration of 12 s each were necessary to accommodate about three handwritten repetitions per word). Hence there were four training blocks for each task grouped into two continuous eight min periods (see right side of Figure 1). The number of audio repetitions per presentation was further reduced to one every six s to simplify the process of subjects reading aloud, while keeping overall trial durations unchanged.

As was the case for the testing and copying training conditions in the prior experiments, subjects were run sequentially in groups of four, but in separate rooms for each subject for each session, and with each group assigned to one of four groups featuring counterbalanced assignment of list to condition (reading or T+FB) and counterbalanced starting order (reading or T+FB first). The test session was conducted in the same computer-based manner as Experiment 3, and used a modified version of the exit survey in which reading was mentioned in place of copying.

Reading. In each reading block, 10 words were presented and read aloud three times per presentation by subjects. This cycle of presentation (and reading aloud) was repeated three more times within each block. Individual words were presented both aurally through the computer speakers and visually on the computer screen at a rate of six s per word. For each presented word, the onscreen instructions told subjects to speak the word out loud three times in a row. Reading out loud is consistent with prior spelling research (e.g., Ehri, 1997).

Testing with feedback. Procedures were identical to Experiment 3, except that the instructions were presented by computer rather than on a classroom projector screen. Due to the elimination of copying blocks, subjects were only provided with one double-sided worksheet which was evenly divided into four quadrants of 10 blank lines each and folded in half, along with a ballpoint pen. Subjects turned the folded worksheets over after the first two phases of each T+FB block. After completion of the first two T+FB blocks, the experimenter collected the worksheet and reversed the fold, exposing the unused reverse side, and returned it to the subject for use in the remaining two T+FB blocks.

Results and Discussion

Training. The training results mirrored those of the prior experiments (Figure 2); the mean proportion of words spelled correctly increased with each block (from first to last: 0.31, 0.62, 0.68, 0.80). The proportion of spelling attempts accurately scored during feedback was 0.95.

Delayed test. On the delayed free recall test, the now familiar incidental testing effect was observed, $t(57) = 11.18$, $p < 0.0001$, $d = 1.47$. Results for the test session diverged from that of the prior experiments, however. A marginally significant testing effect was observed with respect to proportion of free-recalled words spelled correctly (Figure 4), $t(44) = 1.96$, $p = 0.056$, $d = 0.29$. More importantly, a highly significant testing advantage was observed on the cued recall test (Figure 5), $t(57) = 8.2$, $p < 0.0001$, $d = 1.08$. Thus, in line with the typical result in the literature, there is a robust testing effect for spelling relative to reading.

The primary inference from the four experiments – that among adults copying, but not reading, is competitive with T+FB for learning to spell – is buttressed by a two sample *t*-test performed on the training condition cued recall difference scores of Experiments 3 (testing

minus copying) and 4 (testing minus reading), $t(117) = 7.30$, $p < 0.0001$, $d = 1.34$.

Questionnaire results. On the exit survey, 62% of respondents reported previously having used reading to learn spelling and 74% reported previously having used testing with feedback to learn spelling. In terms of relative preference for either technique, 74% of respondents gave ratings that were in the direction of testing. The preference for testing over reading was highly significant, $t(53) = 6.60$, $p < 0.0001$, $d = 0.89$. Respondents also strongly endorsed testing as more effective than reading, $t(57) = 6.32$, $p < 0.0001$, $d = 0.83$.

General Discussion

We investigated the relative effectiveness of three strategies for learning to spell among adults, each of which is used frequently: testing with feedback, copying, and reading. Testing with feedback proved to be as effective as writing and more effective than reading, reinforcing a large experimental data base in which the effectiveness of retrieval practice as a learning strategy has been demonstrated. However, testing with feedback did not produce more learning than did writing (i.e., there was no testing effect in that comparison), despite our adaptation of an experimental paradigm with which we have previously shown a testing advantage for spelling among young children (Jones et al., 2015), and despite progressive optimization of testing from Experiments 1 to 3. Only in Experiment 4, in which the control task was reading, was a testing effect observed. Overall, these results raise important questions regarding the learning mechanisms that underlie spelling, and perhaps test-enhanced learning more generally.

The largely successful *retrieval practice* account of the testing effect, which assumes that the act of successful memory retrieval during testing is a particularly potent learning event (Bjork, 1975; Roediger & Karpicke, 2006a), can readily explain why reading (which presumably involves no retrieval practice) is relatively ineffective for spelling. By the same logic, it can

explain the robust advantage of testing over copying among children. It cannot, however, straightforwardly explain the absence of an advantage for testing over copying observed here for adults, nor the strong interaction between testing vs. copying in Experiment 3 and testing vs. reading in Experiment 4. In the testing conditions of Experiments 2 and 3, the majority (82.5%) of words had at least two correct spelling retrievals per word by the end of training, and final block training accuracy was approximately 0.85. Those experiments should thus have been highly sensitive to the retrieval practice effect, were that effect a consistently more potent basis for learning than copying.

Several other current theories of the testing effect that make more specific claims about retrieval practice appear to encounter similar difficulties in explaining the overall pattern of results. Both the *elaborative retrieval* (Carpenter, 2009) and *mediator effectiveness* (Pyc & Rawson, 2010) theories, as examples, could account for the absence of a testing effect relative to copying in the current experiments, under the assumption that elaborative processing and use of mediators is less effective when learning occurs at the level of phoneme-to-grapheme mappings (for which elaborative processing and keyword mediators likely have limited utility). Those models would not readily account, however, for the cases in which there is a testing effect for spelling (i.e., our prior testing experiments with children and Experiment 4 of the current paper). The apparent equivalence of copying and testing with feedback among adults suggests that retrieval practice itself is not the primary mechanism by which spelling skill acquisition occurs, and it invites consideration of an alternate theoretical account.

Learning through Focused Study

We propose here that study-based processing, rather than inherently test-based processing, may be the primary mechanism for learning to spell for all types of training tasks and

across all populations, and that some training tasks afford more opportunity for focused study than do others. The term *focused study* refers here to study that efficiently and selectively draws attention and learning effort toward the new knowledge that must be acquired to support accurate performance.

We do not advance the focused study hypothesis as a global alternative to retrieval practice as an account of learning in the testing effect paradigm. Indeed, the centrality of retrieval practice is indisputable in our view given the strong evidence that even testing without feedback can produce learning that is superior to restudy (Roediger & Butler, 2011); when there is no feedback on a test trial, answer retrieval itself rather than any type of study activity can safely be presumed as basis of learning. Rather, we propose that the relative efficacy of retrieval practice and study for learning can vary substantially over task domains. In the domain of spelling at least, we suggest that focused study may be the more potent of those two learning mechanisms. Below we advance a focused study account that has the potential to explain the spelling results across tasks for both adults and children.

For adults learning to spell difficult words, it is generally not the entire letter sequence that is difficult to remember but rather one or more subsets of irregular letter sequences within the words (Alper, 1942), usually those that have atypical or easily confused phoneme-to-grapheme mappings (Ehri, 1997; Frith, 1980). Consider the example of *questionnaire*. For most adults the difficult section is likely to be "*-nnaire*." The reader may agree that most if not all of the words used in this study (see Appendix A) have a similar property. By our hypothesis, training strategies that facilitate selective study of those letter sequences are more effective than those that do not. Clearly testing with feedback can be one such strategy. In our experiments, subjects received correct spelling feedback on each test trial, to which they compared their

spelling of the word. If their spelling was incorrect, there was an opportunity to identify and study the incorrectly spelled letter sequences, while ignoring other parts of the word that they spelled correctly (i.e., focused study). Such a learning mechanism would be consistent with error-correction accounts of the testing effect (e.g., Mozer, Howe, & Pashler, 2004).

Copying is also likely to promote focused study among adults, who have extensive spelling expertise. As the adult subjects wrote each letter of a presented word, they are likely to have sub-vocally pronounced the word (cf. Zago, Poletti, Corbo, Adobatti, & Silani, 2008) or to have compared the visual presentation of the word to their copying of it. Based on their spelling expertise, incongruous phoneme-to-grapheme sequence mappings are likely to become salient during that process, providing an opportunity to focus attention on learning those mappings while copying. Less attention would be needed to sections of a word that have standard and highly familiar phoneme-to-grapheme mappings.

The reading task, in contrast, may be less likely to promote focused study on the word's spelling, because accurate reading of a word form that is familiar (even if the exact spelling is not) is possible without attention being drawn to all of the component letters or to unusual phoneme-to-grapheme sequence mappings (Bosman & Van Orden, 1997; Conrad, 2008). The well-known word superiority effect, wherein a word can be recognized prior to recognition of some of its letters (e.g., Healy & Drewnowski, 1983; Reicher, 1969), is another case in point. Similarly, subjects may sometimes have associated the whole-word orthographic structure with its pronunciation, particularly for difficult-to-spell words (Bosman & Van Orden, 1997). In either case, reading would not be an efficient means of promoting focused study of spelling. With regards to the present study, although the focused study hypothesis does not necessarily predict equivalent effectiveness of testing with feedback and copying among adults, it is

consistent with that result and it predicts the superiority of both methods to reading among adults.

For spelling, testing with correct answer feedback can be expected to engage focused study for all populations, and hence it can be effective among both adults and children. For other training tasks, the extent to which focused study can occur is likely to depend in part on the domain expertise of the subject. Young children have much less spelling expertise than do adults (Treiman & Cassar, 1997), and in particular they have less knowledge about common letter strings and their phoneme-to-grapheme mappings (Ehri, 1997). Thus, children may not have the capacity to perform focused study during copying or other study activities. The focused study hypothesis thus provides a plausible account for the large testing effect relative to copying that was observed among children (Jones et al., 2015), but not adults.

The focused study hypothesis avoids circularity by making novel and testable predictions. For example, it predicts that vocalized letter-by-letter pronunciation of a visually presented word will be relatively ineffective for learning to spell (vs. testing with feedback or copying among adults). In that task, attention is drawn to the phonemic properties of each letter when it is named, and away from the actual phoneme-to-grapheme sequence mappings that correspond to word pronunciation, thus hampering focused study of mappings that do not replicate pronunciation based on letter naming. Hence neither reading nor letter pronunciation focus study at the level necessary for efficient spelling skill acquisition. In contrast, a pure restudy condition, in which subjects are visually and aurally presented with spelling words and are allowed to study them however they would like in preparation for a future spelling test, would be expected to be as effective, or possibly more effective, than testing with feedback for adults, at least as testing is implemented in the current experiments. This is most likely because subjects' attention is not

diverted to information that is largely irrelevant for spelling, as can occur in the aforementioned uncompetitive learning tasks. This last prediction is provocative given that learning through testing with feedback is superior to restudy in the great majority of the testing effect literature. Finally, because of their lower domain expertise, pure restudy should be less effective for children than is testing with feedback.

In the discussion above, the goal was to demonstrate the potential viability of the focused study hypothesis as a sufficient mechanism to account for the recent spelling results. Another possibility is that it is only one of two or more operative learning mechanisms. For example, it may be that focused study is promoted more by copying than by testing with feedback, but that correct retrieval practice during testing also produces learning, yielding by coincidence nearly equivalent final test spelling performance in those two conditions for adults.

Design Limitations and Further Optimization of Training

There are also potential limitations of our experimental designs that may have influenced the current results. One possibility is that inefficiencies in the testing procedure remained despite efforts at optimization while maintaining strict control of time on task in Experiments 1 through 3. In particular, the dosage of training, as measured by repetition, was higher for copying than it was for testing; that is, subjects completed more handwritten repetitions during training in the copying condition (typically six repetitions per word in each block) than occurred during training test trials in the testing condition (one repetition per word in each block in Experiment 1; two repetitions per word in Experiments 2 and 3). Because each testing with feedback trial involves two steps rather than one, we believe that most implementations of testing for spelling are inherently more time consuming at the trial level than is a single copying iteration, making simultaneous control of repetition and time on task difficult.

Nevertheless, it might be possible to further increase the number of testing repetitions relative to copying, such as by further reducing testing trial times (and accepting that subjects will not always have enough time to finish memory retrieval or their writing attempts), or by having subjects write out correct spellings during each feedback trial rather than simply checking their answers (and accepting that written copying of visually presented words then occurs in both training conditions). Yet another option would be to relax the constraint of equating time on task, and to equate number of handwritten repetitions in the two tasks instead. Under such a scenario, however, a testing advantage could be obtained even if testing yields less learning per unit time on task, with uncertain educational implications.

When considering the issue of training dosage as indexed by handwritten repetitions, it should also be noted that the doubling of testing in Experiments 2 and 3 did not produce a commensurate spelling improvement, relative to a copying condition that was left unchanged. Moreover, the implementation of copying and testing with feedback was very similar to that which we used for children (Jones et al., 2015), and a large testing effect was observed in that study. Testing was also clearly superior to reading in Experiment 4, despite an identical implementation of testing in Experiments 3 and 4 and an almost identical procedural implementation of copying and reading in those two experiments. Accordingly, it seems unlikely in our view that inefficiency of the testing condition is the main driving factor of the current results. At a minimum, our results show that copying is much more competitive with testing than is reading, independently of any issues related to efficiency of task implementation.

Another modification that might enhance the effectiveness of testing is the inclusion of an initial study phase for all to-be-learned spelling words. Such a study phase would allow subjects to recall their prior study experiences and would likely enhance performance on the first

training test block. Some theories of testing effect (e.g., Karpicke, Lehman, & Aue, 2014), treat that event as critical to the optimal implementation of test-enhanced learning. One reviewer also pointed out that in the absence of an initial study phase, the first training test trial in the testing condition might be construed as a pretest. Research on the effects of pretests, or the *pretesting effect* (e.g., Hays, Kornell, & Bjork, 2013; Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009), has produced mixed results. When feedback is given immediately or after a delay of up to two minutes (longer than the feedback delay in most of our experiments), pretesting can be more effective for learning than study (Richland et al., 2009). However, if feedback occurs at longer delays, pretesting can be less effective than restudy (Hays et al., 2013). It is also worth noting that, unlike pretesting studies, in which there is typically a single pretest trial, in the current experiments had one or three subsequent testing with feedback trials. We believe that these additional trials are likely to have attenuated any pretesting effect or any effects of reduced performance on the first training test trial, although a future study which includes initial study exposure for all spelling words would circumvent any complication that would arise due to such effects.

A third training-based explanation of the current results is that subjects engaged in covert retrieval practice in the copying condition, and that covert practice is responsible for the competitiveness of copying with testing. Covert retrieval has been shown to yield testing effects on a par with those observed when answer retrieval is overt, at least for the cases of cued recall (Putnam & Roediger, 2013) and free recall (Smith, Roediger, & Karpicke, 2013) on a final test. It is also conceivable that subjects spontaneously engage in covert retrieval even when, as with copying, answer retrieval is not part of the assigned learning task. Covert retrieval in the copying condition seems unlikely in our view given that the spelling of each word was visually

available throughout all copying trials. During each of those trials, subjects first attended to the visually presented word and immediately copied it. For the remaining (majority) of the trial time, they continued copying with the correct spelling already written immediately next to the current copying iteration, and hence directly available to assist with spelling. In Experiment 3, the word to be copied appeared on the computer screen precisely in front of each subject, making visual word spellings even more easily accessible than in Experiments 1 and 2.

Moreover, if covert retrieval had occurred during copying trials, we might expect to see frequent misspellings. However, misspellings in the copying task were exceedingly rare, estimated based on random sampling to be less than 1.3% of all copies in Experiment 1 through 3. Finally, if covert retrieval occurred in the copying condition, we would also expect it to have occurred in the reading condition. After first attending to visually presented words in the reading condition, subjects could have engaged in relatively automatic repeated reading without looking at the stimulus. While doing so they could have attempted to recall the presented spelling using visual memory, and then check their recall accuracy against the visually presented word as the trial ended. Yet, reading yielded markedly less learning than did either copying or testing with feedback.

A final training-based account of the current results, first broached in the discussion of Experiment 2, is that there was motivational carry-over between the testing and copying conditions. We had no a priori reason to expect that carryover, however, and even if some form of motivational carry-over did occur, it seems unlikely that it would have yielded equivalent spelling performance for the two conditions on the cued recall test. If it did, that result would imply that testing effects are primarily a consequence of higher motivation during testing than during other training activities, a possibility that has not gained much traction in the literature.

Nevertheless, there is some indication that testing effects can be reduced in within-subjects designs (e.g., Rowland, 2014; Soderstrom & Bjork, 2014), and it remains for future work to conclusively assess that possibility.

Besides the optimizations discussed above, there may be further ways to improve testing, such as exploring ways to better engage subjects throughout testing trials and consideration of alternative feedback schemes (e.g., timing and nature of feedback). When considering such modifications, however, it is important to emphasize that where testing effects have been observed in the literature, there has been minimal evaluation of whether those effects hold beyond the particular control task used, nor have there been systematic efforts to determine whether the control task used was implemented efficiently. Showing that testing with feedback produces more learning after extensive optimization is in our view not very meaningful if the reference condition has not been similarly optimized. Thus, the issue of task selection and efficiency is a challenge to the broader testing effect literature and not exclusively to the current experiments.

Metacognitive Judgments of Relative Task Preference and Effectiveness

When asked to rate their relative preference for the learning techniques that were used in the current study, subjects in Experiments 2 through 4 chose testing. Subjects also gave testing higher effectiveness scores than copying in Experiment 3 and reading in Experiment 4. This strong preference for testing and the belief that it is more effective than reading contrasts with the majority of the testing effect literature, in which a study control task is generally preferred to testing (Roediger & Karpicke, 2006b; Tullis, Finley, & Benjamin, 2013). Moreover, the metacognitive judgments in the present study mirror those that we found for testing vs. writing with young children, who also preferred testing and endorsed it as more effective (Jones et al.,

2015).

What accounts for this result? We considered two possibilities. First, in the metacognitive literature, it is commonly observed that learners prefer easier, less challenging learning tasks because they perform better on them during training (i.e., there is greater fluency for easier tasks), and tend to underappreciate the fact that training conditions that are more difficult tend to produce better learning and retention in the long term (Bjork, 1999; Tullis et al., 2013). Accordingly, it may be that in the current study, testing is less tiring than copying, due to less continuous motor output being required for testing. However, the same is presumably not true of reading, implying that the preference for testing is somehow driven by the testing experience. A second possibility is that the act of repeated cycles of testing with feedback for spelling is intrinsically motivating. Subjects were likely aware of their performance improvement across test repetitions, whereas no such improvements are readily observable with copying or reading, nor in experiments with only one test or restudy trial per item during training, which are more common in the literature. It may be that awareness of improved retrieval fluency or accuracy across test trials resulted in students both preferring testing and endorsing it as more effective than copying or reading.

Incidental Effects of a Spelling Test on Free Recall of Words

While secondary to the spelling results, the finding of far superior word free recall in the testing condition than in the copying or reading conditions was surprising, particularly in light of the finding of no testing effect for spelling relative to copying. An analogous effect of greater free recall for cue words of previously tested items than for the words of previously restudied items was observed by Carpenter et al. (2006) for the case of paired associates. The theoretical basis for those effects is not yet clear. Regarding the current results, one speculative possibility

is that neither the copying nor the reading task results in sustained activation of the presented word representations, whereas the testing task does. This could be related to the fact that only in the testing condition were words not presented visually at the outset. With respect to copying, processing may be dominantly or exclusively at the phonemic and letter sequence levels rather than the lexical (or semantic) levels, given that the word was always visually available. Similarly, although the reading task does involve processing initially at the lexical level, reading repetition may be supported primarily by motor sequence repetition and may not necessitate sustained lexical activation. Testing, on the other, seems much more likely to engage lexical (and perhaps semantic) processing for a more extended period of time as the spelling attempt is made. Another factor that may be unique to the testing task, and that may improve spelling word memory, is the need to transcode from the auditory input to a visual word representation to assist with spelling. In the copying and reading tasks, words are presented in both visual and auditory form during training and thus no modality transcoding is required.

Regardless of the theoretical basis of the free recall effect, those results suggest that the learning processes supporting word free recall are distinct, and possibly independent of, the learning processes involved in spelling. The most straightforward, albeit speculative, account is that learning to spell exclusively engages learning mechanisms that map between phonological and graphemic levels of representation, whereas the learning that underlies word recall occurs exclusively at the lexical and semantic levels.

Educational Implications

Our results unambiguously show that, among adults, reading is a poor strategy for spelling acquisition, whereas testing and copying are both effective. This conclusion supports the continued widespread use of both techniques for adult students. Moreover, given that

repeated cycling through one learning strategy can induce boredom, a strategy of interleaving testing with feedback may be optimal for adults. Alternatively, it may be productive to include supplemental copying trials for words that are incorrectly answered on test trials—essentially, combining both testing with feedback and copying for the study of difficult spelling words. On the other hand, the empirical evidence to date for children clearly identifies testing with feedback as the most effective strategy. Children are exposed to a myriad of training tasks for spelling, however (see Graham, 1983; Jones et al., 2015), and it remains to be seen whether testing with feedback is the only highly effective strategy.

Another potentially important translational question raised by our results is whether the greatly enhanced free recall of words following a spelling test extends to naturalistic settings. In a variety of educational settings, to-be-learned spelling words are not a part of students' spontaneously used vocabularies, even if the words' meanings are known. Do spelling tests enhance incorporation of those words into natural speech or copying? If, as we have hypothesized, the superior word free recall in the testing condition reflects greater semantic elaboration during the spelling test than in either the copying or reading condition, then we expect that it would. Given the large free recall effects in our experiments, the educational value of spelling tests may prove to be as substantial for expansion of spontaneously available vocabulary as it is for spelling.

Conclusions

The current results indicate that testing with feedback does not stand alone as an effective learning strategy for spelling. For adults at least, comparable levels of learning occur with copying. This finding invites examination of other candidate techniques for learning to spell. Ultimately, it may prove to be the case that a range of different instructional techniques which

afford opportunities to perform focused study of difficult phoneme-to-grapheme mappings are able to substantially facilitate spelling acquisition.

References

Alper, T. G. (1942). A diagnostic spelling scale for the college level: Its construction and use. *Journal of Educational Psychology, 33*(4), 273-290. doi:10.1037/h0057905

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge: MIT Press.

Bosman, A. M. T., & de Groot, A. M. B. (1991). De ontwikkeling van woordbeelden bij beginnende lezers en spellers. (The development of orthographic images in beginning readers and spellers). *Pedagogische Studiën, 68,* 199-215.

Bosman, A. M. T., & de Groot, A. M. B. (1992). Differential effectiveness of reading and non-reading tasks in learning to spell. In F. Satow and B. Gatherer (Eds.), *Literacy without frontiers*. (pp. 279-289). Widnes, Cheshire: United Kingdom Reading Association.

Bosman, A. M. T., & Van Orden, G. C. (1997). Why spelling is more difficult than reading. In C. Perfetti, L. Riben, & M. Fayol (Eds.), *Learning to spell: Research, theory, and practice across languages.* (pp. 173-194). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Burt, J. S. (2006). Spelling in adults: The combined influences of language skills and reading experience. *Journal of Psycholinguistic Research, 35*(5), 447-470. doi:10.1007/s10936-006-9024-9

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of

elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. doi:10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268-276. doi:10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*(6), 760-771. doi:10.1002/acp.1507

Conrad, N. J. (2008). From reading to spelling and spelling to reading: Transfer goes both ways. *Journal of Educational Psychology, 100*(4), 869-878. doi:10.1037/a0012544

Cronnell, B., & Humes, A. (1980). Elementary spelling: What's really taught. *The Elementary School Journal, 81*(1), 59–64.

Cunningham, A. E., & Stanovich, K. E. (1990). Early spelling acquisition: Writing beats the computer. *Journal of Educational Psychology, 82*(1), 159-162. doi:10.1037/0022-0663.82.1.159

Ehri, L. C. (1997). Learning to read and learning to spell are one and the same, almost. In C. Perfetti, L. Riben, & M. Fayol (Eds.), *Learning to spell: Research, theory, and practice across languages.* (pp. 237-269). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Frith, U. (1980). Unexpected spelling problems. In U. Frith (Ed.), *Cognitive Processes in Spelling.* (pp. 495-515). New York, NY: Academic Press.

Graham, S. (1983). Effective spelling instruction. *The Elementary School Journal, 83*(5), 560-567.

Grskovic, J. A., & Belfiore, P. J. (1996). Improving the spelling performance of students with

disabilities. *Journal of Behavioral Education, 6*(3), 343-354.

Hartmann, G. W. (1931). The constancy of spelling ability among undergraduates. *Journal of Educational Research, 24*(4), 303-305.

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), 290-296. doi:10.1037/a0028468

Healy, A. F., & Drewnowski, A. (1983). Investigating the boundaries of reading units: Letter detection in misspelled words. *Journal of Experimental Psychology: Human Perception and Performance, 9*(3), 413-426. doi:10.1037/0096-1523.9.3.413

Heckendorn, R. (2014). Robert Heckendorn's list of hard to spell words. Moscow, ID: University of Idaho. Available from: http://marvin.cs.uidaho.edu/misspell.html

Horn, E. (1967). *Teaching spelling: what research says to the teacher*. Washington, D.C.: Department of Classroom Teachers, American Educational Research Association of the National Education Association.

Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., and Rickard, T. C. (2015). Beyond the rainbow: retrieval practice leads to better learning than does rainbow writing. *Educational Psychology Review* (in press). doi:10.1007/s10648-015-9330-6

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 61) (pp. 237-284). San Diego, CA: Elsevier Academic Press.

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), 989-998. doi:10.1037/a0015729

Longcamp, M., Boucard, C., Gilhodes, J., Anton, J., Roth, M., Nazarian, B., & Velay, J. (2008). Learning through hand- or typewriting influences visual recognition of new graphic shapes: Behavioral and functional imaging evidence. *Journal of Cognitive Neuroscience, 20*(5), 802-815. doi:10.1162/jocn.2008.20504

Longcamp, M., Zerbato-Poudou, M., & Velay, J. (2005). The influence of writing practice on letter recognition in preschool children: A comparison between handwriting and typing. *Acta Psychologica, 119*(1), 67-79. doi:10.1016/j.actpsy.2004.10.019

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 494-513.doi:10.1080/09541440701326154

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, 20(4), 516-522. doi:10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200-206. doi:10.3758/BF03194052

McGuffin, M. E., Martz, S. A., & Heron, T. E. (1997). The effects of self-correction versus traditional spelling on the spelling performance and maintenance of third grade students. *Journal of Behavioral Education, 7*(4), 463-476.

McNeish, J., Heron, T. E., & Okyere, B. (1992). Effects of self-correction on the spelling

performance of junior high school students with learning disabilities. *Journal of Behavioral Education, 2*(1), 17-27.

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In *Proceedings of the Twenty Sixth Annual Conference of the Cognitive Science Society.* (pp. 975-980). Hillsdale, NJ, Lawrence Erlbaum Associates Publishers.

Murphy, J. F., Hern, C. L., Williams, R. L., & McLaughlin, T. F. (1990). The effects of the copy, cover, compare approach in increasing spelling accuracy with learning disabled students. *Contemporary Educational Psychology*, *15*(4), 378-386.

Ormrod, J. E. (1986). Learning to spell: three studies at the university level. *Research in the Teaching of English, 20*(2), 160-173.

Ormrod, J. E., & Jenkins, L. (1989). Study strategies for learning spelling: Correlations with achievement and developmental changes. *Perceptual and Motor Skills, 68*(2), 643-650.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Available from http://ncer.ed.gov.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*(1), 3-8. doi:2004-22496-001

Putnam, A. L., & Roediger, H. L.. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36-48. doi:10.3758/s13421-012-0245-x

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81,* 275-280.

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), 243-257. doi:10.1037/a0016496

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382-395. doi:10.1037/a0026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248. doi:10.1016/j.jarmac.2012.09.002

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin,* doi:10.1037/a0037559

Scripps National Spelling Bee Consolidated Word List (2004). Cincinnati, OH: E. W. Scripps Company.

Shahar-Yames, D., & Share, D. L. (2008). Spelling as a self-teaching mechanism in orthographic learning. *Journal of Research in Reading, 31*(1), 22-39. doi:10.1111/j.1467-9817.2007.00359.x

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology, 87*(4), 267-298. doi:10.1016/j.jecp.2004.01.001

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1712-1725. doi:10.1037/a0033569

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99-115. doi:10.1016/j.jml.2014.03.003

Treiman, R., & Cassar, M. (1997). Spelling Acquisition in English. In C. Perfetti, L. Riben, & M. Fayol (Eds.), *Learning to spell: Research, theory, and practice across languages.* (pp. 61-80). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*(3), 429-442. doi:10.3758/s13421-012-0274-5

Venetsky, R. I. (1980). From Webster to Rice to Roosevelt: the formative years for spelling instruction and spelling reform in the U.S.A. In U. Firth (Ed.), *Cognitive Processes in Spelling.* (pp. 9-32). New York, NY: Academic Press.

Wilson, M.D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary,

Version 2. *Behavioural Research Methods, Instruments and Computers*, *20*(1), 6-11.

Wirtz, C. L., Gardner, R., Weber, K., & Bullara, D. (1996). Using self-correction to improve the spelling performance of low-achieving third graders. *Remedial and Special Education, 17*(1), 48-58.

Zago, S., Poletti, B., Corbo, M., Adobbati, L., & Silani, S. (2008). Dysgraphia in patients with primary lateral sclerosis: A speech-based rehearsal deficit? *Behavioural Neurology, 19*(4), 169-175.

Block No.	Word List	Experiment 1 Testing vs. Copying	Experiments 2, 3 Testing vs. Copying	Experiment 4 Testing vs. Reading
1	A1	(T + FB)	(T + FB) + (T + FB)	(T + FB) + (T + FB)
2	B1	C + C	C + C	R + R + R + R
3	A2	(T + FB)	(T + FB) + (T + FB)	(T + FB) + (T + FB)
4	B2	C + C	C + C	R + R + R + R
5	A1	(T + FB)	(T + FB) + (T + FB)	(T + FB) + (T + FB)
6	B1	C + C	C + C	R + R + R + R
7	A2	(T + FB)	(T + FB) + (T + FB)	(T + FB) + (T + FB)
8	B2	C + C	C + C	R + R + R + R

Figure 1. Training session block design for Experiments 1-4. Subjects trained on 40 words in eight four-minute blocks of 10 words each, using techniques that alternated with each successive block. Experiments 1-3 featured testing with feedback (T+FB) vs. copying (C); Experiment 4 featured T+FB vs. reading (R). Note: the figure shows one example of the word lists used per block, plus ordering of training conditions; in all experiments, assignment of word list to training condition and training technique to odd or even numbered blocks was counterbalanced.

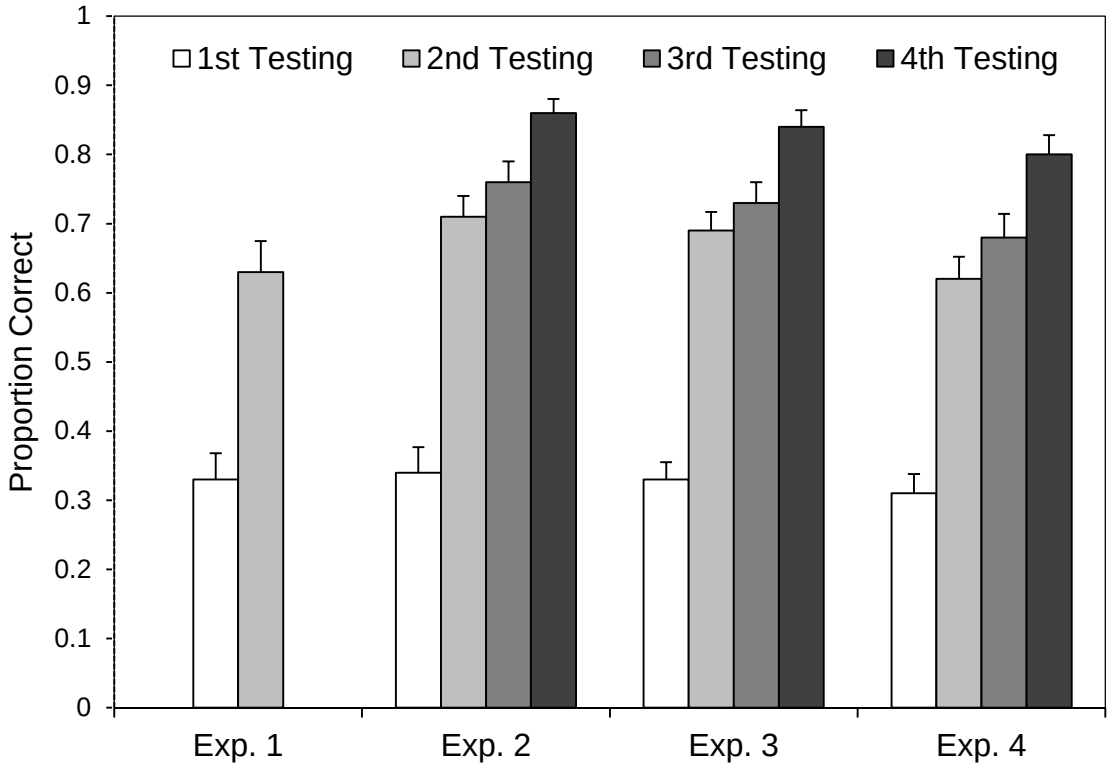

Figure 2. Training data for words trained using testing with feedback for Experiments 1, 2, 3, and 4. Error bars for each experiment are within-subject standard errors.

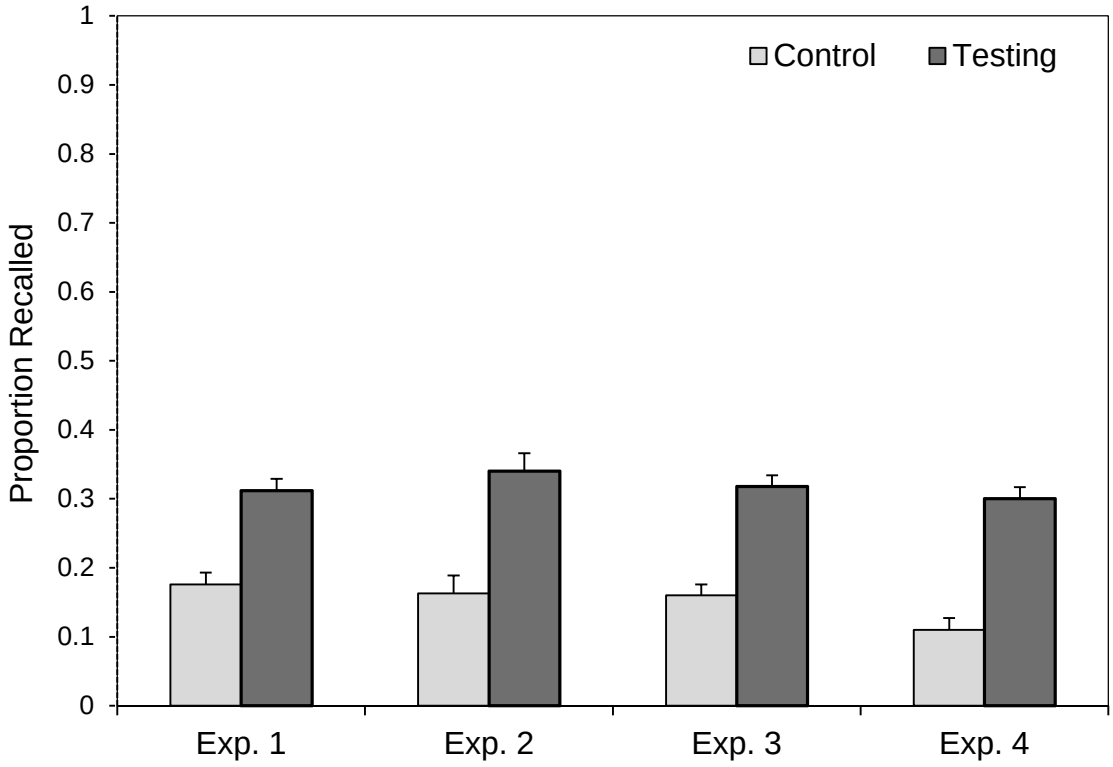

Figure 3. Proportion of phonetically identifiable words recalled in the free recall test of Experiments 1, 2, 3, and 4. Control refers to copying for Experiments 1-3 and reading for Experiment 4. Error bars are within-subject standard errors.

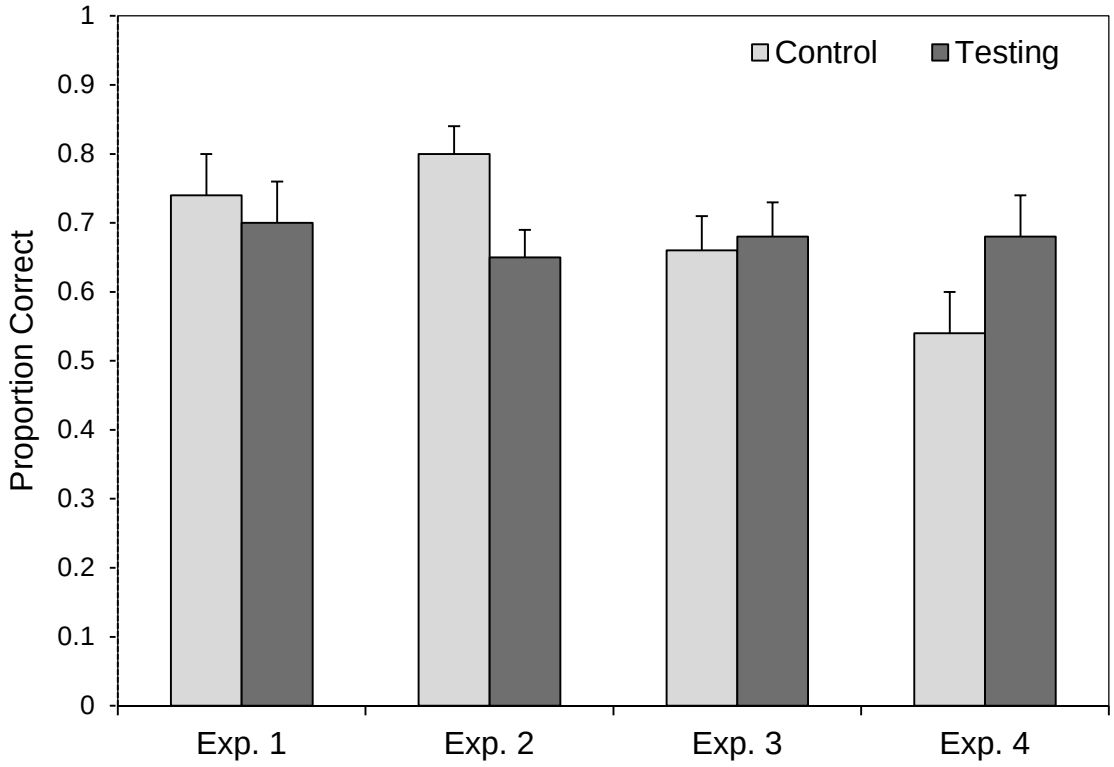

Figure 4. Proportion of words correctly spelled in the free recall test of Experiments 1, 2, 3, and 4. Control refers to copying for Experiments 1-3 and reading for Experiment 4. Error bars are within-subject standard errors.

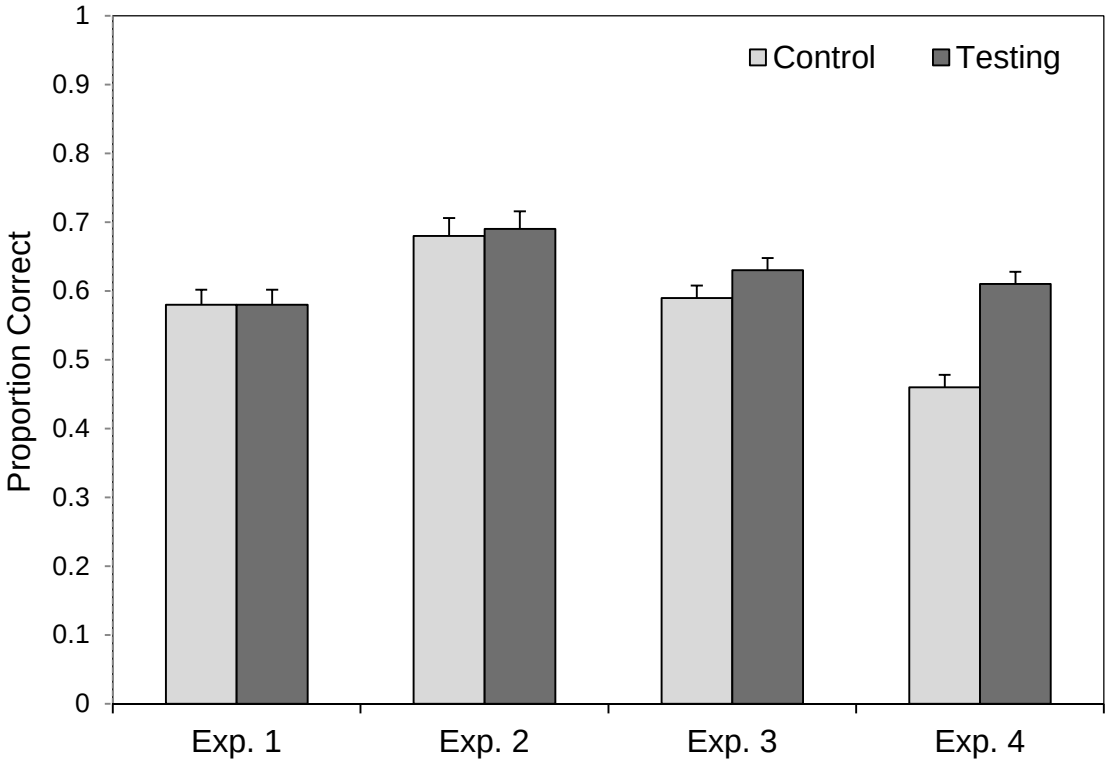

Figure 5. Proportion of words correctly spelled in the cued recall test of Experiments 1, 2, 3, and 4. Control refers to copying for Experiments 1-3 and reading for Experiment 4. Error bars are within-subject standard errors and are thus appropriate for interpreting the condtion differences within each experiment.

Appendix A

Lists of Spelling Words Used in Experiments 1, 2, 3, and 4

	List A		List B	
	A1	A2	B1	B2
1.	bourgeoisie	boulevard	accommodation	boutonniere
2.	camaraderie	camouflage	cataclysm	chameleon
3.	colloquium	cantaloupe	chauffeur	cornucopia
4.	hallelujah	corduroy	connoisseur	diaphragm
5.	mayonnaise	daiquiri	lieutenant	entrepreneur
6.	porcelain	embarrassment	lingerie	handkerchief
7.	questionnaire	laryngitis	mannequin	masquerade
8.	racquetball	limousine	penicillin	perseverance
9.	sauerkraut	renaissance	turquoise	schizophrenia
10.	zucchini	spaghetti	ukulele	souvenir

Appendix B

Metacognitive Questionnaire Used in Experiments 1, 2, 3, and 4

Experiments 1-3:

Last week, you were taught to spell vocabulary words using the following two techniques:

- **Testing with feedback** - any form of quizzes or tests, and getting the correct answers
- **Copying** - writing words multiple times in succession

1. Please rate your relative preference for the two spelling training techniques that you experienced last week: (circle one number)

Testing with feedback						Copying
-3	-2	-1	0	+1	+2	+3

2. On a scale of 1 to 5 (weakest to strongest), please rate how effective you think the following techniques are for learning spelling: (circle one number for each technique)

	Not effective				Very effective
Testing with feedback	1	2	3	4	5
Copying	1	2	3	4	5

Experiment 4:

Last week, you were taught to spell vocabulary words using the following two techniques:

- **Testing with feedback** - any form of quizzes or tests, and getting the correct answers
- **Reading** – speaking words out loud

Questions 1-2, not shown here, follow the same format as the questions used in the preceding experiments above, but with Reading in place of Copying.

Note: In each experiment, subjects were randomly administered one of two surveys, the sole difference being the order of which condition was consistently presented first (e.g., *Testing* vs. *Copying*, or *Copying* vs. *Testing*). In Experiment 1, the word 'relative' and a rating scale was not used in question 1 (subjects chose one technique over another) and question 2 was omitted. In Experiments 1-3, the term *Repetitive Writing* was used in place of *Copying.*