Title
Investigations of Retrieval Practice and Transfer of Learning for Rearranged Stimulus-Response Elements

Author
Pan, Steven C.

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Investigations of Retrieval Practice and Transfer of Learning for Rearranged Stimulus-Response Elements

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Experimental Psychology

by

Steven C. Pan

Committee in charge:

Professor Timothy Rickard, Chair
Professor James Cooke
Professor Tamar Gollan
Professor Craig McKenzie
Professor Hal Pashler
Professor John Wixted

2018

The Dissertation of Steven C. Pan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

DEDICATION

To my family – Li, Wen, and Jeffrey – whose generous support and loving understanding has been an invaluable source of strength all these years.

EPIGRAPH

Exercise in repeatedly recalling a thing strengthens the memory.

— Aristotle, *De Memoria et Reminiscentia*

A curious peculiarity of our memory is that things are impressed better by active than by passive repetition. I mean that in learning by heart (for example), when we almost know the piece, it pays better to wait and recollect by an effort from within, then to look at the book again. If we recover the words in the former way, we shall probably know them the next time; if in the latter way, we shall very likely need the book once more.

— William James, *Principles of Psychology*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Tim Rickard, for his wonderful support and guidance throughout my graduate career. I have learned a great amount from your thoughtful teachings, wisdom, and outstanding example – and am a better scientist for it. You have my everlasting thanks and appreciation!

Thanks also to my committee members – James Cooke, Tamar Gollan, Craig McKenzie, Hal Pashler, and John Wixted – for your guidance. All of you have provided helpful input on this dissertation and have assisted me in developing further aspects of my research career. I am deeply grateful for your assistance.

This dissertation benefited both directly and indirectly from my fellow lab members' efforts. A special thanks to Jarrett Lovelett for his encouragement and generous support; it has been a privilege to be your colleague. Thanks also to John Barry, Daanish Unwalla, Dominic D'Andrea, and Andrew Nguyen, Senior Research Coordinators; Jessica Osuna and Gayathri Sridhar, Assistant Research Coordinators; Sarah Hutter, Carol Wong, and Ben Rubin, Honors Students; and Anastasia Bogozova, Annie Vasishta, Arpita Gopal, Bijan Malaklou, Carissa Gaw, Charles Dupont, Christine Diao, Courtney Lukitsch, Cullin Taggard, Danielle Emmar, Darian Parsey, Ikjot Thind, Jahan Tajran, Jason Chang, Jessica Osuna, Jianyi Liu, Jonathan Clausen, Jonathan Mejia, Joshua Lozano, Kaiqi Guo, Kayla Hartman, Kellie King, Max Bishop, Max Deinitchenko, Megan Iida, Michelle Lauren, Milo Chan, Muzi Chen, Natalia Cameroni, Peter Wong, Minnie Lin, Shaheen Modir, Thomas Chi, Thomas Ting, Trinity Carwile, Vicky Phun, Yangyang Liu, Zachary Potter, and other Research Assistants.

Thanks also to my friends and colleagues at University of California San Diego. From my fellow graduate students and professors to Rachael Lapidis and other treasured members of the staff, thank you for making my experience here a warm and welcoming one. It has been a privilege to be part of this program.

Finally, a million thanks to my parents, Li and Wen, my brother Jeffrey, and my godmother Kay for your kindness, understanding, and everlasting support. You inspired me to pursue my dreams and helped me attain them.

Chapter 1, in full, is a reprint of material as it appears in Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., and Rickard, T. C, Does test-enhanced learning transfer for triple associates?, *Memory & Cognition 44*(1), 24-36, 2016. The dissertation author was the primary investigator and author of this paper. Permission for use of this material has been obtained from Springer.

Chapter 2, in full, is a reprint of material as it appears in Pan, S. C., Gopal, A., and Rickard, T. C., Testing with feedback yields potent, but piecewise, learning of history and biology facts, *Journal of Educational Psychology 108*(4), 563-575, 2016. The dissertation author was the primary investigator and author of this paper. Permission for use of this material has been obtained from the American Psychological Association (APA). Per APA requirements, it should be noted that: this chapter may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association. For further details, please see the acknowledgments section of that chapter.

Chapter 3, in full, is a reprint of material as it appears in Pan, S. C. and Rickard, T. C., Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts?, *Journal of Experimental Psychology: Applied 23*(3), 278-292, 2017. The dissertation author was the primary investigator and author of this paper. Permission for use of this material has been obtained from the American Psychological Association (APA). Per APA requirements, it should be noted that: this chapter may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association. For further details, please see the acknowledgments section of that chapter.

Chapter 4, in full, has been submitted for publication as Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., and Rickard, T. C., In search of transfer following cued recall practice: the case of biology concepts. The dissertation author was the primary investigator and author of this paper.

This research was supported by a National Science Foundation Graduate Research Fellowship and an American Psychological Association Early Graduate Student Researcher Award. Additional funding support was provided by the American Psychological Association, American Psychological Foundation, Psi Chi International Honour Society, and Psychonomic Society.

VITA

2002-2004 National Institute of Mental Health, Career Opportunities in Research, Education, and Training Fellow

2004 Bachelor of Arts in Psychology, Minor in Biology, California State University Los Angeles

2014 Master of Arts in Psychology, University of California San Diego

2014-2018 National Science Foundation Graduate Research Fellow, University of California San Diego

2018 Doctor of Philosophy in Experimental Psychology, University of California San Diego

2018-2019 Chancellor’s Postdoctoral Fellow, University of California Los Angeles

PUBLICATIONS

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., and Rickard, T. C. (submitted). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills.

Pan, S. C., Hutter, S., D’Andrea, D., Unwalla, D., and Rickard, T. C. (submitted). In search of transfer following cued recall practice: the case of biology concepts.

Pan, S. C., Lovelett, J., Stoeckenius, D., and Rickard, T. C. (submitted). Specificity of learning following cued recall does not extend to pretesting.

Rickard, T. C., and Pan, S. C. (submitted). Retrieval practice on pairs, triplets, and facts: when and why does transfer occur?

Pan, S. C., and Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin 144*(7), 710-756. doi:http://dx.doi.org/10.1037/bul0000151

Rickard, T. C., and Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review 25*(3), 847-869. doi:http://dx.doi.org/10.3758/s13423-017-1298-4

Pan, S. C., and Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied 23*(3), 278-292. doi:http://dx.doi.org/10.1037/xap0000124

Rickard, T. C., and Pan, S. C. (2017). Time for considering the possibility that sleep plays no unique role in motor memory consolidation: reply to Adi-Japha and Karni (2016). *Psychological Bulletin 143*(4), 454-458. doi:http://dx.doi.org/10.1037/bul0000094

Pan, S. C., Gopal, A., and Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology 108*(4), 563-575. doi:http://dx.doi.org/10.1037/edu0000074

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., and Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition 44*(1), 24-36. doi:http://dx.doi.org/ 10.3758/s13421-015-0547-x

Pan, S. C., Rubin, B. R., and Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4), 356-369. doi:http://dx.doi.org/10.1037/xap0000062

Pan, S. C. and Rickard, T. C. (2015). Sleep and motor memory: is there room for consolidation? *Psychological Bulletin 141*(4), 812-834. doi:http://dx.doi.org/10.1037/bul0000009

Pan, S. C., Pashler, H., Potter, Z. E., and Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language 83,* 53-61. doi:http://dx.doi.org/10.1016/j.jml.2015.04.001

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 27*(3), 385-400. doi:http://dx.doi.org/10.1007/s10648-015-9330-6

Walker, D., Pan, S. C., Modir, S., and Rickard, T. C. (2014). Sleep enhancement of texture discrimination performance is dependent on training paradigm. *Journal of Vision 14*(10). doi:http://dx.doi.org/10.1167/14.10.664

ABSTRACT OF THE DISSERTATION

Investigations of Retrieval Practice and Transfer of Learning for
Rearranged Stimulus-Response Elements

by

Steven C. Pan

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2018

Professor Timothy C. Rickard, Chair

Retrieval practice is a learning technique which involves taking practice tests on to-be-recalled information. Because it is one of the most potent memory enhancers known to learning science, many cognitive and educational psychologists endorse retrieval practice as a highly effective method for enhancing academic performance. However, retrieval practice's capacity to enhance *transfer of learning*, or the generalization of learning to different contexts, has yet to be fully characterized. Given that learning often needs to be flexibly applied, the importance of transfer to education is paramount. In the current research, four studies investigated retrieval practice and transfer to *rearranged stimulus-response elements*. With this type of transfer, practice and criterial tests target

the same basic information, such as a specific fact or concept, but feature different cues and responses. For example, given the fact "Thomas Jefferson purchased Louisiana from France", a practice test asks, "Thomas Jefferson purchased *what* from France?", whereas a criterial test asks, "*Who* purchased Louisiana from France?" The chief question was whether learning would transfer—as evidenced by better accuracy—to criterial test questions featuring rearranged stimulus-response elements. Studies 1-4 used triple associate words, Advanced Placement History and Biology course facts, term-definition neuroscience facts, and process-based biology concepts, respectively. Throughout these studies, retrieval practice consistently improved accuracy when cues and responses were consistent on practice and criterial tests, relative to a non-retrieval practice reexposure control condition. However, substantially reduced improvements were observed when the criterial test featured rearranged stimulus-response elements. This generally occurred regardless of the amount of, and the format of, retrieval practice. Only the addition of detailed and extensively processed feedback, as was investigated in Study 4, yielded exceptions to the prevailing pattern of minimal-to-no transfer. Overall, the present research yields insights into the transfer properties and mechanisms of retrieval practice, including its often selective enhancement of trained cue-response combinations. It also provides support for dual memory theories of retrieval practice effects. Practically, this research suggests that retrieval practice should be targeted at as many cue-response combinations as possible during training, or should incorporate extensively processed feedback, in order to yield maximum learning benefits.

INTRODUCTION

Investigations of Retrieval Practice and Transfer of Learning for Rearranged Stimulus-Response Elements

Over a century of research on the effects of practice testing, also known as *retrieval practice* or *test-enhanced learning*, has shown that it often generates powerful learning benefits (for reviews see Rickard & Pan, 2017; Roediger & Butler, 2011; Rowland, 2014). Specifically, after initial study of a set of to-be-learned information (e.g., word pairs, facts, concepts, or text passages), the use of retrieval practice to train on that information typically results in improved recall on a delayed criterial test (also called a final test). That improvement is commonly observed relative to a control condition in which an equivalent amount of training time is spent on non-testing activities such as restudy or rereading. That sequence of events (Phase 1: initial study; Phase 2: retrieval practice vs. non-testing control; Phase 3: delayed criterial test) constitutes the standard *retrieval practice paradigm* or *testing effect paradigm*. Approximately 300 studies to date have successfully used that paradigm, or close variants thereof, to demonstrate the benefits of retrieval practice on materials ranging from textbook chapters to maps (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rawson & Dunlosky, 2011; Rowland, 2014). The terms *testing effect* and *retrieval practice effect* are commonly used to describe the memory enhancement that results from the use of retrieval practice.

Current Retrieval Practice Research

Interest in the potential applicability of retrieval practice for enhancing education and training, as well as its underlying theoretical mechanisms, has grown dramatically in the past decade (Roediger & Karpicke, 2018). From the standpoint of application, many cognitive and educational psychologists now regard retrieval practice as among the most potent of all evidence-based learning techniques (e.g., Dunlosky et al., 2013; Pashler, et al, 2007). Accordingly, a wide range of potential uses of retrieval practice have been suggested for different age groups and for a variety of subject materials (e.g., Agarwal, Roediger, McDaniel, & McDermott, 2013; Pan & Agarwal, 2018). Numerous theories of retrieval practice effects have also been proposed. These range from *retrieval effort* (Bjork, 1975) and *distribution-of-memory-strength* accounts (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011) to *error correction* (Mozer, Howe, & Pashler, 2004), *elaborative retrieval* (Carpenter & DeLosh, 2006), *mediator effectiveness* (Pyc & Rawson, 2010), *episodic context* (Karpicke, Lehman, & Aue, 2014), and *dual memory* models (Rickard & Pan, 2017).

As the field of retrieval practice research has grown, it has branched into an increasingly wide range of subtopics. This includes investigations of its applicability in new domains (e.g., spelling as in Jones et al., 2015; Pan, Rubin, & Rickard, 2015); the potential role of individual differences in memory ability, strategy use, and item difficulty in moderating its effects (e.g., Minear, Coane, Boland, Cooney, & Albat, 2018; Pan, Pashler, Potter, & Rickard, 2015); its ability to enhance subsequent learning from feedback or of new materials (e.g., Arnold & McDermott, 2013; Szpunar, McDermott, & Roediger, 2008), and its capacity to yield transfer of learning (e.g., Butler, 2010; Rohrer, Taylor, & Sholar, 2010). That final research type has grown to become the largest subtopic in the current literature and is the focus of this dissertation.

Retrieval Practice and Transfer of Learning

Transfer of learning, or simply *transfer*, can be defined as the successful generalization of prior learning that occurred in one context in a different context (and, per Perkins & Salomon, 1994, such contexts should involve more than that which would be regarded as ordinary learning). In contrast with other types of retrieval practice research, studies of retrieval practice and transfer use practice and criterial test questions that are not identical to one another. Instead, the questions vary on one or more dimensions, such as different test formats (e.g., multiple-choice vs. short answer as in Kang, McDermott, & Roediger, 2007), assessments of related information (e.g., Little, Bjork, Bjork, & Angello, 2012), application or inference questions (e.g., Butler, 2010), new problem-solving scenarios (e.g., Larsen, Butler, & Roediger, 2013), generating new directional responses on visuospatial tasks (e.g., Rohrer, Taylor, & Sholar, 2010), classifying new exemplars of learned categories (e.g., Jacoby, Wahlheim, & Coane, 2010), and more. The transfer issue is critical to the efficacious application of retrieval practice given that (a) instructors eschew reusing practice questions on criterial tests (e.g., Balch, 1998), and (b) in real world circumstances one cannot expect to always retrieve the same exact information. Research on retrieval practice and transfer also has implications for the cognitive mechanisms that underlie retrieval practice's memorial benefits.

Pan and Rickard (2018) recently performed a comprehensive meta-analysis of the literature on retrieval practice and transfer. That meta-analysis found that, across 192 effect sizes drawn from 122 experiments in 67 published and unpublished studies, retrieval practice did yield positive transfer (Cohen's d = 0.40). That transfer was moderated by three main factors: the level of *initial test performance* (i.e., the degree of retrieval success on the practice test), whether there was *response congruency* or not (i.e., whether the correct answers on the practice and criterial tests were the same or similar), and whether *elaborated retrieval practice* was used or not (i.e., whether retrieval practice involved recalling more than a brief, single answer and/or whether it was accompanied by various forms of feedback that involved more than just correct answers, or not). Further, in secondary meta-analyses limited to specific transfer types (e.g., application and inference questions only, or transfer across test formats only), weighted mean effect sizes varied substantially; robust transfer was observed for some types and not others (ranging from $d = 0.16$ to 0.61). That meta-analysis included experiments from all four studies that comprise this dissertation, all of which involved the transfer type that is described next.

Retrieval Practice and Rearranged Stimulus-Response Elements

In this dissertation, retrieval practice and transfer to *rearranged stimulus-response elements* was investigated. With this type of transfer, the same overall information is targeted on practice and criterial tests, but different cues and different responses are involved on each test. For example, given a triple associate (e.g., *gift, rose, wine*), retrieval practice might occur on one word of that triple associate (e.g., *gift*, *rose, ?*), whereas a criterial test might require recall of a different, previously untested word from that triple associate (e.g., *gift, ?*, *wine*). The issue of interest here is whether there will still be a benefit of retrieval practice, relative to a restudy control, on that criterial test. In another example, given a neuroscience term-definition fact (e.g., "*Synaptogenesis is the formation of synapses between neurons*"), a practice test might provide a concept term (e.g., *synaptogenesis*) as a cue to retrieve other parts of that fact (e.g., "*What is the definition of synaptogenesis?*"), whereas the reverse might occur on a criterial test (e.g., "*The formation of synapses between neurons is known as…?*"). The same issue, namely whether there will still be a retrieval practice benefit on the criterial test, is again the focus in that example.

Prior to the present research, published studies on retrieval practice and transfer to rearranged stimulus-response elements consisted of two lines of research. The first involved word pairs (e.g., Carpenter, Pashler, & Vul, 2006; Vaughn & Rawson, 2014), in which retrieval practice on one word of a word pair was shown to generate successful transfer to the second word on a criterial test (e.g., given the word pair *beach, blanket*, retrieval practice on *beach, ?* results in better performance on *?, blanket*, relative to restudy). The second involved classroom studies of fact learning (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013) in which repeated practice tests in conjunction with outside study activities or extended and detailed feedback was shown to promote successful transfer to criterial test questions featuring various forms of rearranged stimulus-response elements (e.g., recall of different key terms from a given fact).

Those two lines of research represent contrasting points in the space of possible scenarios under which transfer and rearranged stimulus-response elements can be investigated. In the first, the to-be-learned materials apparently have a fairly simple associative structure (e.g., $A \rightarrow B$) and relatively basic methods of retrieval practice have been assessed in controlled laboratory conditions; in the second, the to-be-learned materials apparently have a more complex structure (e.g., facts comprised of associations ranging from $A \rightarrow B \rightarrow C$ to *A given B but not C*, etc.) and multiple sessions of retrieval practice have been implemented in conjunction with pre- and post-retrieval activities and in less controlled classroom settings. Additionally, a restudy control was employed in the former but not always in the latter (for related discussion on the importance of the reference condition in retrieval practice research, see Carrier & Pashler, 1992).

Overview of the Present Research

This dissertation is comprised of four sets of studies on retrieval practice and transfer to rearranged stimulus-response elements. A separate chapter is devoted to each. These are: *triple associates* (Chapter 1), *multi-term facts* (Chapter 2), *term-definition facts* (Chapter 3), and *process-based concepts* (Chapter 4). Besides differences in stimulus types, these studies differed in terms of the subject domain from which to-be-learned materials were drawn (from words to Advanced Placement History and neuroscience facts), retention intervals (from 24 hrs to one week), amounts of information that had to be retrieved during training, and implementations of retrieval practice. During training, retrieval practice in the form of cued recall (i.e., fill-in-the-blank or short answer) or multiple-choice was commonly used, and always in conjunction with correct answer feedback (which bolsters the benefits of retrieval practice, e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005), and in some cases more than the correct answer was provided during feedback. Collectively, these studies were designed to comprehensively address the question of whether retrieval practice yields transfer to rearranged stimulus-response elements, and under what conditions such transfer might be obtained.

References

Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2013). *How to use retrieval practice to improve learning.* Teachers' and students' practice guide. Available at: http://www.retrievalpractice.org/library/

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 940.

Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology, 25*(3), 181-185.

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale: Erlbaum.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36(5), 1118-1133. doi:http://dx.doi.org/10.1037/a0019902

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:http://dx.doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review,* 13(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

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

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:http://dx.doi.org/10.1177/1529100612453266

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 37(4), 801-812. doi:http://dx.doi.org/10.1037/a0023219

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), 1441-1451. doi:http://dx.doi.org/10.1037/a0020636

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 27*(3), 385-400. doi: http://dx.doi.org/10.1007/s10648-015-9330-6

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4-5), 528-558.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *The psychology of learning and motivation (vol 61).* (pp. 237-284). San Diego: Elsevier Academic Press.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85-97. doi:http://dx.doi.org/10.1016/j.jml.2011.04.002

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education, 47*(7), 674-682. doi:http://dx.doi.org/10.1111/medu.12141

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, 23(11), 1337-1344. doi:http://dx.doi.org/10.1177/0956797612443370

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:http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, 27(3), 360-372. doi:http://dx.doi.org/10.1002/acp.2914

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of experimental psychology. Learning, Memory, and Cognition.* doi:http://dx.doi.org/10.1037/xlm0000486

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

Pan, S. C. & Agarwal, P. K. (2018). *Retrieval practice and transfer of learning: fostering students' application of knowledge*. Teachers' and students' practice guide. Available at: http://www.retrievalpractice.org/library/

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language 83,* 53-61. doi:http://dx.doi.org/10.1016/j.jml.2015.04.001

Pan, S. C., and Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin 144*(7), 710-756. doi: http://dx.doi.org/10.3758/s13423-017-1298-4

Pan, S. C., Rubin, B. R., & Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4). doi:http://dx.doi.org/ 10.1037/xap0000062

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:http://dx.doi.org/10.1037/0278-7393.31.1.3

Perkins, D. N., & Salomon, G. (1994). Transfer of learning. In T. Husen & T. N. Postelwhite (Eds.). *International Handbook of Educational Research* (Second Edition, Vol. 11; pp. 6452-6457). Oxford: Pergamon Press.

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

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283-302. doi:http://dx.doi.org/10.1037/a0023956

Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review 25*(3), 847-869. doi: http://dx.doi.org/10.3758/s13423-017-1298-4

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:http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2018). Reflections on the resurgence of interest in the testing effect. *Perspectives on Psychological Science, 13*(2), 236-241. doi: https://dx.doi.org/10.1177/1745691617718873

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36(1), 233-239. doi:http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

Szpunar, K. K., McDermott, K. B., & Roediger III, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392.

Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language*, *75*, 14-26. doi:http://dx.doi.org/10.1016/j.jml.2014.04.004

CHAPTER 1

Does Test-Enhanced Learning Transfer for Triple Associates?

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., and Rickard, T. C. (2016).

Memory & Cognition 44(1), 24-36.

Abstract

Test-enhanced learning and transfer for triple associate word stimuli was assessed in three experiments. In each experiment, training and criterial test trials involved the presentation of two words per triple associate (triplet), with the third word having to be retrieved. In agreement with the prior literature on different stimuli, training through testing with feedback yielded markedly better criterial test performance than did restudy. However, in contrast to the positive transfer reported for paired associate stimuli, minimal or no positive transfer was observed, relative to a restudy control, from a trained cue combination (e.g., *A*, *B*, *?*) to other cue combinations from the same triplet that required a different response (e.g., *B*, *C*, *?*). That result also held when two unique cue combinations per triplet were tested during training, and for triplets with low and high average associative strength. Supplementary analyses provided insight into the overall transfer effect. An incorrect response during training appears to yield positive transfer relative to restudy, whereas a correct response appears to yield no or even negative transfer. Cross-experiment analyses indicate that test-enhanced learning is not diminished when two or three cue combinations are presented during training. Thus, even though learning through testing is highly specific, testing on all possible stimulus-response combinations remains

the most efficient strategy for the learning of triple associates.

Introduction

The act of recalling information during a test, otherwise known as *retrieval practice*, enhances memory for that information, above and beyond an equivalent period of time spent restudying the same materials. This *testing effect*, or *retrieval practice effect*, has been replicated numerous times with stimuli ranging from paired associates to prose passages (Roediger & Karpicke, 2006a), and is widely regarded as one of the most robust phenomena in learning science (Butler, 2010; Carpenter, 2012). However, a limitation of the testing effect literature to date is that the vast majority of published studies have used identical materials during the initial and criterial tests (for discussion see McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). Comparatively less is known about testing's effectiveness for transfer to novel contexts.

Alongside retention, transfer has been described as the ultimate objective of learning (e.g., Carpenter, 2012; Rohrer, Taylor, & Sholar, 2010). Instructors hope that the information that they impart to students will be flexibly accessible and will generalize to different contexts. Therefore, it is crucial to determine the extent to which test-enhanced learning transfers and whether it has limitations (Anderson & Biddle, 1975; Hinze & Wiley, 2011). In a recent review of transfer and the testing effect literature, Carpenter (2012) concluded that test-enhanced learning transfers well in three categories: (1) across temporal contexts (e.g., different retention intervals), (2) across test formats (e.g., question format alterations such as short-answer to multiple-choice), and (3) across knowledge domains (e.g., from biology to aeronautics). This classification scheme adheres to recommendations that research on transfer requires definitional clarity (Barnett & Ceci, 2002).

Of the categories delineated by Carpenter (2012), the second and third include studies in which the stimulus materials used on a criterial test are different from those used on prior tests; among these, transfer has been demonstrated for word lists (Carpenter & DeLosh, 2006), prose passages (Butler, 2010; Chan, 2009; 2010; Chan, McDermott, & Roediger, 2006; Karpicke & Blunt, 2011), and science facts (Hinze, Wiley, & Pellegrino, 2013; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel et al., 2013; Rawson, Dunlosky, & Sciartelli, 2013). Test-enhanced learning also exhibits transfer for less commonly examined materials such as math functions (Kang, McDaniel, & Pashler, 2011), spatial learning (Carpenter & Kelly, 2012), map learning (Rohrer et al., 2010), and even medical diagnoses (Larsen, Butler, Lawson, & Roediger, 2013). Moreover, transfer has been found when the difference between the originally tested and transfer test materials is relatively small (*near* transfer involving questions on the same subject, e.g., detail questions as in Hinze et al., 2013) as well as when it is quite large (*far* transfer involving application or inference questions on separate topics, e.g., Chan et al., 2006; Johnson & Mayer, 2009; McDaniel, Howard, & Einstein, 2009).

Conversely, there have been some failures to find transfer of test-enhanced learning (e.g., Agarwal, 2011; Hinze & Wiley, 2011; Tran, Rohrer, & Pashler, 2014). Such studies have employed materials similar to those in the studies described above (e.g., textbook passages, science facts, logical premises) and the transfer measured has ranged from near to far. As of this writing, however, failures to find transfer of test-enhanced learning remain the exceptions in the literature.

Positive Transfer for Paired Associates

Carpenter's (2012) second category of transfer also includes studies in which the same materials are presented on the initial and criterial tests but in which the stimulus and response elements have been rearranged on the criterial test. Strong positive transfer of that type had been observed for paired associate words. Carpenter, Pashler, and Vul (2006) administered cued recall tests for paired-associate English word lists in one cue-to-target direction (e.g., *beach* → *blanket*), and observed substantial transfer to the reversed, previously untested direction (e.g., *blanket* → *beach*) on a delayed criterial test. Similarly, Vaughn and Rawson (2014) adapted the Carpenter et al. paradigm using English-English word pairs to assess criterion-level testing (cf. Vaughn & Rawson, 2011). They replicated the finding of positive transfer for the case of subjects trained to a criterion level of one correct trial for each tested item, although the extent of transfer decreased at higher levels of criterion learning.

In Carpenter et al. (2006) and Vaughn and Rawson (2014), transfer was assessed on the criterial test by the reversal of two elements (e.g., *cue* → *target* reversed to *target* → *cue*). This raises an important yet unaddressed question in the testing literature: Does test-enhanced learning transfer for stimuli with more than two elements, any of which could constitute the answer on the transfer test? This question has not been thoroughly addressed within the testing effect paradigm using accuracy as the dependent variable, in particular for the type of transfer in which previously studied stimulus and response elements are rearranged on the criterial test. There is related research, however, as summarized below.

Transfer Research on Triple Associates

Kahana and Caplan (2002; Experiment 1) explored transfer for word triplets using serial presentation in which each of three words was presented, one at a time, per study trial (all three words were never shown at one time). That study phase was immediately followed by a criterial test in which either a single word or two words were presented, with subjects being asked to vocalize the missing word or words. A recall advantage, in terms of both accuracy and response time (RT), was observed both for single word and two word stimuli in the trained forward direction over the untrained reverse direction (i.e., performance on *A-B-?* was faster and more accurate than performance on *?-B-C*). Kahana and Caplan regarded this finding as evidence of associative asymmetry for the case of triplets, and noted its contrast to the associative symmetry that is characteristic of paired associates (Kahana, 2002). That study did not utilize a testing effect paradigm, however (i.e., training did not involve a test), and the positive transfer was observed relative to no training. Thus, that finding does not directly address the possibility of transfer for tested materials relative to restudied materials.

Another line of work has investigated the transfer of test-enhanced learning for multi-element stimuli following extended retrieval practice. In these studies, transfer was evaluated relative to a control case of no training and using RT as the dependent measure. Rickard, Healy, and Bourne (1994; see also Bajic, Kwak, & Rickard, 2011; Rickard, 2005; Rickard & Bourne, 1996) showed that, for both children and adults, RT gains following retrieval practice on arithmetic facts do not transfer to complement facts (e.g., the RT improvement for $5 \times 7 =$ __ does not transfer positively to $35 \div 7 =$ __, or vice versa). Of more direct relevance to the present study, the same lack of transfer following extended retrieval practice has also been demonstrated for multi-element verbal stimuli.

Rickard and Bajic (2006) had subjects study a set of 24 word triplets (e.g., *cup, hand, tea*). Each triplet could form three test items, wherein a *test item* refers to a two-element test stimulus in which two of the three words are presented as cues and the missing third word is the to-be-recalled target; ignoring the spatial arrangement of the presented words, three test items are possible for each triplet (e.g., for the triplet *cup, hand, tea* the three test items are: test item 1: *cup, hand, ?*; test item 2: *cup, tea, ?*; test item 3: *hand, tea, ?*). In the subsequent training phase, there were 20 repetitions of testing with feedback for 18 test items from the 36 triplets, and the remaining triplets received no training. On the criterial test, RT performance was assessed for all three possible test items per triplet, which yielded three conditions: (1) tested triplets with the trained response (*tested-identical* items), (2) tested triplets with an untrained response (*tested-inverted* items), and (3) triplets that were not trained at all (*untrained* items). Tested-identical items exhibited more than 650 ms of speedup. That learning, however, did not transfer positively to tested-inverted items; there was no difference in accuracy across the criterial test conditions in those experiments (by design, accuracy was high for all test items on the criterial test).

Specificity of learning and testing on triple associates. To explain the specificity of learning in the above RT studies, Rickard et al. (1994; Rickard & Bajic, 2006) proposed an identical elements (IE) model of retrieval practice effects for multi-element stimuli. According to that model, successful retrieval through a study representation establishes a new and separate recall representation for the particular stimulus-response configuration that was tested. That representation, which is assumed to form over the first few successful retrieval attempts, is accessible only for trained stimulus-response combinations. Over the course of retrieval practice, the recall representation becomes the dominant pathway to retrieval, and faster retrieval through that representation as a function of repetition accounts for the RT speedup. Thus, retrieval practice for the test item *cup, tea, ?* will result in substantial RT gains. However, when the tested-inverted item *tea, hand, ?* is later presented, the recall representation for *cup, tea, ?* cannot, according to the IE model, be accessed. Rather, answer retrieval for that test item must rely solely on the initially encoded study representation. The predicted and observed result is no positive transfer of practice effects to tested-inverted items. The IE model also correctly predicts the finding that RT improvement during the transition from reliance on the study representation to reliance on the hypothesized recall (test-based) representation does not follow the nearly ubiquitous power law (Newell & Rosenbloom, 1981) for data averaged over subjects, but rather adheres to a mixture function that also characterizes RT learning in other strategy transition contexts (Rickard, 1997; Rickard & Bajic, 2006).

The studies on multi-element stimuli summarized above and the IE model raise the possibility that transfer of learning will not occur for triplets under the testing effect paradigm (i.e., assessing transfer in the context of performance on tested vs. restudied materials), despite evidence for positive transfer in that paradigm for paired associates. It is important to emphasize, however, that the aforementioned experiments on word triplets and arithmetic facts differ from those in the testing effect literature in at least four critical respects that may yield divergent transfer results. First, the focus in most of those studies was transfer of learning as measured by RTs, whereas testing effect studies primarily use accuracy as the dependent measure. It is plausible that RT gains are highly specific to practiced test items whereas accuracy gains are not. Second, those studies either included no comparison of testing versus restudy (e.g., Rickard & Bajic, 2006), or testing was entirely absent during training (Kahana & Caplan, 2002). It is possible that transfer may be observed in a testing effect paradigm that involves testing and restudy during training. Third, training in several of the aforementioned studies (e.g., Rickard et al., 1994; Rickard & Bajic, 2006) involved numerous repetitions (e.g., 20) of each trained test item, on which the observed RT-based specificity of learning effects might have depended. In the present experiments, transfer was assessed after only one test trial per test item. Finally, all experiments described above involved a test immediately after training. In the testing effect literature, longer delays, such as the one-week delay in the present experiments, have been associated with increased testing effect magnitude (Roediger & Butler, 2011; Roediger & Karpicke, 2006a). It is possible that longer delays also yield greater transfer.

The Current Study

In the present work we examined test-enhanced learning and transfer for triple associates, relative to a restudy control. In session 1, subjects studied a complete set of 36 triplets. During the subsequent training phase they restudied half of the triplets and were tested with feedback on the others. The criterial test occurred one week after training, and the dependent variable was accuracy (proportion correct). The critical question was whether the expected test-enhanced learning would transfer from tested-identical items to tested-inverted items (where a *test item* refers to a two-element test stimulus that takes the form of two presented words with one missing word to be recalled, as in Rickard & Bajic, 2006). To assess criterial test performance on all three possible test items that were derivable from each triplet (i.e., *A, B, ?*; *A, C, ?*; and *B, C, ?*; the spatial order of elements was randomized), we used a three-block design. In each block, one test item (e.g., *A, B, ?*) per triplet was presented, and across all three blocks each of the three test items (e.g., test item 1: *A, B, ?*; test item 2: *A, C, ?*; test item 3: *B, C, ?*) corresponding to each triplet was presented once.

To explore possible moderating factors on both the magnitude and transferability of the testing effect, in the first two experiments we examined the effects of testing on one versus two test items from each triplet during training. In Experiment 3, we assessed the effects of training on all three test items per triplet during training (thus complementing the first two experiments but not directly assessing transfer), as well as an instructional manipulation. The modifications across experiments were motivated in part by prior suggestions that testing on more than one variant of a stimulus (e.g., Goode, Geraci, & Roediger, 2008) and employing instructional manipulations (e.g., Chan et al., 2006) can enhance the transferability of the testing effect.

Experiment 1

Subjects first studied 36 word triplets. Each triplet was then presented once again during training, half for restudy and half for testing with feedback. In this and all other experiments, all test trials during both training and the criterial test always involved presentation of two of the three words as the stimulus, with the third word having to be recalled. As we mentioned in the introduction, a given pair of word stimuli on a test trial is referred to as a *test item* (i.e., two of three words being presented, with the missing third word to be retrieved, constitutes one test item). After a one-week delay, a criterial test assessed recall for all 108 possible test items across the 36 studied triplets (three blocks, with one test item per triplet being presented in each block). The overall design was modeled closely on prior work (Carpenter et al., 2006; Rickard & Bajic, 2006).

Method

Subjects. Forty-six University of California San Diego undergraduate students participated for course credit. All but four of the subjects completed both sessions of the experiment. The sample size selected for this study was comparable to those of prior laboratory studies focused on transfer (e.g., Hinze & Wiley, 2011; McDaniel et al., 2007).

Materials. The stimuli consisted of 36 triplets, each containing three words of 3-7 letters and 1-2 syllables in length (see the Appendix 1A). All triplets were designed to facilitate the formation of an interactive mental image (e.g., *gift, rose, wine*).

Design and procedure. In each session, subjects were individually seated at personal computers running Windows 7 or Windows XP (Microsoft, Redmond, WA) at a viewing distance of 30-40 cm from the computer screen. All experiments were programmed using Adobe Flash Professional CS6 (Adobe Systems, San Jose, CA) and presented using the Mozilla Firefox web browser (Mozilla Foundation, Mountain View, CA) equipped with the Adobe Flash Player 12 plugin.

The first session consisted of two phases:

In the study phase, subjects read instructions stating that they were to "*learn and* *memorize three concepts (words) at a time*," and for each presented triplet to "*use your imagination to visualize how the three concepts might interact with one another*." After pressing the spacebar to proceed, they were shown all 36 triplets one at a time for 8 seconds each. The order of presentation was random and there was no delay between trials. Each triplet appeared in columnar fashion (one word above the other), in a large serif font (Garamond, size 40) at the center of the screen, with the instructions "*study this triplet*" in smaller font underneath. The ordering of each word per triplet (top, middle, or bottom) was randomized on each trial for each subject.

In the training phase, subjects were tested with feedback on one test item from each of 18 randomly selected triplets and restudied each of the remaining 18 triplets, for a total of 36 trials within one uninterrupted block. On each test trial, one test item from a triplet was shown (e.g., *wine*, *gift*, *?*); on each restudy trial, all three words from a triplet were shown. Test and restudy trials were randomly interleaved. At the beginning of the phase, on-screen instructions appeared stating that each of the previously studied triplets would be presented with directions for either restudy (*"you will be asked to study, which gives you more time to memorize"*) or testing with feedback (*"you will be shown two words from a triplet and asked to type the third word"*; *"the correct word will then appear*"). Subjects pressed the space bar to proceed to the first trial. For all trials, the stimulus was presented for 8 s, with no breaks between trials.

On trials involving restudy, triplets were presented in a columnar format identical to that of the study phase. On trials involving testing with feedback, test items also appeared in the columnar format, but with the following changes: (1) one word (the response to be retrieved) was removed and "*???*" appeared in its place (with the word and "*???*" ordering in top, middle, or bottom positions randomized on each trial for each subject); (2) an empty text box with a blinking cursor appeared directly underneath the column; and (3) the instructions stated "*type the third word of this triplet*". Subjects had 6 s to type their answer into the text box, after which no new input was accepted and "*???*" was replaced by the correct word for 2 s, which constituted feedback. During that 2 s period, the two words of the test item and any typed characters continued to be displayed. Hence both restudy and testing with feedback trials lasted for 8 s. At the conclusion of training, the experimenter reminded subjects to return at the same time the following week. To minimize the possibility of practice between sessions, subjects were also told that they would be learning a new set of test items.

The criterial test was administered in the second session. First, subjects were informed that their memory for the triplets that they had learned in the prior week was about to be tested. After pressing the spacebar to proceed, subjects were tested on three 36-trial blocks, with one test item from each of the 36 triplets appearing per block, in random order. Thus, no test item was repeated over blocks. Consider the triplet *ant*, *spray*, *trash*. On the first criterial test block for a given subject, the test item might be *spray*, *ant*, *?*; on the second block: *ant*, *trash*, *?*; and on the third block, *trash*, *spray*, *?*.

Each criterial test trial involved the following: (1) Two words were displayed, while the third was absent and replaced by a "*???*"; (2) the arrangement of those three stimulus components was in an upside-down triangle format with two elements at the same (upper) level, and the third element below. This change in test format from training to criterial test was made in order to minimize the possible influence of word spatial location effects from training (as in Rickard & Bajic, 2006); (3) the locations of the three elements (two words and "*???*") were randomized on each trial to any of the three positions; and (4) an empty text box with a blinking cursor appeared immediately below. Subjects had 15 s to type an answer into the text box before the program automatically advanced to the next trial. No feedback was provided, and there were no breaks between blocks. The experiment ended after subjects had completed all three blocks.

Across the three criterial test blocks, each of the three possible test items for each of the 36 triplets was tested once, yielding 108 test items over 108 consecutive trials (see Figure 1.1). In every block, 18 test items were drawn from triplets that had previously been trained using restudy (*restudied* items), and 18 test items were drawn from triplets that had previously been trained using testing. Six test items from the previously trained triplets were identical to the items that had been presented during training. For example, if the test item *tea, hand, ?* was presented during training, the test item *tea, hand, ?* was presented in the criterial test. Following Rickard and Bajic (2006), those items on the criterial test will be referred to as *tested-identical* items. Twelve criterial test items from the previously trained triplets were inverted (i.e., one or both of the two presented words were not the same as on a previously trained test item). For example, if the test item *tea, hand, ?* was presented during training, the inverted test item, *hand, cup, ?* was presented in the criterial test. Those items on the criterial test will be referred to as *tested-inverted* items. Thus, there were three test item type conditions on each block of the criterial test: tested-identical, tested-inverted, and restudied. Positive transfer of test-enhanced learning would correspond to higher accuracy in the tested-inverted condition than in the restudied condition.

Results and Discussion

Training. The mean accuracy over subjects for the 18 tested triplets was 0.60, SE = 0.032.

Criterial test. A within-subjects factorial Analysis of Variance (ANOVA) was performed on subject-level proportion correct (Figure 1.2) with the factors of Criterial Test Condition (tested-identical vs. tested-inverted vs. restudied) and Block (1 vs. 2. vs. 3). We found highly significant effects of criterial test condition, $F(2, 82) = 18.38$, $p < .001$, $MSE = 0.025$, $\eta_p^2 = 0.31$, and block, $F(2, 82) = 206.01$, $p < .001$, $MSE = 0.024$, $\eta_p^2 = 0.83$, as well as a significant interaction, $F(4, 164) = 3.83$, $p < .005$, $MSE = 0.021$, $\eta_p^2 = 0.085$. The main effect of block corresponds to the overall pattern of improvement during the criterial test, a result that will be discussed further below. The criterial test condition by block interaction reflects the decreasing performance difference among criterial test conditions over successive blocks.

Of greatest interest is the main effect of criterial test condition. Inspection of Figure 1.2 suggests that the training effect is primarily driven by the difference between the tested-identical condition and the other two conditions. A follow-up ANOVA limited to the tested-inverted and restudied conditions confirmed no significant main effect of criterial test condition, $F(1, 41) = 0.41$, $p = .53$, $MSE = 0.021$, $\eta_p^2 = 0.0099$, and no significant interaction with block, $F(2, 82) = 2.12$, $p = .13$, $MSE = 0.011$, $\eta_p^2 = 0.049$. Thus, in marked contrast to the procedurally analogous paired-associate task of Carpenter et al. (2006), test-enhanced learning for triplets appears to transfer minimally, if at all, to tested-inverted items.

The substantial performance improvement over the three criterial test blocks (see Figure 1.2) occurred despite that fact that no feedback was provided. That pattern, which we replicated in Experiments 2 and 3, implies transfer from one test item from a triplet to another across blocks, and thus appears to be inconsistent with the finding of little or no transfer from tested-identical to tested-inverted items within each block. Those two transfer patterns, however, are likely in our view to reflect independent memory phenomena. The lack of transfer to tested-inverted items after the one-week delay clearly reflects the specificity of learning in long-term memory, and that specificity is also likely to drive the more compressed condition effects in blocks 2 and 3. The transfer across the three tested items from a triplet over blocks is more likely to reflect priming, or increased response strength (e.g., Bjork & Bjork, 1992), for individual elements, or perhaps "learning to learn" effects (Postman & Stark, 1967). On the first block, two of the three elements from each triplet were presented as the stimuli on each trial, and those elements always constituted the answers to the other two test items from the same triplet that would be presented on blocks 2 and 3. It is likely in our view that priming of those elements on block 1 made them more available for retrieval on blocks 2 and 3. If priming for both of those stimulus elements was largely saturated on block 1, then the performance improvement due to priming would be expected to be relatively large from block 1 to block 2, and smaller from block 2 to block 3, as was observed.

Experiment 2

Having found no evidence for the transfer of test-enhanced learning after training on one test item per triplet, we next examined the possibility that training on two different test items per triplet could enhance transfer performance on the third test item. It may be that testing on two different test items yields triplet representations that are more flexibly accessible and can support transfer to the third. For example, training on a second test item of a triplet may reactivate memory for the first test item that was presented on the preceding block, which may result in a more integrated, or holistic representation that can promote transfer to a third test item on the criterial test.

Method

Subjects. The sample size was increased in this experiment. Sixty-one University of California San Diego undergraduate students participated for course credit, and all but three subjects completed both sessions of the experiment. On the basis of the standard deviation observed in Experiment 1, a sample size of 58 in this experiment yields statistical power greater than 0.90 to detect a proportion correct advantage of 0.05 or larger in the tested-inverted versus the restudy condition (based on a one-tailed test on the difference scores averaged over criterial test blocks, $\alpha = 0.05$).

Materials, design, and procedure. The design of this experiment was nearly identical to that of Experiment 1, with the primary exception that a second 36-trial training block was added to the training phase. Eighteen of the trials in the second block involved a second restudy opportunity for the 18 triplets restudied in the first block. The remaining eighteen trials in the second block involved testing on a different test item taken from the same triplet that had been tested in the first block (i.e., if *wine*, *gift*, *?* was the test item in the first block, then either *rose*, *gift*, *?* or *wine*, *rose*, *?* was the test item in the second block). The position of each word (and "*???*" for test trials) remained randomized on each trial.

As in Experiment 1, there were three criterial test blocks, each with 36 trials, such that all 108 possible test items were tested once over blocks. As a consequence of the changes to the training phase, however, each criterial test block included 12 tested-identical items, six tested-inverted items, and 18 restudied items. Of the 12 tested-identical items per block, half each were from the first and second training blocks. The order of presentation remained randomized within each block.

The criterial test was also modified to feature no imposed time limit on each trial. This change avoided instances of subjects being in the midst of typing when a trial advanced, which occurred on a few occasions in Experiment 1. Subjects were instructed to press the Enter key after they had finished typing their answer on each trial. To prevent skipped trials, the responsiveness of the Enter key on each trial was conditional on either (1) subjects having typed a minimum of three letters, or (2) three s of time having elapsed.

Results and Discussion

Training. The mean accuracies for tested items were 0.73, SE = 0.059 (first block) and 0.75, SE = 0.057 (second block). No significant difference in performance was apparent on the two blocks, $t(57) = 1.15$, $p = .25$, $d = 0.15$, suggesting minimal transfer of learning across training blocks from one test item to another from the same triplet. The apparent priming effect over the criterial test blocks of Experiment 1 was not evident across the two training blocks of this experiment. That result, however, is not necessarily problematic for the priming account; it may be that all elements were already primed during the immediately preceding study phase, just as we hypothesized was the case after the first criterial test block in Experiment 1.

Criterial test. An ANOVA identical to that from Experiment 1 was performed (see Figure 1.3), yielding highly significant main effects of criterial test condition, $F(2, 114) = 32.20$, $p < .001$, $MSE = 0.024$, $\eta_p^2 = 0.36$, and block, $F(2, 114) = 196.80$, $p < .001$, $MSE = 0.026$, $\eta_p^2 = 0.78$, as well as a significant interaction, $F(4, 228) = 5.92$, $p < .001$, $MSE = 0.019$, $\eta_p^2 = 0.094$. These results closely replicated the findings of Experiment 1.

In the follow-up ANOVA limited to tested-inverted and restudied items, we found no main effect of criterial test condition, $F(1, 57) = 0.076$, $p = 0.78$, $MSE = 0.034$, $\eta_p^2 = 0.0013$, as had been observed in Experiment 1. There was, however, a significant criterial test condition by block interaction, $F(2, 114) = 4.12$, $p = .019$, $MSE = 0.020$, $\eta_p^2 = 0.067$. That interaction mirrors the same (but non-significant) crossover pattern observed in Experiment 1. Post-hoc *t*-tests revealed a statistically significant performance advantage for restudied over tested-inverted items on the third block, $t(57) = 2.47$, $p = .017$, $d = 0.32$, but no significant difference between those conditions on the first or second blocks.

Overall, despite testing on two of the three possible test items during training, no evidence remained of positive transfer from tested-identical to tested-inverted items. These results diverge from the prior finding that encoding variability facilitates the transfer of test-enhanced learning (e.g., Goode et al., 2008), although the previous design involved anagram solving and can arguably be viewed as a skill-learning rather than memory recall paradigm. Moreover, the results of this experiment again contrast sharply with prior work showing excellent transfer for paired associates.

Experiment 3

In the third experiment we examined the effects of training on all three possible test items from each triplet involving two stimulus elements and one response. During the training phase either a triplet was restudied three times or each of its three test items was tested once. Therefore, this experiment included no transfer manipulation. Rather, it served to evaluate the efficacy of testing, relative to restudy, when all possible test items were tested once during training, and it promoted cross-experiment analyses that provided insight into whether testing yields increasing or decreasing enhancement of learning as a function of the number of items tested.

In the training phase, we also examined the effects of adding reminders for subjects to use interactive images, as well as increasing the number of unique interactive images that subjects were instructed to form for each triplet. This manipulation was motivated by prior evidence that instructing subjects to form additional memory associations during training can enhance transfer performance (e.g., Chan et al., 2006; Experiment 3).

Method

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

Materials, design, and procedure. This experiment's design was identical to that of Experiment 2, with the exception of two changes to the training phase. First, training featured three blocks of 36 trials instead of two. For eighteen of the triplets, restudy occurred on each of the three blocks. For the remaining 18 triplets, each of the three test items from the triplet was tested once across blocks. Thus, six tested-identical items per criterial test block had been presented, respectively, in the first, second, and third block of training (constituting the 18 tested-identical items).

Second, we also explored whether the type of interactive imagery instruction might moderate either overall performance or the testing effect. There were three levels of instruction type during training, manipulated between-subjects with random assignment: (1) standard interactive imagery, in the exact manner as in the preceding two experiments; (2) reinforced interactive imagery, in which subjects were reminded of the image-forming process at the start of the second phase of the training session; and (3) multiple interactive imagery, in which subjects were told to form new interactive images at the start of each training block.

Results and Discussion

Training. The mean accuracies (and *SE*s) for test items were 0.68 (0.060), 0.78 (0.054), and 0.72 (0.058) on the first, second and third blocks, respectively. A one-way within-subjects ANOVA revealed a significant difference in performance over training blocks, $F(2, 118) = 9.46$, $p < .001$, $MSE = 0.013$, $\eta_p^2 = 0.14$. Thus, in contrast to the training results for Experiment 2, accuracy in this experiment depended on the training block. The effect was non-monotonic over blocks, however, and the difference in proportions correct between the first and third blocks was small (0.04).

Criterial test. A mixed ANOVA on subject-level mean accuracy scores with a between-subject factor Interactive Imagery (standard vs. reinforced vs. multiple) and within-subject factors of Criterial Test Condition (tested-identical vs. restudied) and Block (1 vs. 2. vs. 3) was performed. In that analysis, no significant main effects or interactions involving the interactive imagery manipulation were observed (all *p*'s $\geq .57$). Neither reminders to use interactive imagery nor instructions to construct a new image for the test items presented on each block appreciably affected performance. Given prior work supporting the effectiveness of interactive imagery (Bower, 1970), those null results are surprising. One plausible account is that our triplet stimuli, which were designed to facilitate imagery, resulted in spontaneous use of imagery-based representation regardless of the instructional condition. Alternatively, the imagery manipulations may not have been heeded by subjects.

We therefore removed imagery instruction as a factor and performed a within-subjects ANOVA on data combined from subjects in all three imagery groups (Figure 1.4). Analogously to Experiments 1 and 2, we found highly significant main effects of criterial test condition, $F(1, 59) = 67.060$, $p < .001$, $MSE = 0.021$, $\eta_p^2 = 0.53$, and block, $F(2, 118) = 187.10$, $p < .001$, $MSE = 0.012$, $\eta_p^2 = 0.76$, as well as a significant criterial test condition by block interaction, $F(2, 118) = 52.23$, $p < .001$, $MSE = 0.0095$, $\eta_p^2 = 0.47$. These results confirm and extend the results of the prior experiments. A large testing effect was observed for triplets regardless of the number of unique test items per triplet that were presented during training.

Cross-Experiment Analyses

To assess the cross-experiment patterns for the testing effect we performed a mixed factorial ANOVA on the mean accuracy scores from the first criterial test block of each experiment, with the between-subjects factor Experiment (1, 2, and 3) and the within-subjects factor of Criterial Test Condition (combined items vs. restudied items). In this analysis, *combined* refers to both tested-identical and tested-inverted items. By combining these two types of test items into the same condition, and comparing it against the restudied condition, in this analysis we focused on a question of potential educational importance, namely: Does test-enhanced learning for a triplet in its entirety accelerate or decelerate, relative to restudy, when one, two, or all three test items per triplet are trained? In other words, what is the effect of taking tests on an increasing number of test items per triplet on performance for all three test items per triplet? For Experiments 1 and 2, the combined condition included data averaged over both tested-identical and tested-inverted items for each tested triplet. All three test items from each triplet were tested in Experiment 3, and hence all test items were classified into the combined condition.

The results are shown in Figure 1.5. We observed highly significant main effects of experiment, $F(2, 157) = 26.15$, $p < .001$, $MSE = 0.068$, $\eta_p^2 = 0.25$, and criterial test condition, $F(1, 157) = 143.37$, $p < .001$, $MSE = 0.018$, $\eta_p^2 = 0.48$, as well as a significant experiment by criterial test condition interaction, $F(2, 157) = 13.31$, $p < .001$, $MSE = 0.018$, $\eta_p^2 = 0.14$. The main effect of experiment corresponds to the increased overall performance levels for both combined and restudied triplets as the number of training exposures increased in Experiments 1 to 3. The main effect of criterial test condition corresponds to the overall advantage for combined versus restudied triplets. Inspection of Figure 1.5 shows that the experiment by criterial test condition interaction reflects an increasing difference between combined and restudied triplets over the progression from Experiments 1 to 3. Thus, the testing effect for overall triplet learning does not appear to be diminished (at least under the present conditions in which ceiling effects are unlikely to constitute a confound), and may be accelerated as the number of test items per triplet increases. These results should be interpreted with caution, however, given that the assignment of subjects to experiments was not random.

General Discussion

We explored test-enhanced learning relative to restudy for triple associates, as well as transfer of that learning to inverted test items. A robust testing effect was observed across all experiments, extending the effect to triple associates and likely to the larger class of multi-element stimuli. Moreover, the testing effect occurred despite randomization of word order and spatial position between training and the criterial test, which eliminates a superficial perceptual learning account. In stark contrast to the robust testing effect, we found no evidence for positive transfer from tested-identical to tested-inverted items relative to restudy, neither when one nor two of the three possible test items were tested during training.

The absence of transfer persisted across variations in pre-experimental associative strength over triplets. This was evident in supplementary analyses using the forward and backward associative strengths from available data (Nelson, McEvoy, & Schreiber, 1998) on all possible word pair combinations per triplet (i.e., *A-B, B-A, A-C, C-A, B-C,* and *C-B*), thus allowing us to calculate an average associative strength score for each triplet. A split-half analysis comparing triplets that scored relatively high versus low in associative strength (averages of 0.036 vs. 0.00079 in forward and backward associative strength across all possible word pair combinations per triplet) did not alter the transfer results observed in Experiments 1 and 2. We observed a large testing effect for both low and high associative strength items in both experiments, as well as a large main effect of associative strength in Experiment 1 (but not Experiment 2). Most importantly, in neither experiment did we see a trend toward an interaction between associative strength (high vs. low) and condition (restudy vs. transfer; tested condition was excluded), *p*'s > 0.50.

Although the finding of no transfer relative to restudy was not anticipated by the testing effect transfer literature, it is consistent with prior work on word triplets and arithmetic facts focused on RTs, as we discussed in the introduction. Given the wide range of stimulus types (both verbal and numerical), populations (both adults and children), skill levels (both single trial testing in the present experiments vs. extended retrieval practice for adult arithmetic), prior learning (none in the present experiments vs. many hours for adult arithmetic), and dependent variables (proportion correct vs. RT) over which high learning specificity for triplets following recall has been demonstrated (Rickard & Bajic, 2006; Walker, Bajic, Mickes, Kwak, & Rickard, 2014), it appears that a substantial degree of learning specificity due to testing is universal over various types of triple associate stimuli.

Theoretical Implications for Triple Associates

The present results suggest that the IE model, which was developed as an account of the specificity of RT learning for multi-element stimuli in the context of extended retrieval practice, may extend to the case of a single retrieval trial and the dependent variable of proportion correct. As we noted in the Introduction, that model assumes that correct retrieval creates a new recall representation that is independent of the study representation and can later be accessed on the criterial test only for the tested-identical item. The present work suggests that even a single correct retrieval trial can establish an independent recall representation that can support subsequent performance on the criterial test.

Because transfer in the present study was evaluated relative to restudy, we cannot conclude that entirely no transfer of learning to tested-inverted items took place in Experiments 1 and 2. Restudy presumably yielded some degree of additional learning (Roediger & Karpicke, 2006b). Thus, positive transfer of testing relative to a hypothetical no-training condition likely did occur in Experiments 1 and 2. The IE model does not anticipate that transfer. A slightly modified version of that model might nevertheless be consistent with the present results, given the possibility that feedback during training played less of a role in the previously described RT studies than in the present study. In the RT studies, accuracy was high throughout training and feedback was typically provided only on incorrect trials. Feedback thus may have played little role in learning, and the learning due to correct retrieval may have been fully asymmetrical in the tested direction. In the present context of accuracy learning, on the other hand, feedback on incorrect trials is likely to have played an important role in enhancing the study memory that is assumed in the IE model. That enhancement might in turn support transfer to tested-inverted items, under the assumption that feedback processing yields roughly symmetrical triplet learning (much as initial study presumably does) that can produce transfer.

According to that account, the nearly equivalent overall performance in the tested-inverted and restudy conditions in Experiments 1 and 2 reflects a balance of minimal or no transfer (even relative to a hypothetical no-training control) following correct training test trials and positive transfer following incorrect training test trials. Here, feedback is assumed to be critical for learning on incorrect trials but inconsequential for learning on correct trials (for supporting evidence see Fazio, Huelser, Johnson, & Marsh, 2010; Pashler, Cepeda, Wixted, & Rohrer, 2005). There is also independent evidence that incorrect initial test trials with feedback yield more learning than does restudy (Butler & Roediger, 2008; Hays, Kornell, & Bjork, 2013).

Transfer as a function of initial test performance. To gain insight into the viability of the hypothesis outlined above, we conducted post-hoc conditional analyses of the first criterial test block data of Experiment 1. Specifically, we explored transfer on that block separately for tested items that had been answered correctly versus incorrectly on the initial test. Thirty-five subjects who had both correct and incorrect initial test observations for both tested-identical and tested-inverted items on the criterial test were included. The results, shown in Table 1.1, are consistent with the IE model outlined above. When the response for a tested item was incorrect on the initial test, proportions correct were nearly identical for the tested-identical and tested-inverted items on the criterial test, consistent with the hypothesis that feedback after an incorrect test trial produces symmetrical learning that supports transfer. In contrast, when the response was correct on the initial test, we found a very large specificity of learning effect, consistent with the IE model predictions.

In summary, the results of the conditional analysis support the hypothesis that, for triplets at least, feedback on an incorrect test trial promotes the transfer of learning, whereas correct answer retrieval (with or without feedback) promotes specificity of learning. A caveat to that conclusion is the possibility of selection bias. Test items that were answered correctly on the initial test presumably correspond, on average, to triplets that are relatively easy to learn, whereas test items incorrectly answered correspond to triplets that are more difficult to learn. That fact raises the possibility that the transfer results are causally related to the intrinsic difficulty of the test items. It is conceivable that triplets that are difficult to learn intrinsically yield strong transfer, regardless of whether they are answered correctly on the initial test, whereas triplets that are easy to learn intrinsically yield weak transfer. Although that possibility cannot be ruled out on the basis of the present data, there is no apparent mechanistic reason to expect it.

Also relevant to the present analyses is performance on restudy test items on the first criterial test block. The mean for those test items overall (M = 0.23, SD = 0.023) was approximately the same for both tested-identical and tested-inverted items from triplets that were incorrectly solved on the initial test, as well as for tested-inverted items from triplets that had been correctly solved on the initial test (see Table 1.1). Those comparisons should be interpreted with caution, however, because restudied triplets cannot be grouped into those that would or would not have been correctly answered on a training phase test. If such grouping could be done, it would almost certainly yield a lower proportion correct (i.e., less than the 0.23 for overall restudied items) for restudied triplets that would have been answered incorrectly on an initial test, and a higher proportion correct (> 0.23) for restudied triplets that would have been answered correctly on an initial test. Using that logic, we can draw the following tentative conclusions for word triplets:

(1) An incorrect initial test followed by correct answer feedback produces roughly symmetric learning that is accessible for both tested-identical and tested-inverted (transfer) items, and (2) correct retrieval (with or without feedback) produces substantial learning for the tested items but minimal transfer of that learning to tested-inverted items.

Transfer for Triple Associates versus Paired Associates

The present findings for triple associate word stimuli contrast strikingly with prior findings of robust transfer for cue-response reversals in the case of paired associates, provided that training did not yield a high level of criterion learning (e.g., Carpenter et al., 2006; Vaughn & Rawson, 2014). Given the design and procedural similarities to the Carpenter et al. study, it appears that the contrasting results stem from the properties of the stimuli themselves (pairs vs. triplets) rather than from other factors.

Why does presenting both a prior stimulus element and a prior response element as retrieval cues – as was the case for tested-inverted items in this study – prevent access to the learning that occurred on the initial test, whereas that learning can be accessed in the case of pure stimulus-response reversal for paired associates? The IE model as it was originally developed and described above does not explain that difference. However, a revised version of the IE model that was proposed to explain positive RT transfer results between multiplication and factoring (Rickard, 2005) provides one candidate account. For those two tasks, the numerical elements of the cues (e.g., *4 x 7 = ?* or *28 = __ x__*) are fully reversed for multiplication versus factoring. The transfer item in that case can be considered a pure reversal of the critical numerical stimulus and response elements (the only elements that varied over items), much like paired associates. To account for the multiplication-factoring transfer, the revised IE model incorporated bi-directional associative links between the stimulus and response elements of a tested item. Thus, in the case of a pure reversal such as multiplication and factoring, positive transfer is expected. That version of the IE model maintained the critical property of the original IE model, however, in that the stimulus elements on the transfer test must be an exact match to either the previously tested stimulus (or, in the revised model, the response) for transfer to occur. Hence, training on, for example, __ *x*__ = *28* should transfer to a "pure numerical reversal", e.g., *4 x 7 =* __, as was observed. However, training on any of the following problems (e.g., *4 x 7 =* _; *28 ÷ 4 =* _; *28 ÷ 7 =* _) should transfer to neither of the other two problems in that set, as was observed in several studies noted in the Introduction. That version of the IE model can thus account for both the strong transfer for paired associates in the testing effect paradigm and the poor transfer for triplets observed here. Nevertheless, that model leaves open the important question of why an exact match of the transfer stimulus elements to either the previously encountered stimulus elements or the previously encountered response element appears to be necessary for positive transfer (relative to restudy) to occur. It also does not explain the recent finding of Vaughn and Rawson (2014) that the extent of transfer for paired associates appears to decrease with increasing levels of mastery.

The difference in the observed transfer patterns for triple associates versus paired associates can also be viewed from the perspective of at least two other theoretical accounts of the testing effect: the elaborative retrieval model (Carpenter, 2009) and the mediator effectiveness model (Pyc & Rawson, 2010). According to the elaborative retrieval model, memory retrieval during testing activates semantic associative paths between the two cue words (in the present experiments) to the response word, potentially resulting in multiple retrieval pathways and facilitating subsequent improved retrieval, relative to restudy. From the perspective of the mediator effectiveness model, a similar process occurs during training tests: Mediators that link cues to targets are more effective on training trials involving a test than on trials involving restudy, thus enhancing later recall of tested items (Vaughn & Rawson, 2014). Neither of those theories in their current forms make strong claims about transfer for triplets. They do, however, explain testing effects at an associative level that lends itself to the development of transfer accounts, and the mediator effectiveness hypothesis has been discussed as a candidate account of the finding for paired associates that asymmetry holds at higher mastery levels (Vaughn & Rawson, 2014).

Educational Implications

Following the recent progression from laboratory materials to applied materials in other testing effect domains, it will be important to determine whether the present results for triple associates generalize to educationally relevant multi-element factual materials in domains such as history (e.g., *who, what, where, when,* and *how* type facts) and biology, among many others. Although triplet stimuli of the type used in this study do not typically appear in classroom contexts, commonly used short answer and fill-in-the-blank test questions often have an analogous multi-element structure (e.g., in the question, "*Winston Churchill was the Prime Minister of what country during World War II?,"* the terms "*Winston Churchill*", "*Prime Minister*", "*United Kingdom*", and "*World War II*" form a quartet). Such questions are often included in textbooks as practice problems, and are a natural application of test-enhanced learning in future educational interventions. It remains to be determined whether transfer of test-enhanced learning from one question to another requiring a different response would occur in that case.

Acknowledgments

This chapter, in full, is reprinted by permission from Springer Customer Service Centre GmbH: Springer Nature, *Memory & Cognition,* Does test-enhanced learning transfer for triple associates? Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., and Rickard, T. C., 2016. Permission was granted for reuse in a dissertation/thesis. The dissertation author was the primary investigator and author of this paper.

Criterial Test Block 1

Triplet Number(s)	Condition (Restudied or Tested-)
1	-Identical
2	-Identical
3	-Identical
4	-Identical
5	-Identical
6	-Identical
7	*-Inverted*
8	*-Inverted*
9	*-Inverted*
10	*-Inverted*
11	*-Inverted*
12	*-Inverted*
13	*-Inverted*
14	*-Inverted*
15	*-Inverted*
16	*-Inverted*
17	*-Inverted*
18	*-Inverted*
19-36	Restudied

Criterial Test Block 2

Triplet Number(s)	Condition (Restudied or Tested-)
1	*-Inverted*
2	*-Inverted*
3	*-Inverted*
4	*-Inverted*
5	*-Inverted*
6	*-Inverted*
7	-Identical
8	-Identical
9	-Identical
10	-Identical
11	-Identical
12	-Identical
13	*-Inverted*
14	*-Inverted*
15	*-Inverted*
16	*-Inverted*
17	*-Inverted*
18	*-Inverted*
19-36	Restudied

Criterial Test Block 3

Triplet Number(s)	Condition (Restudied or Tested-)
1	*-Inverted*
2	*-Inverted*
3	*-Inverted*
4	*-Inverted*
5	*-Inverted*
6	*-Inverted*
7	*-Inverted*
8	*-Inverted*
9	*-Inverted*
10	*-Inverted*
11	*-Inverted*
12	*-Inverted*
13	-Identical
14	-Identical
15	-Identical
16	-Identical
17	-Identical
18	-Identical
19-36	Restudied

Figure 1.1. Example criterial test block design used in Experiment 1, with hypothetical triplet numbers for illustrative purposes. During the criterial test, one test item from each triplet was shown per block, and all 36 triplets appeared once per block. All three test items per triplet were tested over three blocks. Condition indicates whether a test item was previously presented in the first session: *tested-identical* items were trained using testing with feedback, *tested-inverted* items were not presented, and the complete triplet corresponding to each *restudied* item was restudied during training. The same criterial test block design was also used in Experiments 2 and 3, with the sole difference being a different number of tested and tested-inverted items per block.

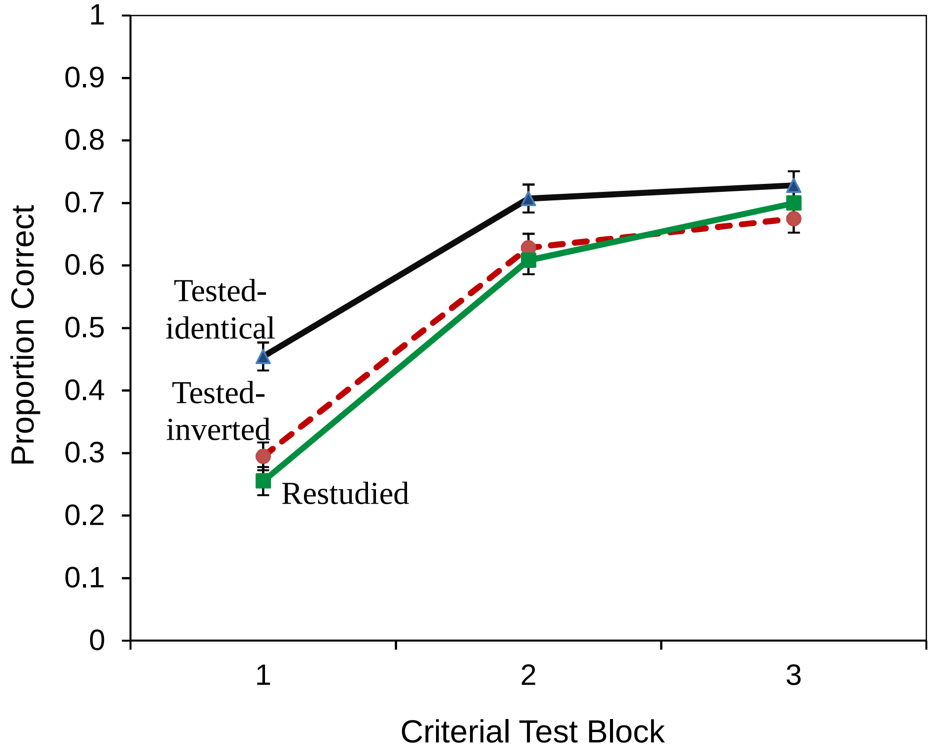


Figure 1.2. Mean accuracy scores (proportion correct) on the criterial test of Experiment 1 as a function of criterial test condition and block. Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on the method outlined by Loftus & Masson, 1994).

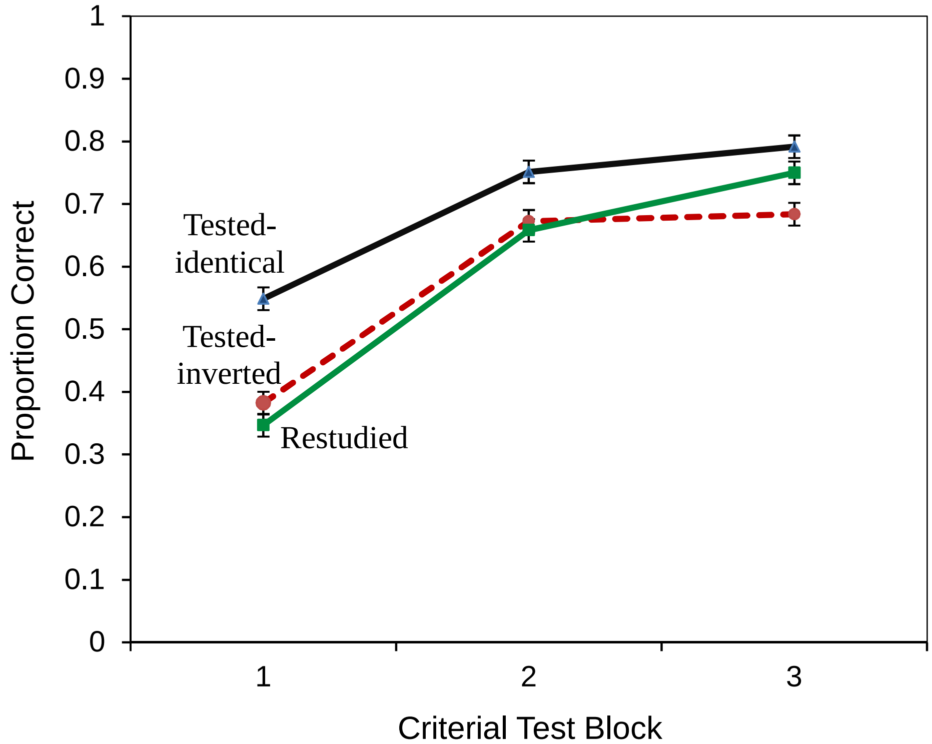


Figure 1.3. Mean accuracy scores (proportion correct) on the criterial test of Experiment 2 as a function of criterial test condition and block. Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores.

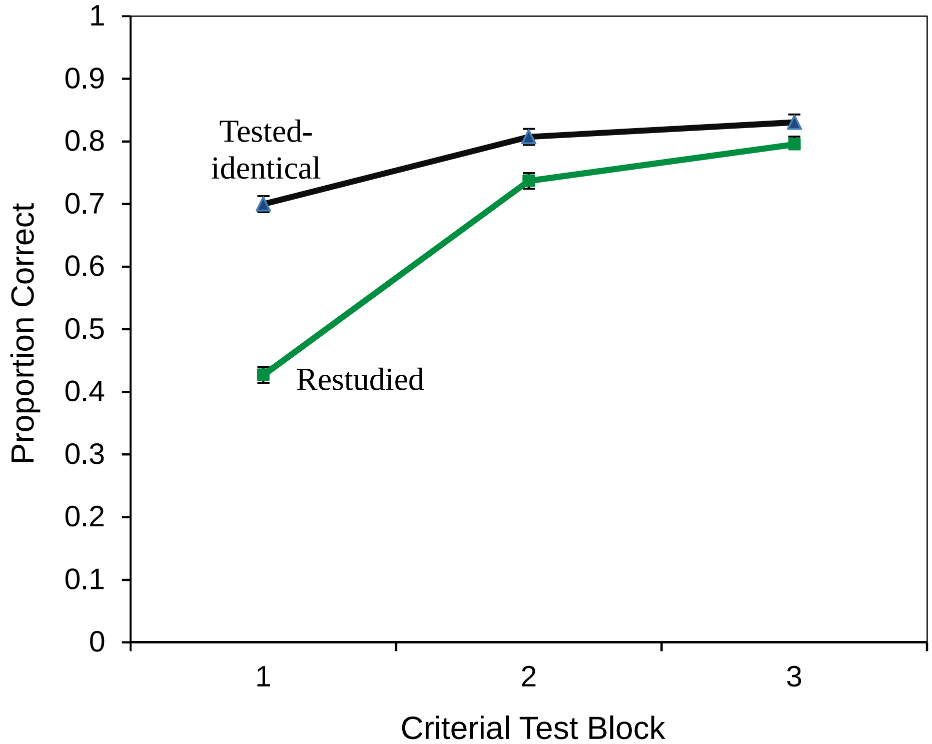


Figure 1.4. Mean accuracy scores (proportion correct) on the criterial test of Experiment 3 as a function of criterial test condition and block. Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores.

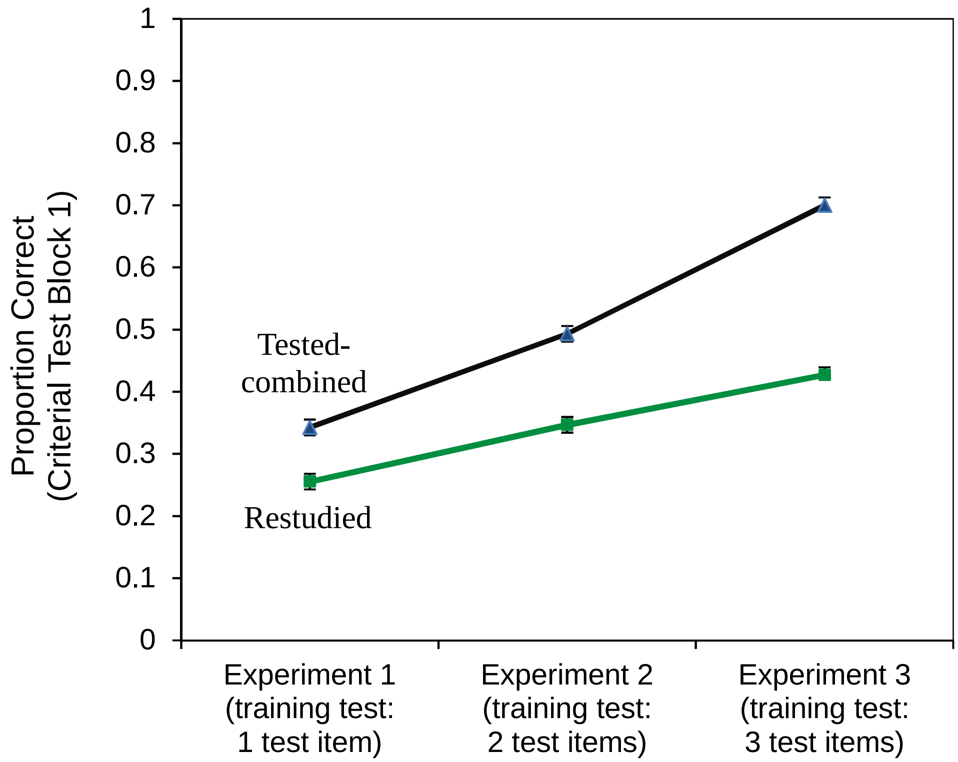


Figure 1.5. Results of the cross-experiment analysis showing mean accuracy scores (proportion correct) on the first criterial test block of Experiments 1, 2, and 3 for tested-identical and tested-inverted items (tested-combined) versus restudied items.

Table 1.1. Conditional analyses of mean proportion correct (SE) for Tested-Identical and Tested-Inverted Items on criterial test block 1 by training phase accuracy in Experiment 1.

Training phase test correct		Training phase test incorrect	
Tested-identical items	Tested-inverted items	Tested-identical items	Tested-inverted items
0.61	0.30	0.26	0.23
(0.050)	(0.034)	(0.051)	(0.050)

Appendix 1A

Triple Associate Word Lists Used in Experiments 1, 2, & 3

ants	spray	trash
bark	dog	stick
bat	dark	cave
beat	drum	march
bleach	stain	shirt
boy	run	field
bus	coins	line
cheer	game	score
clock	rush	late
cloth	soap	sink
cow	grass	milk
doctor	pills	note
farm	sun	sweat
gift	rose	wine
girl	smile	flower
ground	snail	mist
honk	cab	traffic
knight	castle	sword
lion	hunt	meat
map	hike	water
mug	hand	tea
paper	ink	desk
phone	ear	sound
plane	nap	blanket
room	key	door
ship	horn	sea
skate	fall	knee
sky	bird	rain
sofa	laugh	friends
street	bike	car
teeth	bite	wolf
towel	swim	pool
tree	child	swing
voice	sing	guitar
wall	paint	frame
warm	bread	coffee

References

Agarwal, P. K. (2011). Examining the relationship between fact learning and higher order learning via retrieval practice. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3468823)

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The psychology of learning and motivation*: Vol. 9 (pp. 89-132). New York, NY: Academic Press.

Bajic, D., Kwak, J., & Rickard, T. C. (2011). Specificity of learning through memory retrieval practice: The case of addition and subtraction. *Psychonomic Bulletin & Review, 18*(6), 1148-1155. doi:http://dx.doi.org/10.3758/s13423-011-0151-4

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637. doi:http://dx.doi.org/10.1037/0033-2909.128.4.612

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Josslyn, & R. Shiffrin (Eds.), *Essays in honor of William K. Estes: from learning processes to cognitive processes.* (Vol. 2, pp. 35-67). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning & Verbal Behavior, 9*(5), 529-533.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133. doi:http://dx.doi.org/10.1037/a0019902

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. doi:http://dx.doi.org/10.3758/MC.36.3.604

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:http://dx.doi.org/10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. doi:http://dx.doi.org/10.1177/0963721412452728

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:http://dx.doi.org/10.3758/BF03193405

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19*(3), 443-448. doi:http://dx.doi.org/10.3758/s13423-012-0221-2

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153-170. doi:http://dx.doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49-57. doi:http://dx.doi.org/10.1080/09658210903405737

Chan, J. C., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553-571. doi:http://dx.doi.org/2006-20327-005

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory, 18*(3), 335-350. doi:http://dx.doi.org/10.1080/09658211003652491

Goode, M. K., Geraci, L., & Roediger, H. L. (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic Bulletin & Review, 15*(3), 662-666. doi:http://dx.doi.org/10.3758/PBR.15.3.662

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:http://dx.doi.org/10.1037/a0028468

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304. doi:http://dx.doi.org/10.1080/09658211.2011.560121

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*(2), 151-164. doi:http://dx.doi.org/10.1016/j.jml.2013.03.002

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621-629. doi:http://dx.doi.org/10.1037/a0015183

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*(6), 823-840. doi:http://dx.doi.org/10.3758/BF03195769

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition, 30*(6), 841-849. doi:http://dx.doi.org/10.3758/BF03195770

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review, 18*(5), 998-1005. doi:http://dx.doi.org/10.3758/s13423-011-0113-x

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science, 331*(6018), 772-775.

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013). The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18*(3), 409-425. doi:http://dx.doi.org/10.1007/s10459-012-9379-7

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490. doi:http://dx.doi.org/10.3758/BF03210951

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:http://dx.doi.org/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:http://dx.doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360-372. doi:http://dx.doi.org/10.1002/acp.2914

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Available from: http://www.usf.edu/FreeAssociation

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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:http://dx.doi.org/2004-22496-001

Postman, L., & Stark, K. (1967). Studies of learning to learn: IV. Transfer from serial to paired-associate learning. *Journal of Verbal Learning & Verbal Behavior, 6*(3), 339-353.

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

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523-548. doi:http://dx.doi.org/10.1007/s10648-013-9240-4

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126*(3), 288-311. doi:http://dx.doi.org/10.1037/0096-3445.126.3.288

Rickard, T. C. (2005). A revised identical elements model of arithmetic fact representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 250-257. doi:http://dx.doi.org/2005-02160-006

Rickard, T. C., & Bajic, D. (2006). Cued recall from image and sentence memory: A shift from episodic to identical elements representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 734-748. doi:http://dx.doi.org/2006-08497-007

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1281-1295. doi:http://dx.doi.org/10.1037/0278-7393.22.5.1281

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1139-1153. doi:http://dx.doi.org/10.1037/0278-7393.20.5.1139

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:http://dx.doi.org/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:http://dx.doi.org/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:http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233-239. doi:http://dx.doi.org/10.1037/a0017678

Tran, R., Rohrer, D., & Pashler, H. (2014). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*. doi:http://dx.doi.org/10.3758/s13423-014-0646-x

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*(9), 1127-1131. doi:http://dx.doi.org/10.1177/0956797611417724

Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14-26. doi:http://dx.doi.org/10.1016/j.jml.2014.04.004

Walker, D., Bajic, D., Mickes, L., Kwak, J., & Rickard, T. C. (2014). Specificity of children's arithmetic learning. *Journal of Experimental Child Psychology, 122*, 62-74. doi:http://dx.doi.org/10.1016/j.jecp.2013.11.018

CHAPTER 2

Testing with Feedback Yields Potent, but Piecewise, Learning of History and Biology Facts

Pan, S. C., Gopal, A., and Rickard, T. C. (2016).

Journal of Educational Psychology 108(4), 563-575.

Abstract

Does correctly answering a test question about a multi-term fact enhance memory for the entire fact? We explored that issue in four experiments. Subjects first studied Advanced Placement History or Biology facts. Half of those facts were then restudied, while the remainder were tested using “5 W” (i.e., *who, what, when, where,* or *why*) or analogous questions. Each question assessed a specific critical term of the fact. In the first three experiments, one test question was posed per tested fact; in the fourth experiment, up to three different test questions were posed per tested fact. After at least a 24 hr delay, a criterial test involved questions that assessed the same terms that were tested during training, as well as questions that assessed a different term from that previously tested. Results showed that testing produced piecewise fact learning: tested terms benefited relative to restudy, but untested terms did not. That pattern held when either fill-in-the-blank or multiple-choice questions were used during training, when one or two test trials were used during training, for both history and biology facts, and when more than one term from each fact was tested during training. Thus, across a range of circumstances, taking tests on complex facts results in a selective memory benefit for tested terms. In analogous

applied settings, testing on multiple response terms should promote more comprehensive retention.

Introduction

"*Winston Churchill was Prime Minister of the United Kingdom during World War II*." In domains ranging from astronomy to zoology, fact learning is essential for building foundational knowledge. Moreover, fact learning is often the chief goal of learners. The Advanced Placement (AP) exams, which over 2.2 million students in the U.S. and Canada take annually (College Board, 2013a), are a case in point. Those exams typically assess knowledge of 60 to 80 facts each—a subset of the many facts that students typically study for a year or more. Given the prevalence of fact learning, the question follows: what is the best way to learn and retain facts? One answer that is strongly backed by learning science: by taking tests.

The use of tests to improve memory, also known as *test-enhanced learning* or *retrieval practice*, is strongly endorsed by many cognitive and educational psychologists (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler et al., 2007). This recommendation is backed by extensive work over decades and across numerous content domains (e.g., Bjork, 1975; Carrier & Pashler, 1992; Gates, 1917; Glover, 1989; for reviews see Roediger & Karpicke, 2006; Rowland, 2014). Carpenter, Pashler, and Cepeda (2009), for example, showed that taking tests on facts during an 8th grade U.S. history course boosts long-term retention by as much as 41% by the end of the course, relative to restudy. However, most reports of test-enhanced learning (also called the *testing effect*) are subject to a major caveat: the same questions are used during initial training and criterial tests (for discussion see McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Rohrer, Taylor, & Sholar, 2010). The important question of whether test-enhanced learning of facts *transfers* beyond tested content has received comparatively less attention.

In a recent brief review, Carpenter (2012) concluded that testing effects successfully transfers across changes in test format and to application questions. Kang, McDermott, and Roediger (2007), for example, found positive transfer of fact learning when the test format was changed from short answer during training to multiple-choice on the criterial test. Further, Butler (2010) reported positive transfer for application questions which require inferences from initially tested facts (e.g., from animal wings to aircraft wings). Successful transfer to application questions has been reported elsewhere (e.g., Chan, 2009; Chan, 2010; Chan, McDermott, and Roediger, 2006; Foos and Fisher, 1988; Karpicke and Blunt, 2011; McDaniel, Howard, and Einstein, 2009), although no transfer in similar contexts has also been observed by Agarwal (2011) and Tran, Rohrer, and Pashler (2014). Moreover, there remain other categories of transfer of test-enhanced learning that have seen little research to date.

Testing and Transfer Between Fact Terms

The topic of transfer of test-enhanced learning between terms of a fact, which is the focus of this manuscript, is qualitatively distinct from the types of transfer described above. It is an important issue because test questions usually do not assess entire facts, but rather parts of facts. This is especially true for *multi-term* facts, or facts with multiple critical terms (e.g., the fact which begins this article has four: *Winston Churchill*, *Prime Minister*, *United Kingdom*, and *World War II*). In the classroom and in everyday life, one may need to retrieve any of several different terms from a fact. Yet it remains unclear whether a test question covering one, two, or three terms of a fact will be sufficient to memorize the entire fact for later recall. Prior results relevant to that topic are mixed; two studies show positive transfer of learning from one question to another from the same fact, whereas two others suggest no transfer.

McDaniel, Anderson, Derbish, and Morissette (2007) administered online fill-in-the-blank or multiple-choice quizzes on neuroscience facts, and then provided subjects with delayed, detailed feedback on their performance. They observed positive transfer to multiple-choice unit tests in which previously untested parts of each fact were assessed (e.g., from the quiz question "*All preganglionic axons, whether sympathetic or parasympathetic, release _________ as a neurotransmitter*" to the unit test question "*All _________ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter*"). McDaniel et al. (2013; Experiment 1) reported positive transfer for *term-definition* reversals in which multiple-choice classroom quiz questions requiring key terms from science facts as correct responses were switched to final multiple-choice test questions requiring definitions as correct responses, or vice versa (e.g., from the quiz question "*What is the definition of temperature?"* and answer "*The measure of the average kinetic energy of the particles in a substance*" to the criterial test question "*What is the measure of the average kinetic energy of the individual particles in an object?"* and answer "*Temperature*").

In contrast to the above studies, neither Hinze and Wiley (2011; Experiments 1 and 2) nor Pan, Wong, Potter, Mejia, and Rickard (2016) observed transfer. Hinze and Wiley administered fill-in-the-blank tests on portions of factual science texts, using two blanks per sentence or paragraph. Uniquely among the studies covered here, no feedback was provided; the facts themselves were also lengthier (e.g., "*In mitosis, _________ are created from a parent cell. Each new cell contains a complete set of chromosomes which guarantees that they are _________. When the chromatids align this forms the metaphase plate which later becomes the location where the cell is split in two*"). The criterial test entailed presenting the same portions of text, but with the two previous blanks filled and two new ones added; no positive transfer was found. Recently, Pan et al. (2016) demonstrated that test-enhanced learning does not transfer for triple associates (word triplets such as *lion, hunt, meat*), even when correct feedback was provided after each test trial. Performance gains for previously tested questions (e.g., *lion, meat, ?*) did not transfer, relative to restudy, to new questions on previously tested triplets (e.g., *meat, hunt, ?*). Although triplets lack sentence structure, those results reinforce the possibility that the benefits of testing can be highly specific to tested parts of a fact-like concept.

The Current Study

The present experiments, which directly extend the Pan et al. (2016) triplet design using authentic educational materials, assessed the consequences of answering one test question on a critical term of a fact (Experiments 1-3), or up to three questions, each on different critical terms of a fact (Experiment 4), for mastery of that entire fact. In contrast to Hinze and Wiley (2011), correct answer feedback was provided during training. If the lack of feedback in the Hinze and Wiley study was solely responsible for their finding of no transfer, then positive transfer should be observed here. Our feedback is confined, however, to presenting the correct answer after each test trial. If the extensive delayed feedback in the McDaniel et al. (2007) study was critical to their positive transfer effects, then, based on the findings of Hinze and Wiley and Pan et al., no transfer is expected here. If the term-definition structure of the materials in McDaniel et al. (2013) was critical to their positive transfer results, then again no transfer would be expected here.

A secondary goal of the current study was to explore the role of fact domain expertise on testing and transfer effects. In each experiment, subjects with prior AP United States History, World History, or Biology experience in high school (constituting 25-50% of subjects per experiment), as well as subjects with no prior AP course experience, participated. It is possible that the greater expertise for AP students (if confirmed by overall performance) affords a richer knowledge base that will promote more integrative processing through testing, and hence better transfer of learning.

Experiment 1

Method

Subjects. The target sample size in this experiment was 40, which is comparable to that of prior testing effect studies (e.g., Carpenter et al., 2009; Hinze & Wiley, 2011; McDaniel et al., 2007). Undergraduate students were recruited from the subject pool at the University of California, San Diego and received course credit for their participation. Students from both lower- and upper-division courses were eligible to participate. Subject ages were $M = 21.03$, $SD = 2.53$, and ranged from 18 to 31 years. The majority (79%) of the sample was female. Data from two subjects was not analyzed: one due to a computer error generating unusable data and the other due to the subject not returning for the second session. The resulting sample size for analysis was 38.

Materials. Thirty-six history facts were obtained using the AP United States History and World History preparatory texts produced by *Barron's* and *Princeton Review* (Armstrong, Daniel, Kanarek, & Freer, 2014; McCannon, 2014; Meltzer & Bennett, 2014; Resnick, 2014). The history facts, which average 11 words in length, all contain three or more one-word critical terms which address any of the "5 W's" (i.e., *who*, *what*, *when*, *where*, or *why*) that are essential for the comprehension of that given fact. Moreover, most critical terms address different categories (e.g., *who* vs. *what*), and thus exhibit minimal semantic overlap. An example history fact (with terms in italics) is: "*Overlord*, an operation led by *Eisenhower*, began with the invasion of *Normandy*." For three critical terms in each fact, one fill-in-the-blank test question (e.g., "*Overlord, an operation led by Eisenhower, began with the invasion of* _________") and one corresponding short answer test question (e.g., "*Overlord, an operation led by Eisenhower, began with the invasion of WHERE?*") was created; both questions had the same one-word correct answer. While similar in wording and structure to its fill-in-the-blank counterpart, the wording of the short answer test question was modified when necessary to maintain grammatical accuracy. Overall, the full set of materials included three fill-in-the-blank and three short answer questions for each fact (further examples are included in the Appendix 2A).

Design and procedure. As illustrated in Figure 2.1, Session 1 contained two phases: the study phase and the training phase. During the study phase, subjects viewed all 36 history facts, one at a time, for 8 seconds each, and in random order determined anew for subjects. All facts were studied once.

In the subsequent training phase, facts were randomly assigned for each subject to one of two lists: the restudy list or the testing list. Facts in the restudy list were presented for 8 s each, using a procedure identical to that of the study phase. Facts in the testing list were presented as fill-in-the-blank questions with feedback (6 s to type the missing term followed by the correct answer being shown for 2 s, during which subjects were no longer allowed to type responses). For tested facts, only one of the three critical terms per fact was tested, and the tested term per fact was counterbalanced across subjects. All 36 facts were presented in random order. After all facts had been trained once, the training phase ended, and subjects were reminded to return in 48 hrs for Session 2.

Session 2, the criterial test, assessed recall for the entire set of 108 answer terms across all 36 history facts. As illustrated in Figure 2.2, the criterial test involved three 36-trial blocks. Within each block, (1) each fact was assessed once in random order; (2) the assessment was a short answer test question; (3) subjects had unlimited time to provide a one-word answer, and (4) no feedback was provided. In each block, the test question for each fact had a different missing term, such that the three critical terms of each fact were separately assessed over the three blocks. Further, within each block, six of the previously tested facts had questions with the same missing term that was to be retrieved as during training (*tested* questions) and 12 had a different missing term (*transfer* questions); the other 18 questions assessed facts with no prior retrieval practice (*restudied* questions). Thus, the six facts in the tested condition in block 1 (i.e., facts having the same missing term to be retrieved as was the case during training) were assessed in the transfer condition of block 2 (with the missing term being different from that which was tested during training), and analogously, six of the facts in the transfer condition in block 1 were assessed in the tested condition of block 2. Block 3 completed that cycling of questions over the full set of 18 questions that were tested during training. Question assignment to blocks on the criterial test was counterbalanced and block order was randomized for each subject.

The testing of all three critical terms on the criterial test allowed for a comprehensive assessment of the consequences of testing for each fact. The choice of short answer test questions for the criterial test (instead of, for example, multiple-choice) was intended to approximate fact retrieval conditions under everyday circumstances, where cued recall from memory is most common. Moreover, using different types of questions on the training and criterial tests (rather than fill-in-the-blank in both sessions) served to minimize the possibility that any transfer effects, if obtained, could be explained by superficial learning (i.e., learning related to adjacent words or sentence structure that could be used as retrieval cues, as opposed to learning of the fact itself that is generalizable to later contexts in which retrieval cues are not identical to those of prior testing).

After completion of the experiment, subjects were asked whether they had taken an AP History course (on any history topic) in high school.

Data coding and analysis. The sole dependent variable in both the initial and criterial test analyses was accuracy. Because many of the missing terms are easy to misspell, two research assistants who were blind to question type checked subject responses for misspelled answers that could be unambiguously matched to correctly spelled answers; these were coded as correct responses. Here and in Experiments 2 and 3, separate analyses performed with misspelled words coded as incorrect responses yielded lower overall accuracy but no difference in relative performance in the tested, transfer, and restudied conditions.

Results and Discussion

Training phase performance. Overall accuracy on the initial test (involving 18 fill-in-the-blank test questions for each subject) was M = 0.27, SE = 0.03.

Criterial test performance. We performed a within-subjects factorial Analysis of Variance (ANOVA) on subject-level mean accuracy scores (Figure 2.3) with factors of Criterial Test Condition (tested vs. transfer vs. restudied) and Block (1 vs. 2 vs. 3). In this and all subsequent analyses, alpha was set at 0.05. There were statistically significant effects of Criterial Test Condition, $F(2, 74) = 25.90$, $MSE = 0.027$, $p < .0001$, $\eta_p^2 = 0.41$, Block, $F(2, 74) = 41.88$, $MSE = 0.019$, $p < .0001$, $\eta_p^2 = 0.53$, but no significant Criterial Test Condition by Block interaction, $F(4, 148) = 1.21$, $MSE = 0.024$, $p = .31$. The significant improvement over blocks may reflect intermediate priming effects (e.g., Pan et al., 2016); because each test question per block displayed the correct answers to test questions in subsequent blocks, answer priming effects may have occurred for the second and third blocks of the criterial test. Inspection of Figure 2.3 illustrates the main effect of training condition: in all three blocks, proportion correct was higher in the tested condition than in either the restudy or transfer condition. A planned follow-up ANOVA limited to the transfer and restudied conditions yielded a non-significant effect of Criterial Test Condition, $F(1, 37) = 2.75$, $MSE = 0.015$, $p = .11$, and no interaction between Test Condition and Block, $F(2, 74) = 2.50$, $MSE = 0.010$, $p = .089$. Overall, the critical finding of Experiment 1 is that testing strongly enhances fact learning, but affords no, or only very limited, benefits to transfer questions relative to restudy.

Effect of prior AP experience. In exit surveys, nearly half (n = 18) of the subjects reported having previously taken an AP History course. Test performance during the training session was indistinguishable for AP and non-AP students, $t(30) = 1.49$, $p = .15$. Overall performance on the criterial test, however, was substantially better for AP students, $M = 0.40$ vs. $M = 0.28$, $t(29) = 2.32$, $p = .0028$, $d = 0.77$, suggesting better retention for AP students. The overall pattern of no transfer relative to restudy on the criterial test held equivalently, however, for AP and non-AP students. The percent transfer (where 0% means that transfer and restudy performance were equivalent, and 100% means that transfer and tested performance were equivalent) was -3% and -2% for AP and non-AP students, respectively; for both groups, performance on transfer questions was no better, and numerically worse, than that in the restudied condition.

Experiment 2

In the second experiment we investigated whether the transfer properties of test-enhanced learning as established for history facts in Experiment 1 would extend to a physical science domain, namely biology. As evident from a comparison of AP Biology and AP History exams, the structure of biology facts differs substantially from history facts (e.g., reduced frequency of *who* and *when* type terms and more terms referring to objects and processes). It is thus possible that test-enhanced learning for science facts will yield better transfer than it does for history facts.

Method

Subjects. Minimum sample size to detect a modest transfer effect was determined using a priori power analysis. Based on the standard deviation of the criterial test transfer minus restudied condition proportion correct difference scores in Experiment 1, a sample size of at least 47 is needed to achieve power of 0.95 to detect a mean proportion correct difference score of 0.05 or greater (based on a one-tailed, one-sample *t*-test, $\alpha = 0.05$). When subject sign-up and data collection were completed, 58 undergraduates had participated for course credit. Subject ages were $M = 21.12$, $SD = 3.072$, and ranged from 17 to 38 years. Two-thirds (67%) of the sample was female. All subjects finished both sessions of the study.

Materials. Thirty-six biology facts were obtained using the AP Biology preparatory texts produced by *Barron's* and *Princeton Review* (Goldberg, 2014; Magloire, 2014). The biology facts had the same defining characteristics (e.g., at least three one-word critical terms each, average 11 words in length) as the history facts from Experiment 1. However, unlike in the preceding experiment, the information covered by each fact did not include geopolitical concepts, persons, or unique historical events. An example biology fact (with critical terms in italics) is: "The *Krebs* cycle occurs in the *mitochondria* and produces *ATP*." As before, a fill-in-the-blank test question and a short answer test question were created for each of three critical terms per fact (further examples are included in the Appendix 2A).

Design and procedure. Two sessions occurred in the same manner as in Experiment 1, with the following exception: during Session 1, after training on all 36 biology facts, subjects trained on all 36 facts for a second time using identical procedures (i.e., the same fill-in-the-blank question was shown for each tested fact, and the complete fact was displayed for each restudied fact). Hence, for tested facts, subjects were asked to retrieve the same missing term twice. This modification was motivated by pilot tests, which suggested that the biology facts may be more difficult than the history facts.

Results and Discussion

Training phase performance. Mean accuracy on the initial test was $M = 0.37$, $SE = 0.03$ for the first training block, and $M = 0.64$, $SE = 0.03$ for the second training block. The improvement between blocks was statistically significant, $t(57) = 13.63$, $p < .0001$, $d = 1.79$.

Criterial test performance. A within-subjects factorial ANOVA on subject-level mean accuracy scores (Figure 2.4) with factors of Criterial Test Condition (tested vs. transfer vs. restudied) and Block (1 vs. 2 vs. 3) found statistically significant effects of Criterial Test Condition, $F(2, 114) = 57.46$, $MSE = 0.032$, $p < .0001$, $\eta_p^2 = 0.50$, Block, $F(2, 114) = 38.27$, $MSE = 0.017$, $p < .0001$, $\eta_p^2 = 0.40$, and a significant criterial Test Condition by Block interaction, $F(4, 228) = 5.60$, $MSE = 0.018$, $p = .00026$, $\eta_p^2 = 0.090$. The interaction corresponds to an increased pattern of improvement for transfer and restudied questions across blocks; however, even by the final block, performance on tested questions is still greater than for either transfer or restudied questions by $M = 0.10$, $SE = 0.03$. As in Experiment 1, a follow-up ANOVA limited to the transfer and restudied conditions found no significant effect of Criterial Test Condition, $F(1, 57) = 0.26$, $MSE = 0.022$, $p = .61$, a significant effect of Block, $F(2, 114) = 61.62$, $MSE = 0.013$, $p < .0001$,

$\eta_p^2 = 0.52$, and no significant Block by Condition interaction, $F(2, 114) = 0.40$, $MSE = 0.014$, $p = .67$. These results replicate the findings of Experiment 1: testing strongly benefits fact learning relative to restudy, but in a manner that is specific to the exact tested term.

Effect of prior AP experience. There was a trend toward better training phase test performance for AP ($n = 15$) than for non-AP subjects, $M = 0.28$ vs. $M = 0.16$, $t(34) = 1.85$, $p = .07$, $d = 0.50$. Overall performance on the criterial test was again significantly higher for AP subjects, $M = 0.71$ vs. $M = 0.45$, $t(38) = 6.27$, $p < .0001$, $d = 1.48$, suggesting better retention for AP students. The overall pattern of no transfer relative to restudy held equivalently, however, for AP and non-AP students, with the percent transfer being -6% and 1%, respectively.

Experiment 3

The third experiment investigated whether specificity of test-enhanced learning would manifest when multiple-choice test questions, rather than fill-in-the-blank test questions, are used during training. Multiple-choice questions are conveniently graded and are frequently used in educational contexts. The structure of multiple-choice questions requires careful consideration of, and adjudication between, several possible responses.

Method

Subjects. Fifty-eight undergraduates participated for course credit. Subject ages were $M = 20.37$, $SD = 2.38$, and ranged from 18 to 28 years. The majority (83%) of the sample was female. Six subjects did not return to complete the second session; data from the 52 remaining subjects was analyzed.

Materials. The 36 history facts from Experiment 1 served as stimuli for Experiment 3. The only change was the format of test questions during training: instead of fill-in-the-blank test questions, multiple-choice test questions were used (examples are included in the Appendix 2A). Each multiple-choice test question covered one of the three possible missing terms per fact. There were four one-word answer choices per multiple-choice test question (A, B, C, or D; one correct answer and three distractors). Each distractor was designed to be competitive enough to warrant consideration by subjects (Little, Bjork, Bjork, & Angello, 2012; Little & Bjork, 2015).

Design and procedure. Two sessions occurred in the same manner as in the prior experiments, except for the use of multiple-choice test questions during Session 1. For each multiple-choice test question, subjects were given 6 s to select one of the four possible answer choices; the correct answer was subsequently shown for 2 s. There was one training trial per fact.

Results and Discussion

Training phase performance. Mean accuracy on the initial test (18 multiple-choice test questions) was M = 0.73, SE = 0.03.

Criterial test performance. A within-subjects factorial ANOVA on subject-level mean accuracy scores (Figure 2.5) with factors of Criterial Test Condition (tested vs. transfer vs. restudied) and Block (1 vs. 2 vs. 3) yielded statistically significant effects of Criterial Test Condition, $F(2, 102) = 34.19$, $MSE = 0.94$, $p < .0001$, $\eta_p^2 = 0.40$, Block, $F(2, 102) = 67.22$, $MSE = 1.73$, $p < .0001$, $\eta_p^2 = 0.57$, and no significant Criterial Test Condition by Block interaction, $F(4, 204) = 1.70$, $MSE = 0.037$, $p = .15$. A follow-up

ANOVA limited to the transfer and restudied conditions found no significant effect of Criterial Test Condition, $F(1, 51) = 2.48$, $MSE = 0.067$, $p = .12$, a significant effect of Block, $F(2, 102) = 69.59$, $MSE = 1.43$, $p < .0001$, $\eta_p^2 = 0.58$, and no significant Block by Condition interaction, $F(2, 102) = 1.86$, $MSE = 0.024$, $p = .16$. These results replicate and extend the findings of the prior experiments: multiple-choice testing strongly benefits fact learning, but only for the tested term.

Little et al. (2012; see also Little & Bjork, 2015), who also used multiple-choice tests, had distractors during training tests reappear as correct answers to final transfer test questions. They observed partial transfer of learning for multiple-choice training tests relative to cued recall training tests (for which no distractors were shown). Little et al. suggested that the act of considering plausible distractors that were answers to later test questions was the driver of improved transfer performance. This contrasts with the present experiment, where distractors used on multiple-choice training test questions did not reappear as correct answers to transfer questions on the criterial test.

Effect of prior AP experience. There was no difference in mean test performance during the training session for AP ($n = 26$) than for non-AP subjects, $t(49) = 1.30$, $p = .20$. Overall performance on the criterial test was also indistinguishable in this case, $M = 0.41$ vs. $M = 0.40$, $t(49) = 0.18$, $p = .86$. Percent transfer was 1% and -5% for AP and non-AP students, respectively.

Experiment 4

The fourth experiment investigated the effects of training on more than one critical term per fact. For example, given the fact (with critical terms italicized), "*Jefferson* purchased *Louisiana* from the *French*", it is possible to test each critical term separately with a different test question (e.g., by asking about *who, what,* and *whom* regarding that fact). This design addressed two questions. First, does training on two critical terms per fact promote transfer to a third term on the criterial test? If testing with feedback on the second term during training reactivates memory for the preceding test on the first term, then the second test may result in a more integrated, or holistic representation that will promote transfer to a third term on the criterial test. Second, are there increasing or diminishing returns from expanding the number of answer terms that are tested during training? We approached these questions by manipulating the number of different answer terms that subjects were asked to retrieve per fact during training (one, two, or three), and correspondingly the number of times that subjects were asked to restudy facts during training.

Method

Subjects. Fifty-two undergraduates participated for course credit. Subject ages were M = 19.91, SD = 1.68, and ranged from 17 to 26 years. The majority (73%) of the sample was female. All but seven subjects completed both sessions of the experiment; data from the remaining 45 subjects was analyzed.

Materials. Fifty-four AP United States History and AP World History facts were used for this experiment. This included the 36 history facts from Experiments 1 and 3, along with 18 additional facts that were extracted from the same sources. More facts were necessary due to the different training conditions in Session 1. The extra facts met the same selection criteria as in prior experiments, and three fill-in-the-blank and three short answer test questions were also created for each fact.

Design and procedure. Two sessions occurred in the same manner as Experiment 1, with the following modifications. In the training phase of Session 1, of the 27 facts assigned to testing, one-third each were assigned to have one, two, or three critical terms tested. Of the 27 facts assigned to be restudied, a corresponding number were restudied once, twice, or three times. Assignment of facts to restudy or testing, as well as to one, two, or three test or restudy trials during training, was counterbalanced across subjects. As illustrated in Figure 2.6, the training phase featured three blocks of 36 trials each; within each block, there were six trials for the nine facts that were tested on one term or restudied once across the three blocks, 12 trials for the nine facts that were tested on two terms or restudied twice across the three blocks, and 18 trials for the nine facts that were tested on three terms or restudied three times across the three blocks. Fact presentation was randomly ordered within each block, no fact was repeated within each block, and there were no breaks between blocks. By the end of the three training blocks, each third of the 54 facts had been tested or restudied once, twice, or three times. Session 1 ended with a reminder for subjects to return the following day to complete the second part of the experiment.

Session 2, the criterial test, occurred 24 hrs after training and assessed subjects' recall of the entire set of 162 answer terms across all 54 history facts. This involved three 54-trial blocks that were based on the design of the criterial test in Experiments 1-3. As in the prior experiments, each fact appeared once per block, and a different critical term was assessed per fact and per block without repeats. In each block, there were nine facts each per training condition (one, two, or three tested terms or restudy opportunities per block). Of the nine facts that had been tested once during training, there were three tested and six transfer questions per criterial test block. Of the nine facts that had been tested twice during training, there were six tested and three transfer questions per criterial test block.

Results and Discussion

Training phase performance. Mean proportion correct across all three training blocks of the initial test was M = 0.31 (SE = 0.040) for facts that had one term tested, M = 0.39 (SE = 0.031) for facts that had two terms tested, and M = 0.40 (SE = 0.032) for facts that had three terms tested. A within-subjects one-way ANOVA yielded a significant effect of terms trained (i.e., facts that were trained on one, two or three terms across the three training blocks), $F(2, 88) = 6.76$, $MSE = 0.098$, $p = .0019$, $\eta_p^2 = 0.13$. In the context of correct answer feedback, which allows for restudy of the entire fact on each test trial regardless of accuracy, the improved performance over the number of terms trained was expected.

Criterial test performance. We first performed a within-subjects factorial ANOVA on subject-level mean accuracy scores (Figure 2.7) with factors of Criterial Test Condition (tested vs. transfer vs. restudied), Block (1 vs. 2 vs. 3), and Number of Trained Terms or Restudy Trials (one vs. two). As there were no transfer questions for facts that had three terms tested during training, this analysis was confined to data from the facts that had two tested terms or two restudy trials. The ANOVA yielded statistically significant effects of Criterial Test Condition, $F(2, 88) = 35.32$, $MSE = 2.27$, $p < .0001$, $\eta_p^2 = 0.44$, and Block, $F(2, 88) = 28.78$, $MSE = 1.53$, $p < .0001$, $\eta_p^2 = 0.40$, no significant effect of the Number of

Trained Terms or Restudy Trials, $F(1, 44) = 1.94$, $MSE = 0.11$, $p = .17$, and no significant interaction effects (p's $> .17$). A follow-up ANOVA limited to the transfer and restudied conditions (and again limited to facts which were tested on two terms or had two restudy trials) yielded no significant effect of Criterial Test Condition, $F(1, 44) = 3.17$, $MSE = 0.14$, $p = .082$, a significant effect of Block, $F(2, 88) = 28.39$, $MSE = 1.35$, $p < .0001$, $\eta_p^2 = 0.39$, no significant effect of Trained Terms or Restudy Trials, $F(1, 44) = 3.19$, $MSE = 0.21$, $p = .081$, and no significant interactions (p's $> .35$). These results replicate the findings of no transfer relative to restudy observed in the prior experiments. They also extend those findings in an important respect: even when two terms are testing during training, there is no transfer of test-enhanced learning to a previously untested term.

To assess the effect of expanding the number of answer terms tested during training on overall acquired knowledge for each fact, we performed a within-subjects factorial ANOVA on mean accuracy scores (Figure 2.8) from the first criterial test block, with factors of criterial test condition (Combined tested and transfer vs. restudied) and Number of Tested Terms or Restudy Trials (1 vs. 2 vs. 3). Blocks 2 and 3 were not included in this analysis because the overall pattern of results in block 1 was replicated in those blocks in the prior analyses and because block 1 provides the purest and most sensitive measure of training effects. By combining the tested and transfer conditions in this analysis, we were able to assess the effects of testing on one, two, or three terms per fact on memory for the entire fact. That approach to data analysis directly addresses the educationally important question of whether testing with feedback yields diminishing, constant, or increasing learning effects for the fact as a whole, as a function of the number of terms tested.

There were statistically significant effects of Criterial Test Condition, $F(1, 44) = 36.49$, $MSE = 1.38$, $p < .0001$, $\eta_p^2 = 0.45$, and Number of Trained Terms, $F(2, 88) = 20.84$, $MSE = 0.58$, $p < .0001$, $\eta_p^2 = 0.32$. The apparent interaction between those factors in Figure 2.8, however, did not reach statistical significance, $F(2, 88) = 2.33$, $MSE = 0.12$, $p = .10$. Thus, with respect to fact memory as a whole, test-enhanced learning appears to increase linearly as a function of the number of trained terms, whereas repeated restudy may have diminishing returns. The linear increase in proportion correct for tested items is consistent with the hypothesis that learning in the training phase occurs independently for each question-term combination of a fact (for related discussion see Pan et al., 2016).

Effect of prior AP experience. There was significantly better training phase test performance for AP ($n = 14$) than for non-AP subjects, $M = 0.46$ vs. $M = 0.34$, $t(37) = 2.07$, $p = .046$, $d = 0.61$. Overall performance on the criterial test was also significantly higher for AP subjects, $M = 0.60$ vs. $M = 0.44$, $t(37) = 32.26$, $p = .0090$, $d = 0.86$. Extending the results from the preceding experiments, the percent transfer, averaged over all criterial test blocks and across the one, two, and three items tested or restudy trials training conditions, was -9% and -5% for AP and non-AP students, respectively.

General Discussion

The foregoing experiments assessed whether testing on one or more critical terms of a fact benefits long-term memory for the entire fact. The results demonstrate that, under educationally important circumstances, testing can produce potent, but piecewise, fact learning. Despite consistently large testing effects (38%, 38%, 29%, and 41% higher proportion correct for testing vs. restudy in Experiments 1, 2, 3, and 4, respectively), those effects were entirely specific to tested terms, and there was no positive transfer to untested terms relative to restudy. Indeed, in each experiment there was a non-significant trend toward negative transfer. That lack of transfer held for both history and biology facts, when either fill-in-the-blank or multiple-choice questions were used during training, when a single term was tested twice during training (Experiment 2) and when more than one term was tested during training (Experiment 4). Accordingly, we expect that these findings will generalize to other subject materials and types of training tests.

Conditions of Transfer and Theoretical Implications

As noted in the Introduction, the results from the prior literature on transfer between terms of facts are mixed, with two of four studies demonstrating positive transfer relative to restudy. We suggest that those contrasting results can be integrated by two principles. First, if there is extensive feedback about the entire fact between training and the criterial test, positive transfer may be observed. In McDaniel et al. (2007), subjects had unlimited opportunities to view online feedback without any time limit over one week, in preparation for a later test. That extended feedback is likely to have been processed most thoroughly for incorrectly answered test questions, and that processing, which constitutes a form of restudy, may have yielded positive transfer relative to restudy on the criterial test. By that account, the event of testing in itself was not the primary basis for transfer in that report; rather, it is additional learning due to further study opportunities. Second, positive transfer may occur when facts have an explicit term-definition structure, as in McDaniel et al. (2013). If subjects have a learning goal of associating a term and a definition (perhaps because they expect a test of that type), then testing’s benefits may be analogous to that for paired associates, which also exhibit positive transfer (Carpenter, Pashler, & Vul, 2006). In other words, if a given fact can be logically organized as two components, it may be the case that such a structure is more easily learned and hence more transferrable than more complex materials. We have neither evidence nor intuition about that possibility, but rather note it here as a plausible integrative account. Finally, if neither of the conditions above hold, as in the current experiments, the Pan et al. (2016) experiments, and the Hinze and Wiley (2011) experiments, then no transfer will be observed relative to restudy. Further research to test those hypotheses is needed.

From the broader theoretical perspective, the current results support specificity-of-learning accounts of memory representations following retrieval practice for multi-term stimuli, such as the identical elements model (Rickard et al., 1994; Rickard & Bajic, 2006; Rickard & Bourne, 1996). That model, originally developed to account for response time gains following training on arithmetic facts, holds that successful retrieval after initial study establishes a new and separate recall representation for that particular stimulus-response configuration. That memory representation is only accessible when the stimulus-response configuration that was previously trained is presented; for instance, a particular set of operands in a multiplication fact (e.g., Rickard et al., 1994) or specific words from a word triplet (e.g., Pan et al., 2016). The current experiments add a more complex category of stimuli, namely multi-element facts, to the list of materials that are subject to the specificity principles outlined by the identical elements model. It remains to be determined, however, whether the same results would apply if the to-be-learned materials are different from the facts used in the current study (such as, for example, conceptual knowledge).

The current results also suggest that transfer of test-enhanced learning is not substantially modulated by expertise level, at least for multi-term facts. In three of the four experiments, AP students performed significantly and substantially better overall on the criterial test than did non-AP students, confirming the expected higher expertise. However, there were no significant differences in transfer results for those two subject groups. Indeed, performance in the transfer condition for both groups was slightly poorer than performance in the restudy condition in all but one instance (AP students: -3%, -6%, -1%, and -9%; non-AP students: -2%, 1%, -5%, and -5% for experiments 1, 2, 3, and 4, respectively). Based on those results, higher expertise in fact domains apparently does not result in more integrated processing during the initial test that could support transfer.

A caveat to the above conclusion is that specific expertise for the assessed AP facts was not directly manipulated in the current experiments. In future studies, the role of expertise on transfer could be further examined by providing additional time in one condition to review relevant or related subject matter prior to the training phase (thus facilitating greater expertise and contextual knowledge), or by presenting related facts sequentially in one condition, as for example in Chan (2009; Experiment 1).

The trend toward poorer performance in the transfer condition than in the restudy condition, while not significant, raises the possibility of negative transfer. Anderson, Bjork, and Bjork (1994) and others have also observed negative transfer following testing using the retrieval-induced forgetting paradigm. That paradigm differs markedly from that of the current study, however. In the case of retrieval-induced forgetting, testing on specific category exemplars typically results in poorer recall of other category exemplars, relative to exemplars from categories that were not tested or restudied at all, and often only after a relatively short (< 24 h) delay interval (Chan, 2009). In the current study, the terms of each tested fact were from different categories, transfer was compared against restudy rather than no re-exposure, and longer delays were used. It is thus difficult to draw inference about possible relations between those paradigms.

Practical Implications for Testing on Facts

The present transfer results are educationally important given that a large portion, and perhaps the majority, of facts that learners encounter in many educational contexts are multi-term facts rather than paired-associate or term-definition type materials. AP facts exemplify this pattern. In the latest AP World History, U.S. History, and Biology practice tests (College Board, 2011; College Board, 2012; College Board, 2013b), assessed facts have, by our analysis, an average of four testable terms, and few have an explicit paired-associate or term-definition structure (although clearly such facts are also encountered in other educational contexts). Multi-term facts can also be found in a wide variety of high school, college, technical, and other courses. Thus, knowledge of the transfer properties of multi-term facts is essential to efforts to optimize learning in applied contexts.

Of immediate impact to learners, our results indicate that instructors should not expect that one test question will typically be sufficient to memorize an entire fact. Unless training tests target multiple (or in the best-case scenario, all) terms of each fact, a selective memory benefit can result. This finding is particularly important given that practice tests—for instance those at the end of textbook chapters—often assess factual knowledge with a single fill-in-the-blank, short answer, or multiple-choice question per fact. Fortunately, redesigning practice tests to solve this problem is a relatively simple task: ask more than one question per fact, and ensure that questions comprehensively target all parts of each fact. The results from Experiment 4 indicate that learning rate does not diminish as the number of tested terms is increased.

Although the present experiments were not focused on a comparison of test formats, the results from Experiments 1 and 3, which were identical except for the format of test questions during training (fill-in-the-blank vs. multiple-choice), indicate that both types of tests yield potent testing benefits, but that fill-in-the-blank questions with feedback may produce a larger benefit for tested terms (9% greater mean difference in proportion correct vs. restudy). That apparent disparity in criterial test performance is consistent with prior findings of larger testing effects following training with short answer vs. multiple-choice tests (e.g., Butler & Roediger, 2007; Duchastel, 1981; McDaniel et al., 2007).

Maximizing Test-Enhanced Fact Learning

Future work may examine whether there are training conditions that reliably produce transfer of test-enhanced fact learning from one term to another—increasing the degree to which tests benefit fact learning. Allowing more extensive feedback, as in McDaniel et al. (2007), is one possible strategy. Instructional strategies that promote more integrative processing (e.g., Chan et al., 2006; Chan, 2009) are another option. From the perspective of optimizing learning efficiency, however, the extra time required for such approaches, relative to that required for testing on a variety of answer terms, must be factored into any systematic evaluation. Moreover, until such techniques are consistently demonstrated, a strategy of testing on as many parts of each fact as possible stands as the most promising solution for successful fact learning.

Acknowledgments

This chapter, in full, is reprinted with permission from the American Psychological Association, with the required notice embedded in the remainder of this paragraph. Copyright © 2016 American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is: Pan, S. C., Gopal, A., and Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology 108*(4), 563-575. This chapter may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association. The dissertation author was the primary investigator and author of this paper.

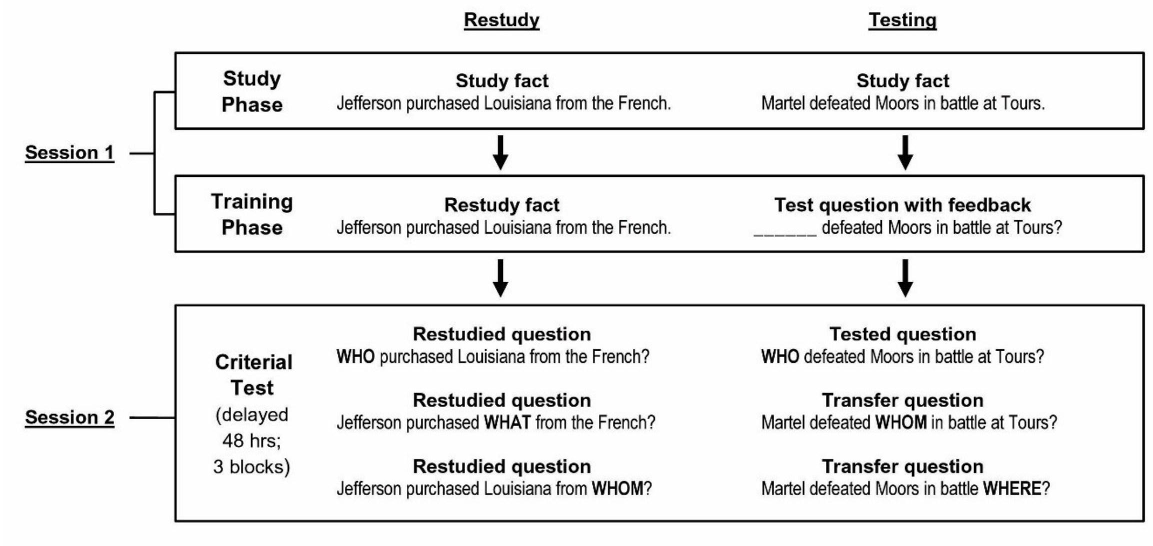


Figure 2.1. Experiment 1-3 procedure and example stimuli (Experiment 1) involving training through testing (right) and training through restudy (left). Training condition was manipulated within-subjects. (1) Study phase: subjects view all 36 facts, one at a time. (2) Training phase: facts trained via testing with feedback or restudy. In Experiments 1 and 2, fill-in-the-blank test questions are used; in Experiment 3, multiple-choice test questions are used. (3) Criterial test: recall for all 108 critical terms across 36 facts is assessed. There are three 36-trial blocks, within each of which one question per fact is shown. For previously restudied facts, all three criterial test questions are in the *restudied* condition; for previously tested facts, two questions are in the *transfer* condition, and one is in the *tested* condition.

Criterial Test Block 1

Fact number(s)	Final Test Condition
1	Tested
2	Tested
3	Tested
4	Tested
5	Tested
6	Tested
7	Transfer
8	Transfer
9	Transfer
10	Transfer
11	Transfer
12	Transfer
13	Transfer
14	Transfer
15	Transfer
16	Transfer
17	Transfer
18	Transfer
19-36	Restudied

Criterial Test Block 2

Fact number(s)	Final Test Condition
1	Transfer
2	Transfer
3	Transfer
4	Transfer
5	Transfer
6	Transfer
7	Tested
8	Tested
9	Tested
10	Tested
11	Tested
12	Tested
13	Transfer
14	Transfer
15	Transfer
16	Transfer
17	Transfer
18	Transfer
19-36	Restudied

Criterial Test Block 3

Fact number(s)	Final Test Condition
1	Transfer
2	Transfer
3	Transfer
4	Transfer
5	Transfer
6	Transfer
7	Transfer
8	Transfer
9	Transfer
10	Transfer
11	Transfer
12	Transfer
13	Tested
14	Tested
15	Tested
16	Tested
17	Tested
18	Tested
19-36	Restudied

Figure 2.2. Example criterial test block design used in Experiments 1-3, with hypothetical fact numbers for illustrative purposes. During the criterial test in Session 2, one critical term from each fact was tested per block, and all 36 facts appeared once per block. All three terms per fact were tested over three blocks. Criterial test condition indicates whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied), during Session 1.

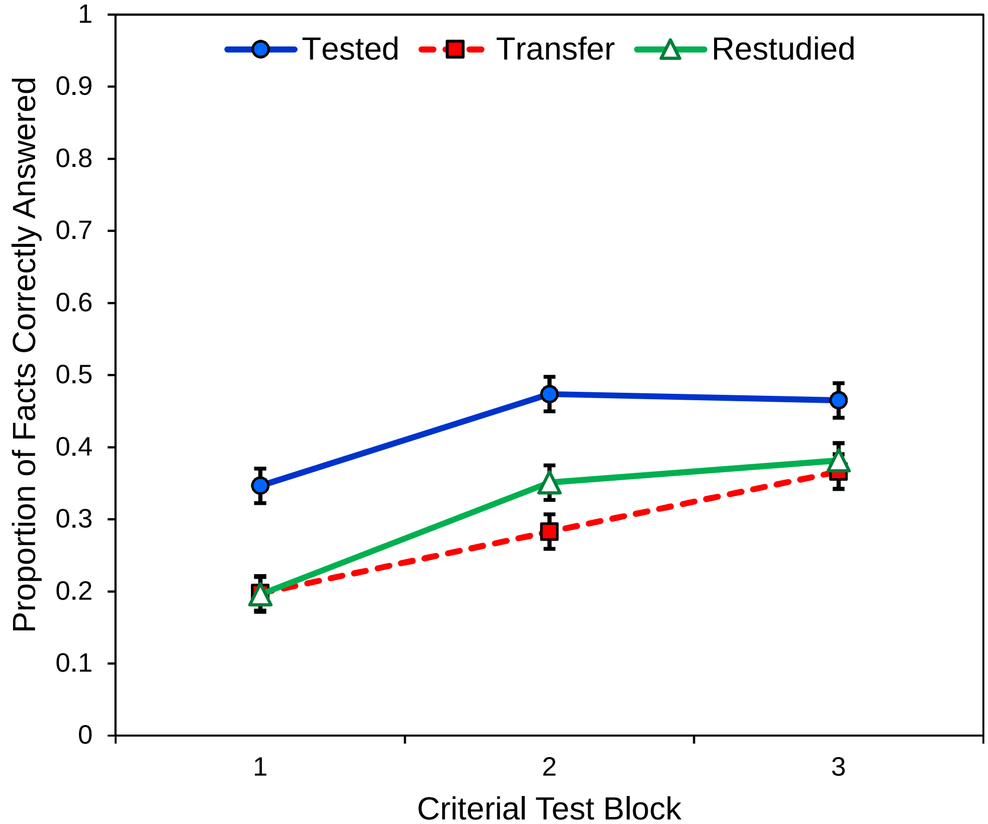


Figure 2.3. Results from Experiment 1: mean accuracy performance for history fact questions on the criterial test as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994).

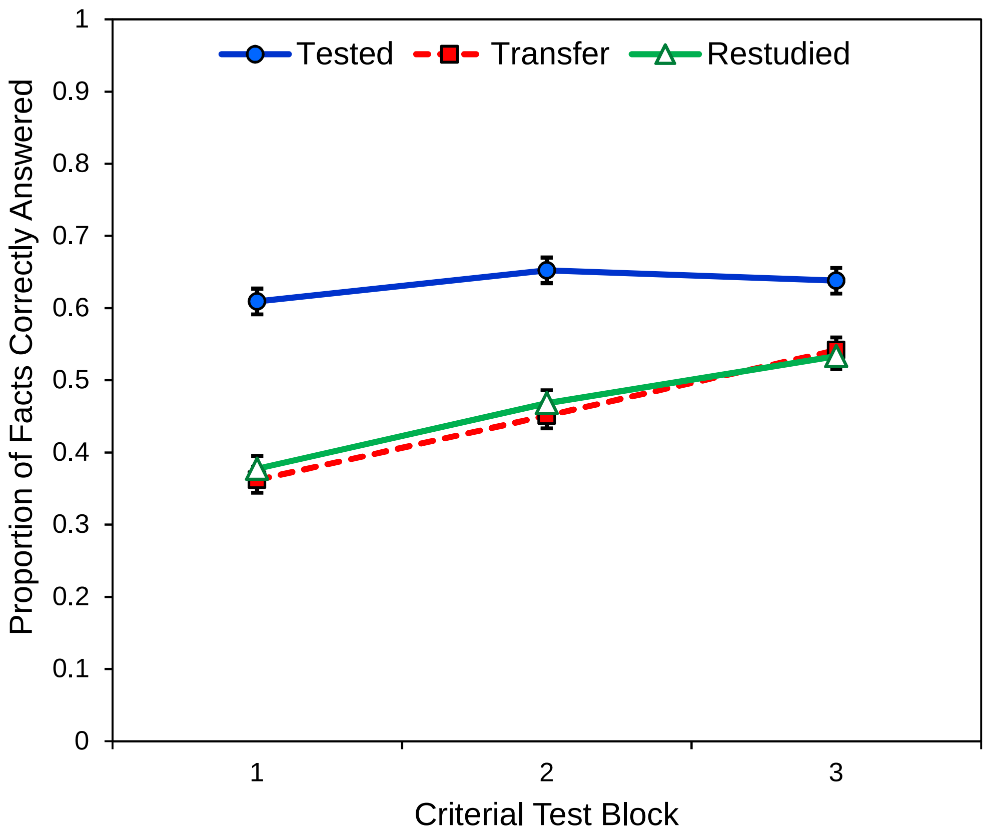


Figure 2.4. Results from Experiment 2, in which subjects trained twice per fact during training: mean accuracy performance for biology fact questions on the criterial test as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994).

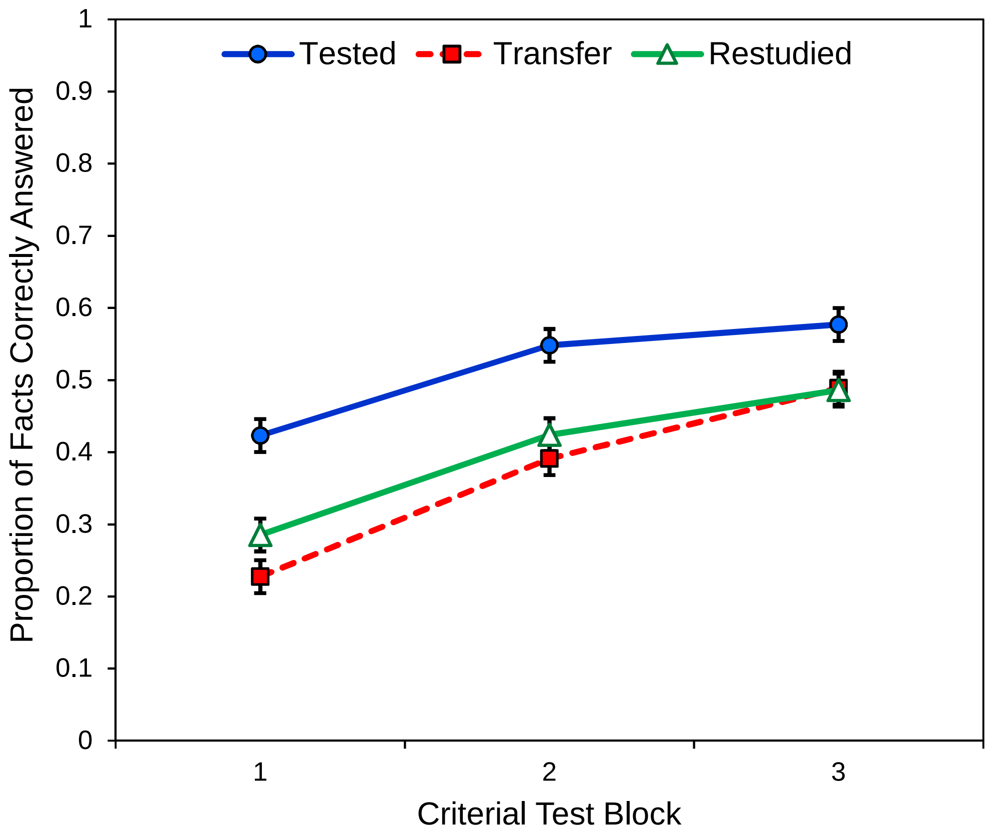


Figure 2.5. Results from Experiment 3, in which multiple-choice test questions were used during training: mean accuracy performance for history fact questions on the criterial test as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994).

Training Block 1

Fact number(s)	Training Condition
1	1 term tested
2	1 term tested
3	1 term tested
10	2 terms tested
11	2 terms tested
12	2 terms tested
13	2 terms tested
14	2 terms tested
15	2 terms tested
19-27	3 terms tested
28	1 restudy trial
29	1 restudy trial
30	1 restudy trial
37	2 restudy trials
38	2 restudy trials
39	2 restudy trials
40	2 restudy trials
41	2 restudy trials
42	2 restudy trials
46-54	3 restudy trials

Training Block 2

Fact number(s)	Training Condition
4	1 term tested
5	1 term tested
6	1 term tested
16	2 terms tested
17	2 terms tested
18	2 terms tested
13	2 terms tested
14	2 terms tested
15	2 terms tested
19-27	3 terms tested
31	1 restudy trial
32	1 restudy trial
33	1 restudy trial
43	2 restudy trials
44	2 restudy trials
45	2 restudy trials
40	2 restudy trials
41	2 restudy trials
42	2 restudy trials
46-54	3 restudy trials

Training Block 3

Fact number(s)	Training Condition
7	1 term tested
8	1 term tested
9	1 term tested
10	2 terms tested
11	2 terms tested
12	2 terms tested
16	2 terms tested
17	2 terms tested
18	2 terms tested
19-27	3 terms tested
34	1 restudy trial
35	1 restudy trial
36	1 restudy trial
37	2 restudy trials
38	2 restudy trials
39	2 restudy trials
43	2 restudy trials
44	2 restudy trials
45	2 restudy trials
46-54	3 restudy trials

Figure 2.6. Example training block design used in Experiment 4, Session 1, with hypothetical fact numbers for illustrative purposes. For tested facts, a different missing term was assessed on each test trial across all three blocks. For restudied facts, the entire fact was presented on each restudy trial. Of the 54 facts in this experiment, 9 each were assigned to the 1, 2, and 3 tested term or restudy trial training conditions. In each 36-trial block, equal distribution of facts in the different training conditions was accomplished by presenting 3 facts in the 1 term tested condition, 6 facts in the 2 terms tested condition, and 9 facts in the 12 terms tested condition. A corresponding number were presented in each of the restudied conditions.

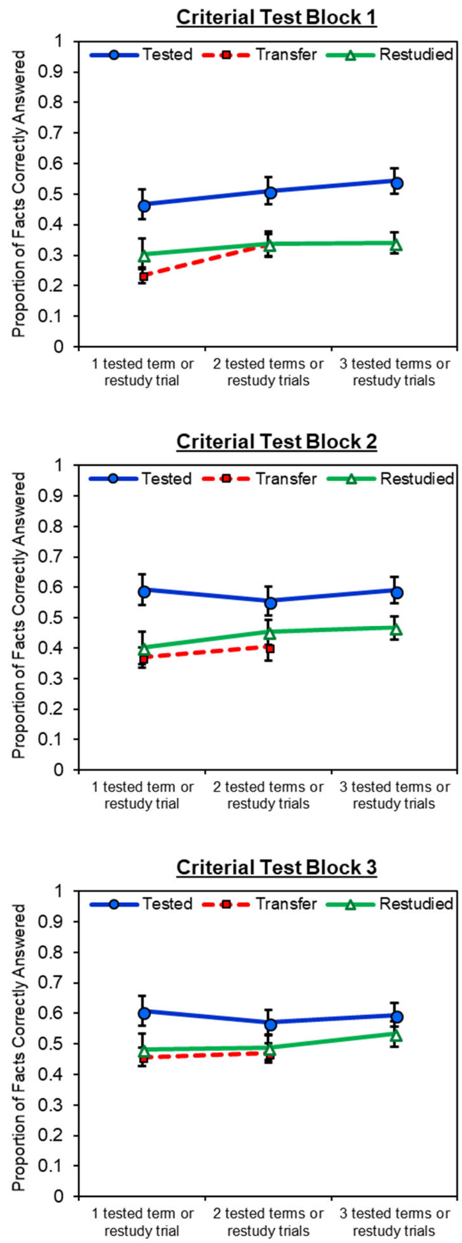


Figure 2.7. Results from Experiment 4: mean accuracy performance for history fact questions by training condition (1, 2, or 3 tested terms or restudy trials) per criterial test block as a function of whether the missing term was previously retrieved (tested), not previously retrieved, but from a tested fact (transfer), or from a fact that was not previously tested (restudied). Error bars are standard error of the mean.

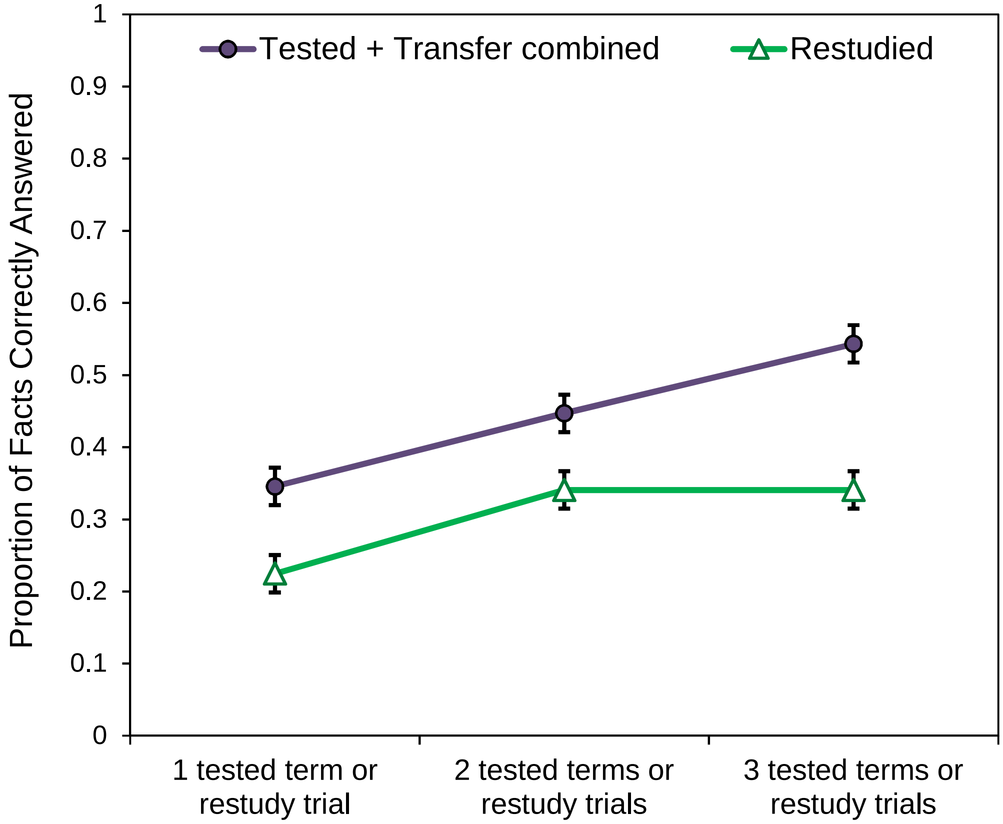


Figure 2.8. Experiment 4 mean accuracy performance for history fact questions by training condition (1, 2, or 3 tested terms or restudy trials) on the first block of the criterial test, as a function of whether the fact was previously tested at least once (tested + transfer questions combined) or not previously tested at all (restudied). Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject mean accuracy scores (based on Loftus & Masson, 1994).

Appendix 2A

Training and Criterial Test Question Examples

Exp.	Subject	Type	Example
1,4	AP History	Training fact	An edict signed by Henry in Nantes gave rights to Calvinists.
		Fill-in-the-blank	An edict signed by _____ in Nantes gave rights to Calvinists.
			An edict signed by Henry in _____ gave rights to Calvinists.
			An edict signed by Henry in Nantes gave rights to _____.
		Short answer	An edict signed by WHO in Nantes gave rights to Calvinists?
			WHERE did Henry sign an edict giving rights to Calvinists?
			An edict signed by Henry in Nantes gave rights to WHOM?
2	AP Biology	Training fact	During glycolysis, sugar is broken down into pyruvate.
		Fill-in-the-blank	During glycolysis, sugar is broken down into _____.
			During _____, sugar is broken down into pyruvate.
			During glycolysis, _____ is broken down into pyruvate.
		Short answer	Sugar is broken down into pyruvate during WHAT?
			During glycolysis, WHAT is broken down into pyruvate?
			During glycolysis, sugar is broken down into WHAT?
3	AP History	Fact	Darrow defended Scopes over his teaching of evolution.
		Multiple-choice	_____ defended Scopes over his teaching of evolution. a. Bryan c. Hunter b. Darrow d. Butler
			Darrow defended _____ over his teaching of evolution. a. Scopes c. Hicks b. Stewart d. Raulston
			Darrow defended Scopes over his teaching of _____. a. religion c. evolution b. creationism d. segregation
		Short answer	WHO defended Scopes over his teaching of evolution?
			Darrow defended WHOM over his teaching of evolution?
			Darrow defended Scopes over his teaching of WHAT?

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063-1087. doi:http://dx.doi.org/10.1037/0278-7393.20.5.1063

Agarwal, P. K. (2011). Examining the relationship between fact learning and higher order learning via retrieval practice. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Order No. 3468823)

Armstrong, M., Daniel, D., Kanarek, A., & Freer, A. (2014). *The Princeton Review: Cracking the AP World History Exam.* New York, NY: Random House.

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133. doi:http://dx.doi.org/10.1037/a0019902

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4-5), 514-527. doi:http://dx.doi.org/10.1080/09541440701326097

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. doi:http://dx.doi.org/10.1177/0963721412452728

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:http://dx.doi.org/10.1002/acp.1507

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

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

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153-170. doi:10.1016/j.jml.2009.04.004

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49-57. doi:10.1080/09658210903405737

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553-571. doi:http://dx.doi.org/2006-20327-005

College Board. (2011). *AP World History: Practice Exam and Notes.* New York, NY.

College Board. (2012). *AP Biology: Practice Exam and Notes.* New York, NY.

College Board. (2013a). Advanced Placement Program Student Score Distributions. Available from: http://media.collegeboard.com/digitalServices/pdf/research/2013/STUDENT-SCORE-DISTRIBUTIONS-2013.pdf

College Board. (2013b). *AP United States History: Practice Exam.* New York, NY.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology, 6*(3), 217-226.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. doi: http://dx.doi.org/10.1177/1529100612453266

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80*(2), 179-183. doi:http://dx.doi.org/10.1037/0022-0663.80.2.179

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399. doi:http://dx.doi.org/10.1037/0022-0663.81.3.392

Goldberg, D. (2014). *Barron's AP Biology (4th ed.)*. Hauppauge, NY: Barron's Educational Series.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304. doi:http://dx.doi.org/10.1080/09658211.2011.560121

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science, 331*(6018), 772-775.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558. doi:http://dx.doi.org/10.1080/09541440601056620

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43*(1), 14-26. doi:http://dx.doi.org/10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337-1344. doi:http://dx.doi.org/10.1177/0956797612443370

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490. doi:http://dx.doi.org/10.3758/BF03210951

Magloire, K. (2014). *The Princeton Review: Cracking the AP Biology Exam.* New York, NY: Random House.

McCannon, J. (2014). *Barron's AP World History (6th ed.)*. Hauppauge, NY: Barron's Educational Series.

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:http://dx.doi.org/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:http://dx.doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360-372. doi:http://dx.doi.org/10.1002/acp.2914

Meltzer, T., & Bennett, J. H. (2014). *The Princeton Review: Cracking the AP U.S. History Exam.* New York: Random House.

Pan, S. C., Wong, C., Potter, Z., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*. doi:http://dx.doi.org/10.3758/s13421-015-0547-x

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.

Resnick, E. (2014). *Barron's AP United States History (2nd ed.)*. Hauppauge, NY: Barron's Educational Series.

Rickard, T. C., & Bajic, D. (2006). Cued recall from image and sentence memory: A shift from episodic to identical elements representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 734-748. doi: http://dx.doi.org/2006-08497-007

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1281-1295. doi:http://dx.doi.org/10.1037/0278-7393.22.5.1281

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1139-1153. doi:http://dx.doi.org/10.1037/0278-7393.20.5.1139

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233-239. doi:http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

Tran, R., Rohrer, D., & Pashler, H. (2014). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review.* doi:10.3758/s13423-014-0646-x

CHAPTER 3

Does Retrieval Practice Enhance Learning and Transfer Relative to Restudy for Term-Definition Facts?

Pan, S. C. and Rickard, T. C. (2017).

Journal of Experimental Psychology: Applied 23(3), 278-292.

Abstract

In many pedagogical contexts, *term-definition* facts which link a concept term (e.g., "*vision*") with its corresponding definition (e.g., "*the ability to see*") are learned. Does retrieval practice involving retrieval of the term (given the definition) or the definition (given the term) enhance subsequent recall, relative to restudy of the entire fact? Moreover, does any benefit of retrieval practice for the term transfer to later recall of the definition, or vice versa? We addressed those questions in four experiments. In each, subjects first studied term-definition facts and then trained on two-thirds of the facts using multiple-choice tests with feedback. Half of the test questions involved recalling terms; the other half involved recalling definitions. The remaining facts were either not trained (Experiment 1) or restudied (Experiments 2-4). A 48 hr delayed multiple-choice (Experiments 1-2) or short answer (Experiments 3a-4) criterial test assessed recall of all terms or all definitions. Replicating and extending prior research, retrieval practice yielded improved recall and positive transfer relative to no training. Relative to restudy, however, retrieval practice consistently enhanced subsequent term retrieval, enhanced subsequent definition retrieval only after repeated practice, and consistently yielded at best minimal positive transfer in either direction. Theoretical and practical implications are discussed.

Introduction

"What is vision?" "What is the ability to see called?" Those two questions refer to one another: the first asks for the definition of a term (e.g., *vision*), while the second asks for the reverse. The underlying fact that links these two questions ("*Vision is the ability to see*") is a *term-definition* fact: an "A-is-B" formatted fact that links a concept (i.e., term) with its corresponding definition. Term-definition facts are ubiquitous across numerous subject areas, ranging from commonly learned academic concepts (e.g., the definition of *photosynthesis* in biology) to more specialized domains (e.g., the definition of *aileron* in aeronautical engineering). Mastery of such facts provides a prerequisite foundation for thinking, problem solving, and other higher order skills (Willingham, 2009). To be competent in a given domain, one must master important terminology and concepts, and the learning of term-definition facts can be essential toward that end.

One promising method for learning term-definition facts is *retrieval practice*, a technique which involves attempting recall of to-be-learned information (e.g., by taking a practice test). The benefits of retrieval practice (also known as the *testing effect* and *test-enhanced learning*, among other appellations), relative to control tasks such as no training or restudy, have been demonstrated across a wide range of materials, for learners both young and older (Meyer & Logan, 2013), and across different levels of memory ability (Pan, Pashler, Potter, & Rickard, 2015). Accordingly, many learning researchers now classify retrieval practice as one of the most robust learning techniques available (e.g., Brown, Roediger, & McDaniel, 2014; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) and recommend its implementation in schools (e.g., Pashler et al., 2007).

The two basic methods of using retrieval practice for term-definition facts—namely to recall the term (*term-response*) given the definition or recall the definition (*definition-response*) given the term—are the focus of this manuscript. More specifically, we address three questions: (a) does retrieval practice in the form of term-response or definition-response confer memory benefits for the same response on a later test?; (b) are any such benefits observed relative to a restudy control condition?; and (c) do any such benefits generalize, or *transfer*, to the reversed case; that is, does learning due to a term-response test transfer to a definition-response test, and vice versa?

The comparison of retrieval practice against a restudy control condition (as opposed to a no training control, which has been used in several prior studies that are relevant to the present topic, as discussed later in this manuscript) addresses the important pedagogical issue of whether there is a benefit of practice testing over alternative learning strategies. Given a finite amount of available learning time, should learners use practice testing or study when learning term-definition facts? Study (and restudy) is used more often by students than is practice testing (e.g., Kornell & Bjork, 2007). Further, unlike a no training condition, the restudy condition allows control for time on task and for reexposure to materials (Carrier & Pashler, 1992).

To date, two studies have addressed the issue of learning and transfer for term-definition facts following retrieval practice (for an overview of the larger literature on retrieval practice and transfer of learning, see Carpenter, 2012; for a meta-analytic review, see Pan & Rickard, 2018). As described in the following section, both studies demonstrated positive transfer. Neither, however, addressed the case of transfer relative to a restudy control for the simplest case implementation of retrieval practice that is most likely to be used in educational contexts. Further, studies involving such basic implementations of retrieval practice on facts that do not explicitly have a term-definition structure have often shown no positive transfer relative to a restudy control (e.g., Agarwal, 2011; Hinze & Wiley, 2011; Pan, Gopal, & Rickard, 2016).

Retrieval Practice and Transfer Between Fact Terms

In a classroom study, McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013; Experiment 1) had 7th grade students take three identical multiple-choice clicker quizzes with feedback on term-definition facts over an approximately 11-day period. The quizzes covered two-thirds of the facts that students were learning across three academic subjects: cells, machines/energy, and animals. Half of the quiz questions required term retrieval, and half required definition retrieval. The remaining one-third of the materials was not practiced. A multiple-choice criterial test, administered one day after the third practice quiz, had three conditions: (a) questions that were identical to those that were used during practice quizzes (e.g., term-response on both), (b) *term-definition reversals* (e.g., if retrieval practice on a given fact involved term retrieval, the criterial test question on that fact involved definition retrieval, or vice versa, hence the label "reversal"), and (c) questions on facts that had not been trained. A substantial testing effect was observed relative to no training, as well as substantial transfer of that effect; that is, there was roughly equivalent performance in the first and second conditions, in the context of much poorer performance in the third, no training condition.

The results of McDaniel, Thomas, et al. (2013) suggest that retrieval practice holds substantial promise at enhancing learning and transfer for term-definition facts. However, the authors acknowledged that their experimental design may limit the generalizability of their findings. It was impossible in that experiment to fully control for time on task, outside study activities, and the effects of student motivation on the final exam (which accounted for half of the course grade). For example, it is likely that students attended particularly well to the feedback for incorrectly answered quiz questions, and used that feedback to guide their subsequent outside study activities. Further and perhaps most importantly, transfer was assessed against a no training control. Although the authors speculated that significant testing and transfer effects would still have been observed had a restudy control condition (which is expected to produce at least some learning) been used, that possibility was not empirically tested.

In relation to that last issue, McDaniel, Bugg, Liu, and Brick's (2015) study of retrieval practice and term-definition facts incorporated a restudy control. In each of two experiments, subjects studied a research methods text and then trained two days later on that text using test-test-test (TTT), test-study-test (TST), or study-study-study (SSS) schedules. Retrieval practice involved answering a multiple-choice term- or definition-response question comparable to those of McDaniel, Thomas, et al. (2013; Experiment 1), as well as a multiple-choice *application* question for each fact. The application question involved identifying the term which best matched a provided example (e.g., *"You take a test and score very high; a week later you take the same test and score moderately; this test lacks _____?"* for which the answer is *reliability*). Feedback was not provided in Experiment 1 but was provided in Experiment 2. On a short answer criterial test administered four days after training, positive transfer to term-definition reversals was observed for the TTT and TST groups in Experiment 1 and only for the TST group in Experiment 2, relative to restudy (the SSS condition).

McDaniel, Bugg, et al. (2015) reached a similar conclusion as McDaniel, Thomas, et al. (2013), namely that retrieval practice can result in transferrable learning for term-definition facts. However, that evidence of positive transfer was obtained with a multi-session paradigm in which training entailed repeatedly answering two types of questions per fact (a term- or definition-response question as well as an application question), plus (for the more potent TST condition) re-reading of text in between practice tests. Thus, the question of whether a more basic implementation of retrieval practice (i.e., without re-reading of materials between tests, additional application questions, or necessarily involving repeated testing) enhances transfer for term-definition facts, relative to restudy, remains unaddressed.

That simplified form of retrieval practice—henceforth referred to as *unelaborated* retrieval practice (i.e., practice testing with brief correct answer feedback but no other elaborative processing or added post-retrieval activities)—is characteristic of studies that use facts without a term-definition structure. In those studies, unelaborated retrieval practice typically does not enhance transfer from one part of a fact to another, relative to a restudy control. For example, Pan, Gopal, et al. (2016) had subjects study facts from Advanced Placement (AP) History and Biology courses (e.g., "*Jefferson purchased Louisiana from the French*") and then take practice tests with feedback on individual words from those facts (e.g., "*WHO purchased Louisiana from the French?*"). Across four experiments, they consistently found robust testing effects on criterial test questions that involved recall of the same words as during training. However, there was no evidence of positive transfer to criterial test questions that involved recall of different words from previously tested facts (e.g., *"Jefferson purchased WHAT from the French?"*), relative to restudy. That result occurred whether training was in short answer or multiple-choice format, with single or multiple practice trials per fact, and for history and biology facts. Accordingly, the authors concluded that retrieval practice confers potent, but piecewise, learning benefits for AP and similar types of facts.

The findings of Pan, Gopal, et al. (2016) build on prior work which suggests that unelaborated retrieval practice yields highly specific and nontransferable learning for complex facts (e.g., Hinze & Wiley, 2011) and other types of stimuli with multiple elements (e.g., word triplets as in Pan, Wong, Potter, Mejia, & Rickard, 2016). It thus appears that transfer from one response element to another following unelaborated retrieval practice is weak for stimuli with multiple terms or elements, and does not exceed the learning that occurs through restudy. None of those studies, however, used term-definition facts as materials.

Despite the counter-evidence discussed above, one reason to suspect that unelaborated retrieval practice for term-definition facts may yield transfer relative to restudy is that such transfer is observed for structurally analogous paired associates. Term-definition facts have an apparent two-element (i.e., *term* → *definition*) associative structure as is the case for paired associates (for prior analogies of fact learning to paired associates, see Mozer, Howe, & Pashler, 2004; Rohrer, Taylor, & Sholar, 2010; for related discussion see McDaniel, Thomas, et al., 2013). Carpenter, Pashler, and Vul (2006) demonstrated potent transfer of learning from one response element of a paired associate to another following cued recall practice tests (e.g., for the word pair *beach* → *blanket*, a cued recall test on *beach* → *?* resulted in better criterial test performance on *blanket* → *?*, relative to restudy). This finding was subsequently replicated by Cheng (2014) using a multiple-choice training format (in which the cue from a word pair was presented on each practice test trial, with subjects having to select one of four possible answers). These results suggest that retrieval practice may yield transfer when the to-be-learned materials involve forming associations between only two elements, as is the case, on the surface at least, with terms and their definitions.

The Current Study

To investigate whether unelaborated retrieval practice enhances learning and transfer for term-definition facts, we conducted four experiments. The design and procedures were adapted from the paradigm used in Pan, Gopal, et al. (2016). In contrast with prior work on term-definition facts, we did not pair each term- or definition-response practice test question with an additional application question, nor did we provide the opportunity for additional exposure to the facts outside the experiment. Our implementation of retrieval practice can be compared to similar implementations of practice test questions in educational contexts, such as those provided at the end of textbook chapters or given by an instructor in advance of a high-stakes final exam.

Experiment 1 was designed as a conceptual replication of McDaniel, Thomas, et al.'s (2013) results, in which retrieval practice and transfer effects were observed relative to a no training control. To foreshadow, the results fully replicated their findings. In Experiments 2-4 we investigated, for the first time, the learning of term-definition facts under conditions of unelaborated retrieval practice vs. a restudy control (which, as we argued above, is more informative from an educational perspective).

Experiment 1

The first experiment aimed to determine whether the positive transfer relative to a no training control, as observed by McDaniel, Thomas et al. (2013; Experiment 1), would replicate in a laboratory experiment with university students and for the case in which retrieval practice consisted of a single practice test trial with feedback per fact. Besides changes in setting, subjects, and amount of retrieval practice, there were four additional design differences between the prior and current experiments: use of university-level factual materials, no extra study opportunities outside the experiment, a retention interval of 48 hrs, and a between- rather than within-subjects manipulation of question type (term- or definition-response) on the criterial test.

Method

Subjects. In this and all subsequent experiments, undergraduate students recruited from the University of California, San Diego subject pool participated in exchange for course credit. The target sample size, which is comparable to that of prior retrieval practice and transfer studies (e.g., Hinze & Wiley, 2011; McDaniel & Fisher, 1991) and also applied to Experiment 2, was 50. Fifty-five undergraduate subjects participated. All but one subject completed both sessions of the experiment; data from the remaining 54 subjects (*term-response* group, n =28; *definition-response* group, n =26) was analyzed.

The entire study was conducted with the approval of the university's institutional review board, and all subjects provided written informed consent prior to participating.

Materials. Thirty-six term-definition facts, each averaging 17 words in length, were extracted from a widely-used undergraduate introductory neuroscience textbook, *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience* (Breedlove & Watson, 2013), and its publisher-provided test bank. Versions of this textbook have been used in prior retrieval practice studies (e.g., McDaniel, Anderson, et al., 2007). Each fact defined a concept in biology or neuroscience (e.g., *consciousness*). One to three facts were extracted from each of the textbook's nineteen chapters, with each fact having a term-definition "A-is-B" structure in which (a) the concept term was at the start of the sentence (either starting with the first word or after the articles "a", "an", or "the"), (b) that term was followed by the verb "is" or a comparable verb (e.g., "refers"), and (c) that verb itself was followed by the definition of the term. An example fact is: "*Consciousness is personal awareness of one's own emotions, thoughts, movements, and experiences*." Terms from all 36 facts are listed in Appendix 3A.

For each fact, two multiple-choice questions were created: a term-response question and a definition-response question. For both questions, four answer choices were created: the correct answer and three plausible lures. The correct answers and lures did not overlap with those for any of the other facts. Examples of facts, test questions, answers, and lures are included in Appendix 3B.

For counterbalancing purposes, three sub-lists of 12 facts each were randomly drawn from the full set of 36 facts, with no fact overlap across sub-lists. Six experimental lists of 24 facts each were then created using all possible pairings of two of the three sub-lists. Within each experimental list, one sub-list of facts were assigned to be trained using term-response test questions, and the other sub-list to be trained using definition-response test questions. An additional 12 facts not in that experimental list were assigned to the no training condition. This design enabled each fact to appear in each of the three training conditions (i.e., trained with a term-response question, trained with a definition-response question, or not trained at all) with equal frequency over subjects.

Design and procedure. The experiment entailed two sessions separated by a 48 hr delay. As shown in Figure 3.1, Session 1 contained two phases: the study phase and the training phase. During the study phase, subjects viewed all 36 facts, one at a time, for 10 seconds each, and in a random order determined anew for each subject. All facts were studied once.

The training phase, in which each subject trained on one experimental list, followed immediately afterward. Subjects were assigned to one of the six experimental training lists in counterbalanced fashion. There was one training phase test trial per fact (24 trials total), and test questions were presented in random order determined anew for each subject. For each training phase test question, subjects were given 12 s to select one of the four possible answer choices by typing the letters A, B, C, or D. Answer choice order was randomized on each trial for each subject. After 12 s had elapsed, the correct answer was subsequently shown for 3 s. That correct answer constituted feedback. When the feedback appeared, the test question, the four possible answer choices, and the subject's selected response remained on screen, thus providing subjects with all relevant information with which to study the feedback (similar to McDaniel, Thomas, et al., 2013). After the training phase ended, subjects were reminded of their Session 2 appointment, falsely informed that they would return to study new facts (a deception designed to prevent outside study), and dismissed.

In Session 2, which occurred after a 48 hr delay, subjects completed the criterial test. The question type on the criterial test (term- or definition-response) was manipulated between-subjects, with random assignment to either the term-response group or the definition-response group. Based on this assignment, subjects were assessed on recall of either all terms or all definitions for each of the 36 facts, with one test trial per fact. On each trial, a multiple-choice test question and four possible answer choices were presented, and subjects input their answer by typing the letters A, B, C, or D. As occurred during Session 1, answer choice order was randomized on each test trial for each subject. No feedback was provided on the criterial test, and subjects had unlimited time to respond on each trial.

Data analysis. Criterial test data (i.e., proportion correct on criterial test questions) were analyzed using a 3 (Criterial Test Condition: tested-same, tested-different, and not trained) x 2 (Question Type: term-response and definition-response) mixed factors design, with Criterial Test Condition and Question Type as within- and between-subjects variables, respectively. *Tested-same* indicates that training and criterial test questions for a given fact were identical (e.g., term-response), while *tested-different* indicates that training and criterial test questions for a given fact were different (i.e., term-definition reversals, such as term-response for training and definition-response on the criterial test).

Results

Training. Proportion correct was M = 0.80, SE = 0.021 for term-response questions and M = 0.70, SE = 0.021 for definition-response questions. The higher accuracy on term-response questions was statistically significant (here and in all subsequent analyses we use $\alpha = 0.05$), $t(53) = 4.35$, $p < .0001$, $d = 0.59$. That pattern suggests that recall of definitions is more difficult than recall of terms, matching an observation previously made by Lipko-Speed, Dunlosky, and Rawson (2014) and patterns observed in McDaniel, Thomas, et al. (2013).

Criterial test. The mean proportion correct across the two question types and the three criterial test conditions is shown in Figure 3.2, Panel a. For comparison, the results of McDaniel, Thomas, et al. (2013; Experiment 1) are depicted in Panel b. A factorial Analysis of Variance (ANOVA) on subject-level proportion correct scores yielded a statistically significant effect of Criterial Test Condition, $F(2, 104) = 30.82$, $MSE = 0.34$, $p < .0001$, $\eta_p^2 = 0.37$, no significant effect of Question Type, $F(1, 52) = 1.12$, $MSE = 0.070$, $p = .29$, $\eta_p^2 = 0.021$, and no significant Criterial Test Condition by Question Type interaction, $F(2, 104) = 2.07$, $MSE = 0.023$, $p = .13$, $\eta_p^2 = 0.038$.

Comparison of Panels a and b shows that the criterial test results of Experiment 1 are highly analogous to those of McDaniel, Thomas, et al. (2013; Experiment 1). Most critically, in both experiments there were large performance differences between the tested-same and tested-different conditions on one hand, and the no training condition on the other, a result that was observed for both the term- and definition-response groups. Hence, in both experiments there was substantial evidence that retrieval practice yielded transfer relative to a no training control. The only statistical difference between the experiments was marginally better performance in the tested-same than in the tested-different condition for the term-response group in our experiment, $t(27) = 2.17$, $p = .039$, $d = 0.41$, but not in the McDaniel, Thomas, et al. experiment.

Effect of prior course experience. In exit surveys, well over half ($n = 38$) of the subjects reported having previously taken a biological psychology or neuroscience course. While prior experience translated into better overall performance during training ($M = 0.77$, $SE = 0.070$ vs. $M = 0.71$, $SE = 0.084$) and on the criterial test ($M = 0.73$, $SE = 0.069$ vs. $M = 0.63$, $SE = 0.067$), there was no effect of that expertise difference in the pattern of testing and transfer effects. Moreover, in Experiments 2-4 (for which the percentage of subjects reporting prior course experience was 47%, 45%, 17%, and 47%, respectively), a similar pattern of better overall performance for experienced subjects was also observed, but again that improved performance did not change the resulting testing effect and transfer patterns. Results of exit surveys are not discussed further.

Discussion

The results of Experiment 1 confirm that, relative to a no training control, retrieval practice (unelaborated practice in this case) enhances criterial test performance for both the previously tested response (i.e., a testing effect) and the reverse response (i.e., a transfer effect). That replication and extension of the findings of McDaniel, Thomas, et al. (2013) shows that positive testing and transfer effects for term-definition facts relative to a no training control are robust across subject populations, different implementations of retrieval practice, and the various other differences in experimental design.

Experiment 2

In this experiment we explored the question, heretofore unaddressed, of whether unelaborated retrieval practice enhances learning and transfer for term-definition facts relative to a restudy control. This experiment retained all design and procedural features of Experiment 1 with one exception: the control condition was changed from no training to restudy. From the applied perspective, this comparison of retrieval practice relative to a non-retrieval reference task is arguably more important than comparison to a no training control.

Method

Subjects. Sixty-one undergraduate students participated for course credit. All but two subjects completed both sessions of the experiment; data from the remaining 59 subjects (*term-response* group, n = 31; *definition-response* group, n = 28) was analyzed.

Materials, design, and procedure. All aspects of this experiment's design and procedure were identical to its predecessor (see Figure 3.1), with the exception that the 12 facts that were not trained during the training phase of Session 1 were instead restudied. Each restudy training trial involved the presentation of an entire fact for 15 s each. Thus, the total duration of each restudy trial during training was identical to each practice test trial. Restudy trials were randomly mixed with term- and definition-response practice test trials, for a combined total of 36 training trials (i.e., one trial per fact). As in the preceding experiment, each fact appeared in each training condition (i.e., trained with a term-response question, trained with a definition-response question, or restudied) with equal frequency over subjects, and question type on the criterial test (term-response or definition-response) was manipulated between-subjects.

Results and Discussion

Training. Accuracy on the initial test was as follows: term-response questions, M = 0.71, SE = 0.037; definition-response questions, M = 0.62, SE = 0.028. Performance on term-response questions during training was again significantly better than definition-response questions, $t(58) = 3.99$, $p < .001$, $d = 0.52$.

Criterial test. An ANOVA identical to that performed for Experiment 1 yielded a statistically significant effect of Criterial Test Condition, $F(2, 114) = 5.60$, $MSE = 0.071$, $p = .0048$, $\eta_p^2 = 0.089$, no significant effect of Question Type, $F(1, 57) = 0.004$, $MSE = 0.00033$, $p = .95$, $\eta_p^2 < 0.001$, and a significant Criterial Test Condition by Question Type interaction, $F(2, 114) = 3.80$, $MSE = 0.049$, $p = .025$, $\eta_p^2 = 0.063$. Inspection of Figure 3.3, Panels a and b provides insight into that interaction. For the term-response group (Panel a) there was a testing effect (tested-same vs. restudied), $t(30) = 3.78$, $p < .001$, $d = 0.68$, but there was no observable transfer (tested-different vs. restudied), $t(30) = 1.18$, $p = .25$, $d = 0.21$. For the definition-response group (Panel b), performance was statistically indistinguishable across conditions; there was no evidence for either a testing effect or a transfer effect.

A cross-experiment analysis of Experiments 1 and 2 confirmed that the transfer results depend critically on control task. In a 2 x 2 x 2 factorial ANOVA with factors of Experiment (1 vs. 2), Criterial Test Condition (control and tested-different conditions only, a comparison which focuses specifically on the transfer issue), and Question Type (term- or definition-response), there was a significant main effect of Criterial Test Condition, F(1,

109) = 24.57, MSE = 0.32, p < .00001, η_p^2 = 0.18, a significant Experiment by Criterial Test Condition interaction, $F(1, 109)$ = 10.47, MSE = 0.14, p = 0.0016, η_p^2 = 0.088, and no other significant main effects or interactions (ps < .41). The Experiment by Criterial Test Condition interaction confirms the large differences in transfer results that are evident in a comparison of Panel a of Figure 3.2 vs. Panels a and b of Figure 3.3. Specifically, transfer results are markedly different when measured against no training than against restudy.

The results described above suggest that the bulk of the test-enhanced learning and transfer effect in the prior term-definition studies (McDaniel, Bugg, et al., 2015; McDaniel, Thomas, et al., 2013) reflects either use of a no training control, or, in the former study, use of various forms of elaborated retrieval practice. It appears that retrieval practice itself, at least in the form of a multiple-choice test with only brief correct answer feedback, can produce a testing effect, at least for term retrieval, but does not yield a transfer effect.

Experiments 3a and 3b

In the first two experiments we employed multiple-choice initial and criterial tests, just as did McDaniel, Thomas, et al. (2013). In Experiments 3a and 3b, we switched to a short answer criterial test format, similar to McDaniel, Bugg, et al. (2015), while retaining multiple-choice practice test questions. The change in criterial test format was motivated by two considerations. First, whereas multiple-choice tests are commonly used in educational settings, and hence constitute a viable test format for the training test, short answer questions are far more common outside of classroom settings, and therefore constitute a more appropriate format for the criterial test. Second, if a short answer criterial test yields larger testing effects than does a multiple-choice test (e.g., Kang et al., 2007; for related discussion see Carpenter & DeLosh, 2006; Halamish & Bjork, 2011), then such a test may also be more sensitive to possible transfer effects.

Method

Experiments 3a and 3b were completed over non-overlapping date ranges using independent samples from the same subject pool. Experiment 3a involved only term-response criterial test questions, whereas Experiment 3b involved only definition-response questions. Otherwise those two experiments were identical.

Subjects. Although the small standard errors in Experiment 2 indicate that its results are reliable, we chose to roughly double the sample size in Experiments 3a and 3b for increased statistical power. One-hundred four undergraduate students, all recruited from the same population as in the prior experiments, participated for course credit. All but three subjects in Experiment 3a completed the entire study. Experiments 3a and 3b each involved the same target sample size, 50, as in the preceding experiments. Data from 49 subjects were collected in Experiment 3a (*term-response* only), and data from 52 subjects were collected in Experiment 3b (*definition-response* only).

Materials, design, and procedure. Both experiments were largely identical to the respective term- or definition-response groups of Experiment 2 (see Figure 3.1), with the primary exceptions that (a) the criterial test involved short answer test questions instead of multiple-choice, and (b) ten of the facts were shortened with easier to score one word terms (e.g., *absolute refractory period* was changed to *refractory*) and practice and criterial test questions altered to match. The short answer questions on the criterial test were minimally modified versions of their multiple-choice counterparts (see Appendix 3B for examples).

Prior to the first criterial test trial, subjects were told that their typed responses should be spelled as accurately as possible, but that if they were unsure of the correct spelling, to still make their best possible attempt at an answer. The format of the short answer test questions was as follows: the question appeared on the screen, while an empty text box with a cursor appeared directly underneath. Subjects typed their answer and pressed the Enter key to advance to the next trial. As in the preceding experiments, subjects had unlimited time to enter a response on each criterial test trial.

Data coding and analysis. The nature of typed short answer responses, in which spelling and grammatical errors were possible and indeed often occurred, necessitated a different scoring procedure than in the preceding experiments. Moreover, as term- and definition-response answers differ in length (single word vs. sentence-length answers), dedicated scoring procedures were developed for each.

Term-response scoring. To avoid penalizing subjects for misspelled responses that could be unambiguously identified as referring to correct answers, all criterial test responses in Experiment 3a were analyzed in Microsoft Excel using the Fuzzy Lookup (Microsoft Research, Redmond, WA) add-in (cf. Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005) before statistical analyses were performed. This add-in compares each response to a master list of correct answers, computes a Jaccard similarity score (Microsoft Corporation, 2011), and if a close match is found, identifies the most closely related and accurately spelled answer choice (in this case, the add-in was configured to do so for similarity scores of 0.6 or higher). A comparison with traditional human scoring methods, which involved scoring one-fifth of all responses using the add-in as well as a human rater, indicated that reliability was high (0.98) between the add-in and human scoring. After close misspellings were analyzed and corrected, we used a letter matching algorithm to score for accuracy (where an exact match was scored as correct and all other cases were scored as incorrect).

Definition-response scoring. To score the sentence-length responses in Experiment 3b, and to provide leeway for unambiguously identifiable misspellings, a point-based coding method was developed. Under this method, three to five idea units per fact were identified (cf. Lipko-Speed et al., 2014; Rawson & Dunlosky, 2011; McDaniel, Bugg, et al., 2015), with each idea unit allotted one point. Scoring involved reading each response and assigning yes/no ratings for each idea unit per fact. Eight research assistants trained on this method were assigned up to five facts each; each assistant scored all of the responses in Experiment 3b for their assigned facts while remaining blind as to training condition assignment. To verify the consistency of this scoring technique, raters also scored an identical subset of the data (comprised of a sample of 5% of the responses for each of the 36 facts). These scores where then compared with scores generated by the authors; there was high reliability (≥ 0.80) between the raters and the authors. For statistical analyses, scores for each response (i.e., number of yes ratings divided by total number of points possible) were dichotomously transformed (i.e., into correct or incorrect, which corresponds with the scoring outcomes on the term-response criterial test) using the following criterion: > 60% of idea units had to be correctly recalled in order for a response to be scored as correct. Summary data for analyses using continuous scores (i.e., untransformed data) are also reported. To foreshadow, the use of either dichotomous or continuous scores did not change the overall pattern of results.

Results and Discussion

Training. Proportion correct on the multiple-choice test in Experiment 3a was M = 0.76, SE = 0.023, for term-response questions and M = 0.72, SE = 0.024 for definition-response questions. While mean performance was numerically better for term- vs. definition-response questions, matching the pattern observed in the prior experiments, in this case the difference was not statistically significant, $t(48) = 1.66$, $p = .10$, $d = 0.24$. In Experiment 3b, accuracy on the initial test was M = 0.68, SE = 0.027 for term-response questions and M = 0.62, SE = 0.030 for definition-response questions. Mean performance on term-response questions was significantly better than on definition-response questions, $t(51) = 2.79$, $p = .0074$, $d = 0.39$, again matching the pattern observed in prior experiments.

Criterial test. Criterial test results for Experiment 3a are shown in Figure 3.4, Panel a. A within-subjects one-way ANOVA on proportion correct scores yielded a statistically significant main effect of Criterial Test Condition, $F(2, 96) = 6.08$, $MSE = 0.12$, $p = .003$, $\eta_p^2 = 0.10$. Inspection of Figure 3.4 reveals that this effect is driven primarily by a testing effect for the tested-same condition in the absence of substantial transfer to the tested-different condition. That pattern was confirmed by two tests. The first, limited to the tested-different and restudied conditions, yielded no significant effect of Criterial Test Condition, $t(48) = 0.62$, $p = .53$, $d = 0.089$. The second, limited to the tested-same and tested-different conditions, yielded a significant effect of Criterial Test Condition, $t(48) = 2.73$, $p = .009$, $d = 0.39$. Experiment 3a thus fully replicated Experiment 2: for the case of unelaborated retrieval practice vs. a restudy control, testing enhances term retrieval, but that enhancement does not transfer to the reverse case of definition retrieval. The same ANOVA for Experiment 3b (with dichotomous scores, as shown in Figure 3.4, Panel b) yielded no significant main effect of Criterial Test Condition, $F(2, 102) = 2.29$, $MSE = 0.026$, $p = .11$, $\eta_p^2 = 0.043$, while the same ANOVA using continuous scores (mean accuracy (*SE*) of 0.30 (0.020), 0.28 (0.022), 0.28 (0.018) in the tested-same, tested-different, and restudied conditions, respectively) also yielded no significant main effect of Criterial Test Condition ($p = .70$). Thus, just as for Experiment 2, there was no statistical evidence for a testing or transfer effect for definition-responses. It should be noted, however, that there was a numerical trend in favor of a testing effect in Experiment 3b.

A cross-experiment analysis indicated that performance was better overall for the case of term-response (Experiment 3a) than definition-response (Experiment 3b) short answer criterial tests. In a 2 x 3 factorial ANOVA with factors of Experiment (3a vs. 3b; using dichotomous scores) and Criterial Test Condition (tested-same, tested-different, and restudied), there was a significant main effect of Experiment, $F(1, 99) = 25.16$, $MSE = 1.90$, $p < .0001$, $\eta_p^2 = 0.20$, and Criterial Test Condition, $F(2, 198) = 8.25$, $MSE = 0.13$, $p < .0001$, $\eta_p^2 = 0.87$, and no Experiment by Criterial Test Condition interaction, $F(2, 198) = 1.24$, $MSE = 0.019$, $p = .29$, $\eta_p^2 = 0.012$. The main effects are apparent upon examination of Panels a and b of Figure 3.4: performance was better overall on the term-response criterial test, and performance in the tested-same condition tended to be better than the tested-different and restudied conditions in both experiments. Moreover, the absence of an interaction indicates that the relative performance among the tested-same, tested-different, and restudied conditions did not significantly differ between the two experiments (although, as noted previously, pairwise analyses did find evidence of a testing effect for the case of term-response in Experiment 3a, but not for definition-response in Experiment 3b).

Overall, Experiments 3a and 3b fully replicate Experiment 2 with respect to both testing and transfer effects. There was statistical evidence of a testing effect for term- but not definition-responses on short answer criterial tests, and there was again no evidence of positive transfer in either direction.

Experiment 4

For the fourth experiment we investigated whether increasing the number of practice test trials from one to three per fact yields different learning and/or transfer results. The use of repeated test trials was motivated in part by the lack of a statistically significant testing effect for definition retrieval in Experiments 2 and 3b. Many demonstrations of the testing effect use repeated training tests (e.g., Butler, 2010; Pan, Rubin, & Rickard, 2015; Roediger, Agarwal, McDaniel, & McDermott, 2011), and it has been suggested the repeated testing strengthens the benefits of retrieval practice (e.g., McDaniel, Thomas, et al., 2013), including when the training test format is multiple-choice (e.g., McDaniel, Wildman, & Anderson, 2012; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014). Thus, by increasing the training "dosage" in this experiment, we sought to increase experimental sensitivity to detect a testing effect for definition retrieval, if it exists in the population. If we observe a testing effect for definition retrieval, then we will also be in a position to further address the issue of transfer for that case.

Method

Subjects. The minimum sample size to detect a small transfer effect was determined using a priori power analysis. Based on the standard deviation of the tested-different minus restudied condition proportion correct difference scores on the criterial test of Experiment 3b, a sample size of at least 54 per group is needed to achieve power of 0.8 or greater to detect a proportion correct difference score of 0.05 or greater (based on a one-tailed, one-sample t test, $\alpha = 0.05$). Accordingly, 124 undergraduate students, all recruited from the same population as in the prior experiments, participated for course credit. All but eleven subjects completed both sessions of the experiment; data from the remaining 113 subjects (*term-response* group, $n = 54$; *definition-response* group, $n = 59$) was analyzed. Subjects were randomly assigned to the term- or definition-response group.

Materials, design, procedure, and data coding. Nearly all aspects of this experiment's design and procedure were identical to that of Experiments 3a and 3b, with the primary exception being that training involved three practice test or restudy trials per fact. This was accomplished by presenting each of the 36 facts once per training block across three training blocks. Assignment of fact to training condition was kept consistent across all three blocks. The design of each training block was functionally identical to that used in the preceding experiment, with the sole exception being that paraphrased questions, reworded lures, and slightly reworded correct definition responses were used for test questions on each block (examples are included in Appendix 3C), along with paraphrased facts. The purpose of presenting subjects with modified questions and answers on each training block was to encourage careful reading of each question or fact on each practice trial. The scoring procedure for the term- and definition-response criterial tests was identical to that used in the preceding two experiments; as before, analyses of definition-response criterial test data using dichotomous scores are reported in their entirety and analyses using continuous scores are summarized (the use of either scoring methods did not change the overall pattern of results).

Results and Discussion

Training. As shown in Figure 3.5, accuracy for both term-response and definition-response questions improved across the three blocks of the initial test. A 2 x 3 factorial ANOVA with factors of Question Type (term- or definition-response) and Block (1 vs. 2 vs. 3) yielded a significant main effect of Question Type, $F(1, 112) = 167.9$, $MSE = 4.45$, $p < .0001$, $\eta_p^2 = 0.60$, and Block, $F(2, 224) = 132.9$, $MSE = 1.86$, $p < .0001$, $\eta_p^2 = 0.55$, as well as a significant Question Type by Block interaction, $F(2, 224) = 3.47$, $MSE = 0.040$, $p = .033$, $\eta_p^2 = 0.030$. Performance on term-response questions was better than definition-response questions, in line with the pattern observed in the preceding experiments. The pattern of improvement across blocks indicates that repeated testing generated additional learning, and the magnitude of block-to-block improvements (from the first to the third block, proportion correct improvement of $M = 0.20$ for term-response and $M = 0.15$ for definition-response) was slightly greater for term-response questions.

Criterial test. An ANOVA identical to that performed for Experiments 1 and 2 (and using dichotomous scores for definition-responses; results shown in Figure 3.6) yielded a statistically significant effect of Criterial Test Condition, $F(2, 222) = 37.19$, $MSE = 0.56$, $p < .0001$, $\eta_p^2 = 0.25$, no significant effect of Question Type, $F(1, 111) = 0.73$, $MSE = 0.082$,

p = .39, η_p^2 < 0.01, and a significant Criterial Test Condition by Question Type interaction, $F(2, 222) = 5.25$, $MSE = 0.079$, $p = .006$, $\eta_p^2 = 0.045$. The same ANOVA using continuous scores for definition-responses (mean accuracy (*SE*) of 0.49 (0.017), 0.43 (0.022), 0.42 (0.022) in the tested-same, tested-different, and restudied conditions, respectively) also yielded the same pattern: a significant effect of Criterial Test Condition and a significant Criterial Test Condition by Question Type interaction (ps < .0001). As inspection of Figure 3.6 shows, there were highly significant performance differences between the tested-same and tested-different conditions (excluding the restudied condition) in both the term-response group, $t(53) = 7.11$, $p < .00001$, $d = 0.97$, and the definition-response group, $t(58) = 2.81$, $p = .007$, $d = 0.37$. However, there was again minimal positive transfer from term retrieval to definition retrieval, or vice versa; as evident in Figure 3.6, mean proportion correct in the restudy conditions was within about one standard error of that in the tested-different conditions. Thus, there were testing effects in this experiment for both term and definition retrieval, but at best minimal transfer in either case.

The source of the Criterial Test Condition by Question Type interaction is evident upon comparison of Panels a and b of Figure 3.6: the testing effect for the term-response group is larger than that for the definition-response group. This was confirmed statistically by comparing the difference scores for the tested-same minus tested-different conditions for both groups; the mean difference score in the term-response group was significantly larger, $t(110) = 2.63$, $p < .01$, $d = 0.49$. That finding of a larger testing effect for term- vs. definition-response is broadly consistent with the fact that, at the lower dosage levels of

Experiments 2, 3a, and 3b, testing effects were observed on the term- but not definition-response criterial tests.

General Discussion

In the present experiments we investigated the utility of unelaborated retrieval practice for learning term-definition facts, including whether it generates testing effects for term- and definition-responses, and whether it yields transfer from terms to corresponding definitions and vice versa. Experiment 1 conceptually replicated McDaniel, Thomas, et al. (2013; Experiment 1), in which positive testing effects and strong transfer was observed with middle school students relative to a no training control condition, for both term- and definition-response criterial tests. Experiment 1 also extended that result from children to adult students, and over changes in setting, materials, extent of training, and other differences in experiment procedure. Thus, testing and transfer effects for term-definition facts relative to a no training condition appear to be robust. However, when the control condition was switched to restudy in Experiments 2-4, there was at best weak evidence for transfer from either term to definition retrieval or the reverse. Further, percent transfer (bounded by restudy and tested-same performance) decreased rather than increased in Experiment 4, suggesting that transfer wanes at dosages at or beyond that which yields robust testing effects for both term and definition retrieval.

Term-Definition Facts and Rearranged Stimulus-Response Elements

Our findings indicate that term-definition facts should be added to the list of materials for which transfer to rearranged stimulus-response elements (i.e., where the required response was previously a cue, and vice versa) on the criterial test is minimal or absent relative to a non-testing reexposure control (e.g., restudy), and particularly when the practice tests do not involve extensive feedback in the form of re-reading (e.g., a text passage) or application questions. Those materials include multi-sentence college biology facts (Hinze & Wiley, 2011), AP History and Biology facts (Pan, Gopal, et al., 2016), and word triplets (Pan, Wong, et al., 2016). Analogous failure of transfer has also been observed for both adult's (Rickard & Bourne, 1996) and children's single-digit arithmetic (Walker, Bajic, Mickes, Kwak, & Rickard, 2014). Overall, it appears that specific and nontransferable learning for rearranged stimulus-response components on a criterial test is a likely outcome of unelaborated retrieval practice for a wide range of facts and other materials with multiple testable components. As detailed in the following section, the only known exception to this pattern involves an even simpler type of stimuli, namely paired associates.

Term-Definition Facts vs. Paired Associates

Some researchers have noted the existence of structural similarities between paired associates and facts (e.g., McDaniel, Thomas, et al., 2013; Mozer et al., 2004; Rohrer et al., 2010), and have speculated that such similarities may drive comparable learning and transfer processes. That hypothesis is motivated by the fact that both paired associates and facts contain two components to be associated (for paired associates: *cue* → *target*; for term-definition facts: *term* → *definition*). Of most interest here, practice test questions involving term-definition facts appear to reflect that structure. That is, the learner is given the term and asked for the full corresponding definition, or is given the full definition and asked for the corresponding term.

In contradiction to that hypothesis is the current finding that, whereas strong positive transfer (in some cases up to 100%) is consistently observed following various forms of retrieval practice on paired associates (e.g., in multiple-choice or cued recall format, and with single or repetition practice; Carpenter et al., 2006; Cheng, 2014; Vaughn & Rawson, 2014), little or no transfer is observed for term-definition pairs. It thus appears that the greater complexity of term-definition factual materials and other authentic educational facts leads to associative learning processes that are different in important respects than those involved in paired associate learning (a possibility also considered by Rohrer et al., 2010, and McDaniel, Thomas, et al., 2013, and now empirically supported by the current work).

The different learning and transfer properties for term-definition facts vs. paired associates may be related to the fact that the definition component of a term-definition fact contains multiple words that are unlikely to have been strongly associated with each other prior to training, and are exceedingly unlikely to have constituted a single, or chunked, memory representation as in the case of a familiar word (other stimuli with three or more elements may also exhibit similar associative properties; e.g., Pan, Wong, et al., 2016). Thus, a simple bi-directional association between a familiar stimulus and response (as presumably forms for word pairs) is not sufficient for learning. The possibility that learning term-definition facts involves forming multiple memory associations between the constituent words (or concepts) of each fact may also account for performance disparities between term and definition retrieval: on term-response questions, multiple words comprising the definition are presented as cues for a single to-be-retrieved target, whereas on definition-response questions, only a single cue (i.e., the concept term) is presented for multiple to-be-retrieved targets. Retrieval is likely to be more difficult in the latter case (given that there are no additional cues to help facilitate recall, and also because multiple targets need to be retrieved, rather than a single word as in the case of term retrieval). Thus, repeated retrieval practice (as in Experiment 4) may be necessary to strengthen associations between the single presented cue and its multiple targets in order to generate a testing effect for definition responses relative to restudy (alternatively, as pointed out by a reviewer, a more overtly effortful retrieval format such as cued recall may be used). Follow-up theoretical work that investigates the differences between paired associates and term-definition facts (and other materials that may have a similar paired structure, such as number digit or sound pairs), as well as the differences between concept term and definition retrieval, is warranted.

Practical Implications for Learning Term-Definition Facts

Retrieval practice has potent memory benefits relative to restudy across a wide range of materials, including those explored here. That fact, along with recent demonstrations of positive transfer of highly elaborated retrieval practice for term-definition facts (McDaniel, Bugg, et al., 2015; McDaniel, Thomas, et al., 2013), might lead educators to assume that retrieval practice will generally also give rise to transfer. More specifically, an instructor who wishes to incorporate evidence-based learning techniques might assign term- or definition-response practice test questions, expecting that transfer effects will result relative to alternative exercises. Similarly, a student could devote time to answering term- or definition-response practice test questions at the end of a textbook chapter, expecting that the resultant learning will generalize to entire facts. Our findings indicate that while unelaborated retrieval practice will generate learning dividends relative to a restudy strategy for term retrieval practice and (with repeated training) definition retrieval practice, it is unlikely to yield enhanced recall for unretrieved portions of such facts. This is an important consequence of retrieval practice on term-definition facts—a consequence of which both instructors and students should be aware.

Will the same testing and transfer effects occur if training involves a potentially more effortful retrieval method such as cued recall (i.e., short answer) rather than multiple-choice, as suggested by Kang et al. (2007)? Recent work provides insights. In Pan, Gopal, et al. (2016), both single and repeated trials of short answer retrieval practice consistently generated substantial testing but no transfer effects for the recall of terms from biology and history facts. Moreover, similar testing effects were observed in that study when the practice test format was switched to multiple-choice. Given the analogous results when the Pan, Gopal, et al. paradigm was adapted for the present experiments, it is unlikely in our view that the use of short answer practice tests will appreciably change the effects of unelaborated retrieval practice (with respect to either the testing effect or transfer) for term-definition facts.

The practical implications enumerated here apply specifically to the case of unelaborated retrieval practice involving term or definition retrieval. That method constitutes the most basic and least time intensive implementation of retrieval practice, and it is the most likely to be used in current educational practice (particularly since multiple-choice tests are relatively easy to score). However, as detailed in the Introduction, more elaborate forms of retrieval practice such as practicing on term- or definition-response and application questions for each term-definition fact (as occurred in McDaniel, Bugg, et al., 2015) appear to hold promise for facilitating positive transfer. This may be due to additional or different cognitive processes engendered by such methods (for related discussions see Hinze, Wiley, & Pellegrino, 2013; Jensen, McDaniel, Woodard, & Kummer, 2014; McDaniel et al., 2012). Other elaborate forms of retrieval practice that incorporate techniques such as spacing or criterion level learning (e.g., Rawson, Dunlosky, & Sciartelli, 2013), as well as repeated quizzing in conjunction with extended feedback opportunities and outside study (e.g., McDaniel, Anderson, et al., 2007; McDaniel, Wildman, et al., 2012) may also yield greater amounts of transfer for similar types of (i.e., factual) materials. However, such elaborated retrieval practice is more time intensive, and it is unclear whether the transfer results engendered exceed the gains that would be observed if that extra time were instead allocated to retrieval practice on both term- and definition-responses. Further research along those lines is needed to optimize uses of both unelaborated and elaborated retrieval practice for learning term-definition facts.

Finally, it should be noted that the current study investigated one educationally valid transfer context (i.e., transfer between fact elements) out of many possible contexts. Instructors and students may also be interested in transfer under different circumstances that do not necessarily involve term or definition retrieval (for examples see Carpenter, 2012). For instance, McDermott et al. (2014; Experiment 3) investigated retrieval practice and transfer of factual knowledge to application questions (i.e., generalizing prior knowledge to new examples); in that case, positive transfer was observed. Further, in a recent meta-analytic review of the retrieval practice and transfer literature, Pan and Rickard (2018) observed that the extent of transfer in that literature varies substantially between different transfer contexts; transfer involving rearranged stimulus-response elements is generally negligible (especially when elaborate forms of retrieval practice are not used), but more substantial transfer is evident for multiple other transfer types (e.g., application and inference questions). Thus, while the extent of transfer in the present study also proved to be minimal, that result does not preclude different findings involving similar materials but under different transfer contexts.

Conclusions

In the present work we demonstrated that unelaborated retrieval practice on term-definition facts, involving recall of previously studied responses followed by brief correct answer feedback, is superior to restudy. However, that retrieval practice benefit does not transfer from either term retrieval to definition retrieval or the reverse. That absence of transfer extends prior results for history and biology facts without a term-definition structure. In conjunction with other recent work, it now appears that, across a wide range of educational materials, retrieval practice *itself* (i.e., excluding any influence of elaborative post-retrieval processing) cannot be expected to yield transfer to rearranged stimulus-response criterial test questions. From the practical standpoint, students and instructors should not expect that taking a practice test on part of a term-definition fact, or any other type of fact with multiple testable terms, will be sufficient to enhance memory (relative to a study strategy) on a later test wherein a different response from the fact is required. Nevertheless, retrieval practice appears to be the best overall learning strategy because it can clearly facilitate later memory for the practiced response.

Acknowledgments

This chapter, in full, is reprinted with permission from the American Psychological Association, with the required notice embedded in the remainder of this paragraph. The official citation that should be used in referencing this material is: Pan, S. C. and Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied 23*(3), 278-292. The dissertation author was the primary investigator and author of this paper.

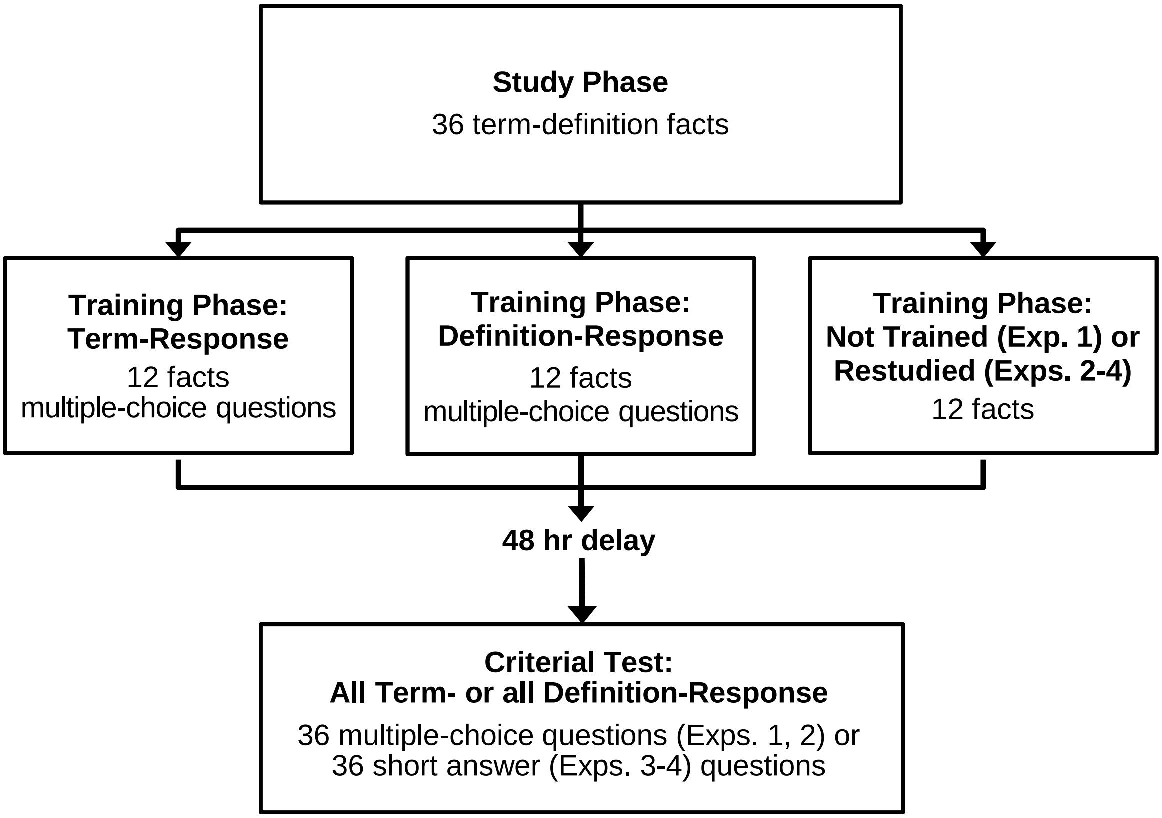


Figure 3.1. Procedure for Experiments 1-4. Session 1 began with the Study Phase, wherein all 36 facts were studied, followed by the Training Phase, wherein 12 facts each were trained using one (Experiments 1-3b) or three (Experiment 4) *term-response* or a *definition-response* test question(s) with feedback; the remaining 12 facts were either not trained (Experiment 1) or restudied (Experiments 2-4). After 48 hrs, subjects returned for Session 2, the Criterial Test, and were tested on all term-response or definition-response questions in multiple-choice (Experiments 1, 2) or short answer format (Experiments 3a-4).

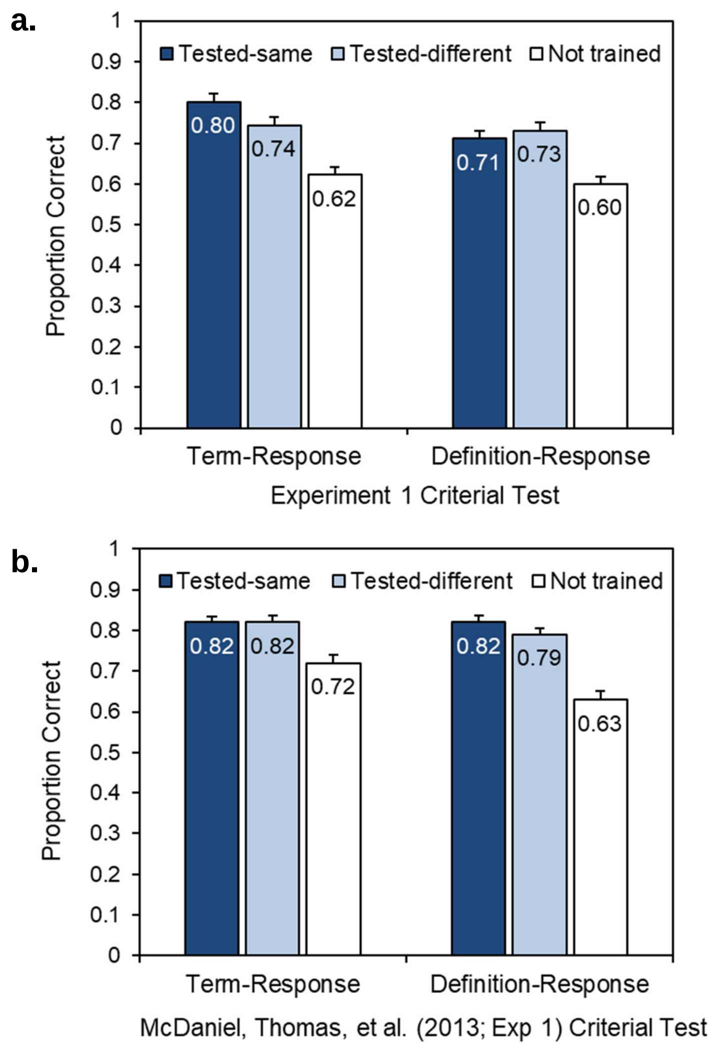
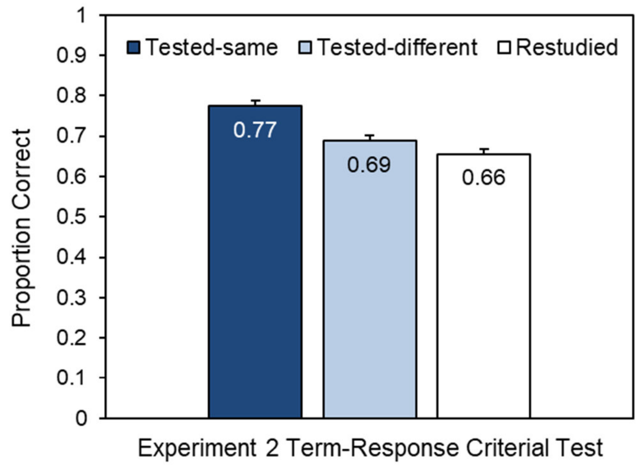


Figure 3.2. Criterial test performance on multiple-choice term-response and definition-response questions. Panel a: results from Experiment 1 of the current study. Panel b: results from McDaniel, Thomas, et al. (2013; Experiment 1). *Tested-same* indicates that the training question type matched the criterial test type; *tested-different* indicates that the training and criterial test question types did not match. Error bars in the Panel a are standard errors based on the interaction error term of a mixed-factors ANOVA on subject mean accuracy scores (based on Loftus & Masson, 1994).

a.



b.

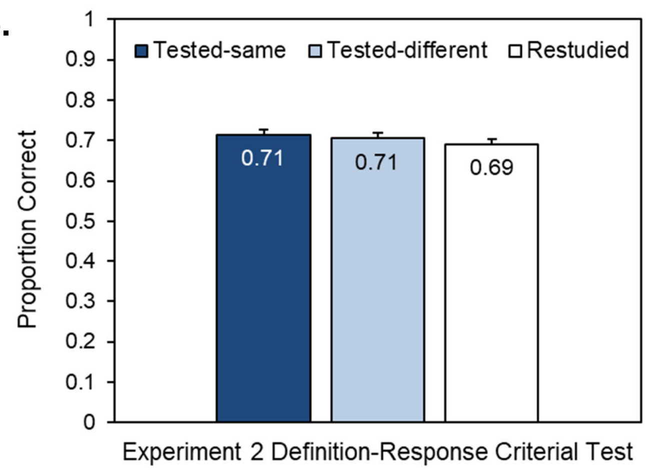


Figure 3.3. Criterial test performance on multiple-choice term-response and definition-response questions in Experiment 2. Panel a term-response. Panel b: definition-response. *Tested-same* indicates that the training question type matched the criterial test type; *tested-different* indicates that the training question type did not match the criterial test type. Error bars are standard errors based on the interaction error term of a mixed-factors ANOVA on subject mean accuracy scores (based on Loftus & Masson, 1994).

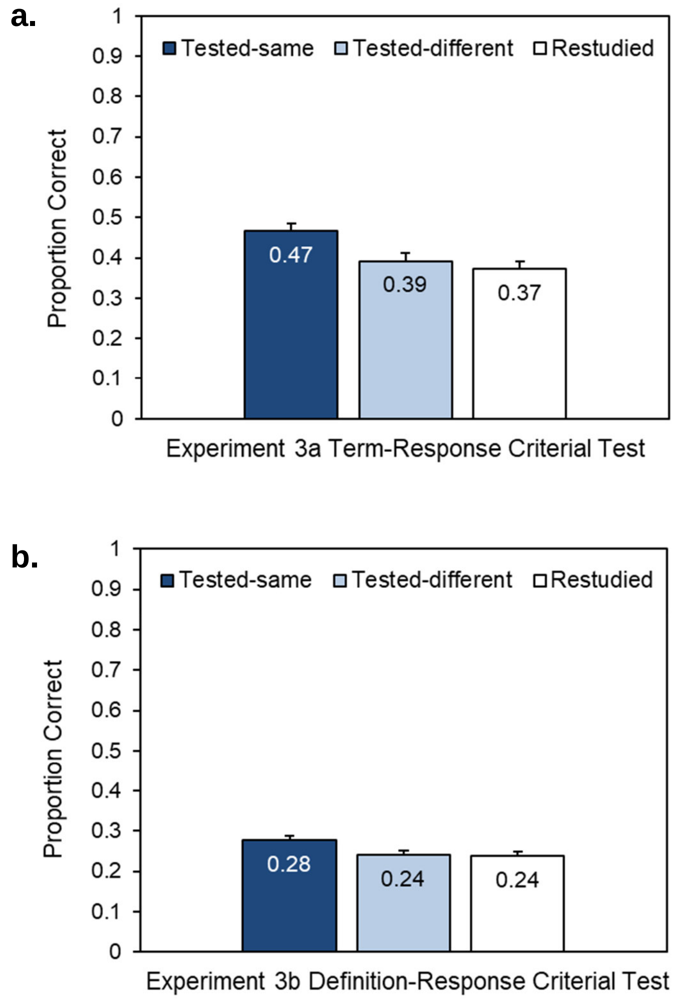


Figure 3.4. Criterial test performance on short answer term-response and definition-response questions in Experiments 3a and 3b. Panel a: term-response. Panel b: definition-response. *Tested-same* indicates that the training question type matched the criterial test type; *tested-different* indicates that the training question type did not match the criterial test type. Error bars are standard errors based on the error term of a within-subjects ANOVA on subject mean accuracy scores, performed separately for the two experiments (based on Loftus & Masson, 1994).

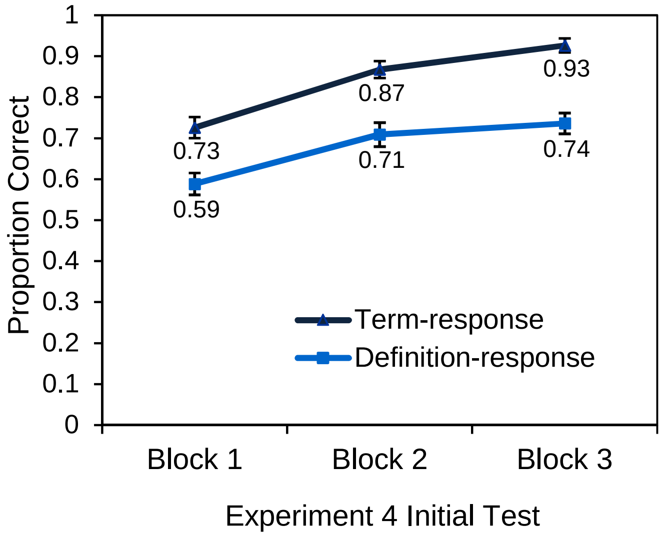


Figure 3.5. Initial test performance for multiple-choice term-response and definition-response questions across the three training blocks of Experiment 4. Error bars are standard errors of the means.

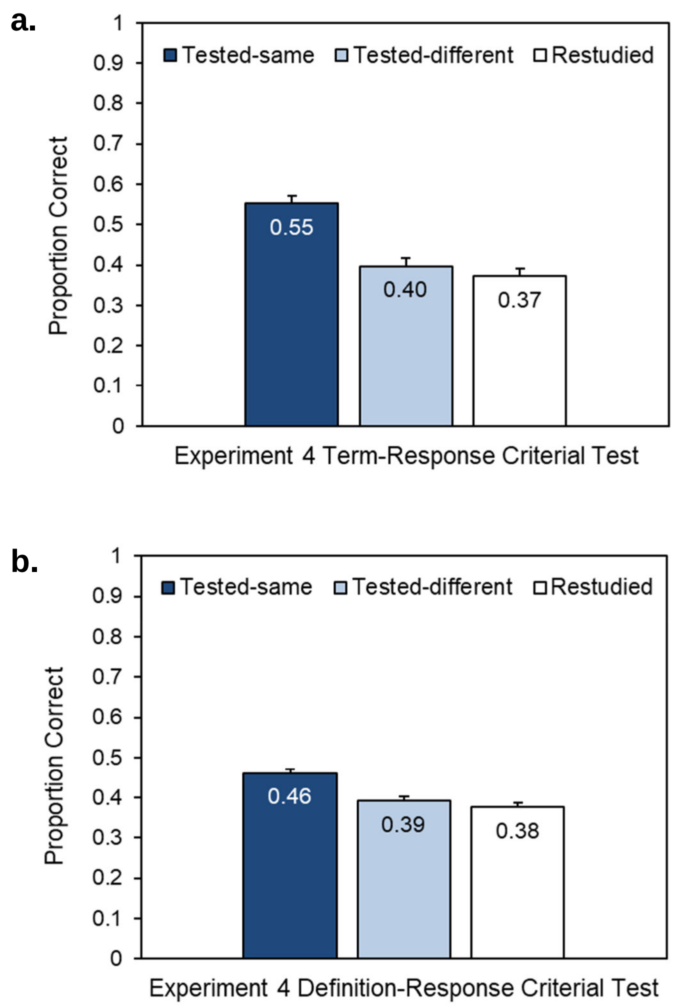


Figure 3.6. Criterial test performance on short answer term-response and definition-response questions in Experiment 4, in which each fact was trained three times. Panel a: term-response. Panel b: definition-response. *Tested-same* indicates that the training question type matched the criterial test type; *tested-different* indicates that the training question type did not match the criterial test type. Error bars are standard errors based on the interaction error term of a mixed-factors ANOVA on subject mean accuracy scores (based on Loftus & Masson, 1994).

Appendix 3A

Term-Definition Fact Terms

Fact number	Term
1	Absolute refractory period
2	Acetylcholine
3	Adaptation
4	Cingulate cortex
5	Cocktail party effect
6	Consciousness
7	Dermatome
8	Dermis
9	Dualism
10	ED-50
11	Encoding
12	Focal seizure
13	Glucose
14	Immunocytochemistry
15	Inattentional blindness
16	Infradian cycle
17	Labeled lines
18	Meninges
19	Neuroleptics
20	Neuron doctrine
21	Nissl stain
22	Oculomotor apraxia
23	Oligodendrocytes
24	Osmosensory neurons
25	Oxytocin
26	Paracrine
27	Plasticity
28	Reductionism
29	Sexual selection
30	Somatic intervention
31	Substantia nigra
32	Supplemental motor area
33	Sylvian
34	Temporoparietal junction
35	Tolerance
36	Wernicke's aphasia

Appendix 3B

Term-Definition Fact and Test Question Examples (used in Experiments 1-4)

Term	Stimulus type	Example
Dermis	Fact	Dermis is the layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood vessels.
	Term-Response (MC training or criterial tests, Exps 1-4)	What is the layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood vessels? a. epidermis b. dermis c. hypodermis d. epithelium
	Definition-Response (MC training or criterial tests, Exps 1-4)	What is the definition of dermis? a. The layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood. b. The layer of skin that contains no blood vessels and must receive nourishment from underlying tissue. c. The deepest layer of skin on the body. d. The protective layer of chitin that covers vulnerable body parts.
	Term-Response (SA criterial test, Exps 3a, 4)	The layer of skin that contains a rich web of nerve fibers in a network of connective tissue and blood vessels is called what?
	Definition-Response (SA criterial test, Exps 3b, 4)	What is the definition of dermis?
ED-50	Fact	ED-50 is the dose at which a drug has a half-maximal response.
	Term-Response (MC training or criterial tests, Exps 1-4)	What is the dose at which a drug has a half-maximal response? a. saturated a. therapeutic c. ED-50 d. LD-50
	Definition-Response (MC training or criterial tests, Exps 1-4)	What is the definition of ED-50? a. The dose at which a drug has a half-maximal response. b. The dose at which a drug is halfway metabolized by the body. c. The dose at which a drug results in a 50 percent mortality rate. a. The maximum dose at which a drug can be safely administered.
	Term-Response (SA criterial test, Exp 3a, 4)	The dose at which a drug has a half-maximal response is called what?
	Definition-Response (SA criterial test, Exp 3b, 4)	What is the definition of ED-50?

Note: Exp = Experiment, MC = multiple-choice, SA = short answer.

Appendix 3C

Paraphrased Term-Definition Practice Question Examples (used in Experiment 4)

Term	Stimulus type	Example
Neuroleptic	Term-Response (version 1)	What type of drug is effective at reducing the symptoms of schizophrenia? b. anxiolytic c. tricyclic c. neuroleptic d. sedative
	Term-Response (version 2)	A drug that is capable of reducing a patient's schizophrenic symptoms is called...? a. analgesic b. neuroleptic c. amphipathic d. neurolytic
	Term-Response (version 3)	A drug that can reduce schizophrenic symptoms effectively is called..? a. neuroleptic b. inhalant c. prophylactic d. inhibitor
Consciousness	Definition-Response (version 1)	What is consciousness? a. personal awareness of one's own emotions, thoughts, movements, and experiences. b. the mental state comprised of the id, ego, and superego. c. the ability to recognize oneself as a separate and distinct entity from other individuals in the environment. d. the ability to observe and empathize with one's actions and those of others.
	Definition-Response (version 2)	What does consciousness mean? a. components of the mind such as the id and the ego. b. the awareness of one's distinct existence apart from other people in an environment. c. self-awareness of personal emotional, physical, and mental experiences. d. the capacity for empathy when observing the actions of oneself and others.
	Definition-Response (version 3)	What does consciousness refer to? a. the perception of self as an entity which is separate from other people. b. id, ego, and superego which comprise the mental state. c. mindfulness of internal and external phenomena. d. mindfulness of one's own emotions, thoughts, movements, and experiences.

Note: Exp = Experiment.

References

Agarwal, P. K. (2011). Examining the relationship between fact learning and higher order learning via retrieval practice. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3468823)

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637. doi:http://dx.doi.org/10.1037/0033-2909.128.4.612

Breedlove, S. M., & Watson, N. V. (2013). *Biological Psychology: An Introduction to Behavioral, Cognitive, and Clinical Neuroscience* (7th ed.). Sunderland, MA: Sinauer Associates.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning* (pp. 23-45). Cambridge, MA: Belknap Press.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118-1133. doi:http://dx.doi.org/10.1037/a0019902

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. doi:http://dx.doi.org/10.1177/0963721412452728

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:http://dx.doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

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

Cheng, C. K. (2014). Effect of multiple-choice testing on memory retention—cue-target symmetry. (Doctoral dissertation). Retrieved from https://tspace.library.utoronto.ca/handle/1807/65649

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:http://dx.doi.org/10.1177/1529100612453266

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 801-812. doi:http://dx.doi.org/10.1037/a0023219

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning* (pp. 24-37). San Diego, CA: Academic Press.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304. doi:http://dx.doi.org/10.1080/09658211.2011.560121

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*(2), 151-164. doi:http://dx.doi.org/10.1016/j.jml.2013.03.002

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test ... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307-329. doi:http://dx.doi.org/10.1007/s10648-013-9248-9

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621-629. doi:http://dx.doi.org/10.1037/a0015183

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528-558. doi:http://dx.doi.org/10.1080/09541440601056620

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219-224.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology, 67*(2), 259-266. doi:http://dx.doi.org/10.1037/h0076933

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition, 3*(3), 171-176.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490. doi:http://dx.doi.org/10.3758/BF03210951

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:http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied, 21*(4), 370-382. doi:http://dx.doi.org/10.1037/xap0000063

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*(2), 192-201.

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:http://dx.doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360-372. doi:http://dx.doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*(1), 18-26. doi:http://dx.doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1), 3-21. doi:http://dx.doi.org/10.1037/xap0000004

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37*(8), 1077-1087. doi:http://dx.doi.org/10.3758/MC.37.8.1077

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging, 28*(1), 142-147. doi:http://dx.doi.org/10.1037/a0030890

Microsoft Corporation (2011). *Microsoft Fuzzy Lookup Add-In for Excel.* Retrieved from: https://atidan.files.wordpress.com/2013/08/fuzzy-lookup-add-in-for-excel.pdf

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

Pan, S. C., Gopal, A., and Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology 107*(4). doi:http://dx.doi.org/ 10.1037/edu0000074

Pan, S. C., Pashler, H., Potter, Z. E., and Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language 83*, 53-61. doi:10.1016/j.jml.2015.04.001

Pan, S. C., Rubin, B. R., and Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4). doi:http://dx.doi.org/ 10.1037/xap0000062

Pan, S. C., and Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin 144*(7), 710-756. doi: http://dx.doi.org/10.3758/s13423-017-1298-4

Pan, S. C., Rubin, B. R., and Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4). doi: 10.1037/xap0000062

Pan, S. C., Wong, C., Potter, Z., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*. doi:http://dx.doi.org/10.3758/s13421-015-0547-x

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:http://dx.doi.org/10.1037/0278-7393.31.1.3

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283-302.

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, *25*(4), 523-548. doi:http://dx.doi.org/10.1007/s10648-013-9240-4

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1281-1295. doi: http://dx.doi.org/10.1037/0278-7393.22.5.1281

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233-239. doi:http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14-26. doi:http://dx.doi.org/10.1016/j.jml.2014.04.004

Walker, D., Bajic, D., Mickes, L., Kwak, J., & Rickard, T. C. (2014). Specificity of children's arithmetic learning. *Journal of Experimental Child Psychology, 122*, 62-74. doi:http://dx.doi.org/10.1016/j.jecp.2013.11.018

Willingham, D. T. (2009). Why don't students like school? Because the mind is not designed for thinking. *American Educator*. Available at: http://www.aft.org/sites/default/files/periodicals/WILLINGHAM%282%29.pdf

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*(3), 214-221. doi:http://dx.doi.org/10.1016/j.jarmac.2014.07.001

CHAPTER 4

In Search of Transfer Following Cued Recall Practice:

The Case of Biology Concepts

Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., and Rickard, T. C. (2018).

Abstract

Previous work has demonstrated that cued recall of a term from a fact yields learning that does not transfer, relative to a restudy control, to recall of another term from the same fact. Here we report six experiments in which a series of manipulations during the initial study and training phases of learning, intended to increase transfer for process-based biology concepts, were investigated. In Experiments 1-2, fill-in-the-blank questions combined with immediate or delayed and repeated correct answer feedback improved learning but not transfer. In Experiments 3-4, practice questions which involved recalling process steps, understanding ordinal relationships, or making inferences, did not improve transfer. Positive transfer was produced, however, in Experiments 5-6 via *retrieval-verification-scoring*, a method in which difficult fill-in-the-blank questions were combined with extensive feedback processing. We discuss implications for transfer in both theoretical and applied contexts.

Introduction

Across virtually all areas of STEM education, the learning and retention of process-based concepts is an essential step for achieving mastery. For current purposes, a process-based concept is defined as a sequence of events that occur over time, leading to a predictable end state. Such concepts are ubiquitous in the target domain of this manuscript, namely introductory biology. A classic example is the concept of *protein synthesis*. It involves the following process: *DNA is first copied into RNA via transcription, and then RNA is coded into protein via translation* (Freeman, Quillin, & Allison, 2014). In that concept, two sequential processes, transcription and translation, result in the formation of protein.

Given the ubiquity and foundational nature of such materials, it is important to understand how students can learn them efficiently (i.e., with the least time invested), durably (i.e., yielding long-term retention), and comprehensively. One promising approach is *retrieval practice with feedback*. Retrieving information from memory, as occurs on a practice test, yields robust learning benefits over restudy and other non-retrieval methods. That finding, commonly known as the *retrieval practice effect* or *testing effect*, has been observed across a wide range of educationally-relevant materials (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), using different test formats (e.g., Kang, McDermott, & Roediger, 2007), with diverse learners (e.g., Pan, Pashler, Potter, & Rickard, 2015), and under different levels of motivation (e.g., Kang & Pashler, 2014). The effect is enhanced if *correct answer feedback* is provided after attempting retrieval, and it persists over time; the relative advantage of retrieval practice over restudy increases with longer retention intervals (e.g., Carpenter, Pashler, & Cepeda, 2009; Rowland, 2014). Given these findings, retrieval practice would appear to be well-suited to enhance the learning and transfer of process-based concepts – a possibility that we explore in this manuscript.

Specificity of Learning Through Cued Recall Practice

In the research literature (see Rickard & Pan, 2017), and quite likely also in current educational practice, the most commonly used variant of retrieval practice is *cued recall*, in which part of a studied item is presented on the initial test as a cue, and the remainder of the item is to be retrieved from memory. For the concept given above, an example cued recall question (in fill-in-the-blank format) is: "*In protein synthesis, DNA is first copied into RNA via _______?*" There is overwhelming evidence that cued recall facilitates, relative to restudy, subsequent retrieval of the same term that was retrieved on the initial test (i.e., the test that occurred during training). However, there is equally strong evidence that for educationally-relevant materials (e.g., fact learning), that facilitated learning does not extend, or *transfer*, to questions from the same stimulus that require a different term response. For example, in Pan, Gopal, and Rickard (2016), cued recall involving the fill-in-the-blank question, "*Thomas Jefferson purchased ______ from the French,*" for which the answer is *Louisiana*, yielded substantial learning relative to restudy when that same question was again asked after a delay, but did not yield transfer to "*Thomas Jefferson purchased Louisiana from WHOM?,*" relative to restudy. Rather, criterial test performance on such items was equivalent to performance after restudy. Similar findings have been reported by Hinze and Wiley (2011) and Pan and Rickard (2017); one exception is

McDaniel, Anderson, Derbish, and Morrissette (2007), which is discussed later in this manuscript. Thus, although retrieval practice appears to be at least as effective as restudy in all instances, it is, in work to date, evidently superior to restudy only for the initially tested term (for a review, see Pan & Rickard, 2018).

It however remains to be determined whether the extreme specificity of learning through cued recall testing is universal for facts and concepts, or whether there are conditions, heretofore unexplored, in which positive transfer will be observed. For instance, in many retrieval practice studies, target materials are studied relatively briefly and without supporting information (e.g., explanatory diagrams); it has yet to be established whether the same patterns would manifest for more enriched and more extensively studied information. Further, researchers have hypothesized that some forms of retrieval practice, including the use of questions that invoke processing of conceptual relationships, yields more transferrable learning (e.g., Jensen, McDaniel, Woodward, & Kummer, 2014). There has also been some research and theorizing which suggests that feedback that is processed for greater lengths of time and/or contains more than just the correct answer can enhance transfer (e.g., Butler, Godbole, & Marsh, 2013; see also Eglington & Kang, 2017; McDaniel & Little, in press). If feedback is the critical factor, then that would potentially constitute an *indirect* transfer effect (i.e., transfer that does not stem wholly from the retrieval event itself; for discussion see Roediger & Karpicke, 2006; Pan & Rickard, 2018). Learning benefits that stem solely from performing retrieval (e.g., as can occur for correctly answered cued recall without feedback) constitute *direct effects* of retrieval practice, whereas learning benefits stemming from activities besides the retrieval event, such as post-retrieval processing of feedback, constitute indirect effects.

Additionally, in the studies of cued recall and transfer noted above, the target materials were typically not process-based. For instance, none of the facts in Pan, Gopal, et al. (2015), which included materials drawn from AP History and Biology courses, described a sequence of processes. Rather, for history they focused on details of the "who, what, when, and where" variety, and for biology they typically took the form of *x of y and z* (e.g., "*The exoskeletons of most insects are made of chitin.*"). By contrast, the current experiments explicitly involved only process-based biology concepts. Thorough learning of such concepts might yield a richer, or deeper level of knowledge representation than does learning of non-process-based concepts or facts, which in turn might support positive transfer of cued recall relative to restudy.

Overall, across the six experiments detailed here, we explored three distinct types of cued recall with multiple forms of feedback that might promote transfer (see Table 4.1 for a summary of training methods). Although this series of experiments is systematic and is largely motivated by prior empirical or theoretical work, it is also exploratory, especially in the latter experiments.

The Current Study

In each experiment, subjects first studied 24 or 36 process-based concepts one at a time in random order. The concepts were drawn from a widely-used undergraduate biology textbook (Freeman et al., 2014; see Appendix 4A for a list). To enhance conceptual understanding and to promote more comprehensive learning, each concept was presented with a visual-conceptual diagram and a glossary of term definitions (see Figure

4.1 for an example). Subjects were instructed to study each concept carefully and to do so at their own pace. The decision to provide diagrams and definitions was motivated by a preliminary experiment in which those materials were not provided during initial study and in which no transfer was observed.

Training methods. After the initial study period, subjects completed a training phase in which each concept was trained with cued recall or restudy. Across six experiments, three retrieval practice methods were used (see Appendix 4B for examples): *term retrieval* (Experiments 1-2; fill-in-the-blank questions in which a missing term was to be retrieved), *higher order questions* (Experiments 3-4; primarily short answer questions in which a process step needed to be retrieved, the relationship between terms needed to be identified, or inferences about the concept had to be made), and *retrieval-verification-scoring* (Experiments 5-6; a new method in which the entire concept, excepting one or two cue terms, had to be retrieved, followed by a series of feedback processing methods).

Feedback methods. We implemented feedback in six different ways. This included three feedback types (see Appendix 4B for examples): *simple correct answer feedback* (Experiment 1) which involved just the term that was supposed to be retrieved (e.g., *transcription*); *process timelines* (Experiment 2) which involved a timeline of the underlying conceptual process (e.g., *DNA* — *transcription* — *RNA* — *translation* — *protein*); and *whole concept feedback* (all except Experiment 1) which involved presentation of the entire concept in text form.[1] Feedback was presented and processed using three different methods: *immediate feedback* (all except Experiment 2); *delayed and repeated feedback* (Experiment 2); and feedback in which subjects checked terms, scored their performance on those terms, and then copied the entire presented concept (i.e., *retrieval-verification-scoring*; Experiments 5-6). Those which involved more than immediate correct answer feedback can be classified as *elaborative* forms of feedback (Butler, Godbole, & Marsh, 2013). Any of these methods might enhance indirect effects of retrieval practice.

Criterial test. Two days after training, subjects completed a self-paced short answer criterial test. This criterial test, modelled after that used in Pan, Gopal, et al. (2016) and Pan, Wong, Potter, Mejia, and Rickard (2016), was identical across all experiments and featured three contiguous blocks in which each concept was tested once per block and with a different to-be-retrieved term per block (e.g., for *protein synthesis*, *transcription, translation*, and *RNA* were assessed). The instructions directed subjects to recall the exact terms that they had previously learned for each concept. No feedback was provided.

Based on prior literature, it was expected that there would be a substantial benefit of cued recall on the criterial test for terms that were previously tested (*tested-same* condition), relative to terms from concepts that were restudied (*restudied* condition). The primary question of interest was whether there would be a benefit of cued recall for previously untested terms from tested concepts (*tested-different* condition) versus the restudied condition, and under what training conditions such a benefit would manifest.

Experiment 1

The first experiment served to assess whether the cued recall and feedback methods used in prior studies of cued recall and transfer across terms from facts could yield transfer for process-based biology concepts that were supplemented by diagrams and definitions during initial study. It was the first of two experiments in which *term retrieval* questions were used during training. *Simple correct answer feedback* was also implemented in this experiment only.

Method

Subjects. In all experiments, subjects were recruited from the same subject pool and identically compensated. In Experiment 1, fifty-eight undergraduate students recruited from a large research university participated for course credit. Data from one subject was excluded due to computer error, leaving 57 subjects' data for analysis. That sample size well exceeded our per-experiment target (a sample size of 45 is needed to achieve 0.85 power to detect a mean proportion correct difference between the tested-same or tested-different condition and the restudied condition of 0.04 or greater on a one-tailed, one-sample *t*-test at $\alpha = .05$, based on the a priori power analysis reported in Pan, Gopal, et al., 2016).

Design and procedure. In this and all subsequent experiments, the independent variable, namely training via cued recall testing versus restudy, was manipulated within-subjects. There were two experimental sessions. In the first session, subjects were first informed that their task was to learn a series of biology concepts. An initial study period followed in which 36 concepts were studied one at a time as previously described. This was followed by a training phase in which each concept was tested or restudied once for 10 s each. Testing and restudy trials were randomly ordered within a single contiguous training block. Testing involved a single fill-in-the-blank test question with immediate correct answer feedback (presented for 8 and 2 s, respectively) and restudy involved viewing the concept in sentence form. After a two-day delay, subjects returned for the criterial test as previously described.

Materials. The materials consisted of 36 process-based biology concepts drawn from Freeman et al. (2014). Each concept was a biological process involving at least two discrete steps and took the form of a single sentence with a mean length of 18 words. The mean Flesch-Kincaid reading grade level of the concepts was 12.8. Three essential terms that represented crucial components or steps were identified for each concept (e.g., *transcription, translation*, and *RNA*). There was largely minimal overlap in content and no terms shared between concepts. Three fill-in-the-blank training questions and three corresponding short answer criterial test questions, each targeting one of those terms, were created for each concept. The criterial test questions were identical to the training questions except for the use of the word "*WHAT*" in place of the answer blank.

Visual-conceptual diagrams and term definitions. During initial study, each concept was presented in sentence form and accompanied by a visual-conceptual diagram and a glossary of term definitions (see Figure 4.1). Both appeared only during that portion of the experiment and were intended to aid subjects' understanding of the concepts, many of which included unfamiliar jargon (e.g., terms such as *cochlea*, *effector*, *phospholipid*). The diagram consisted of a color drawing of conceptual processes with arrows specifying individual steps or movements. Labelled representations of each of the three essential terms, plus any other jargon terms or important steps, were included with each diagram. The glossary defined each of the three essential terms for each concept plus any other jargon terms that were also present. Each term on the diagram and in the glossary was present in exact matching form in the concept sentence.

Training and criterial test lists. Each subject trained using one of six counterbalanced training lists. There was one training trial per concept on each list. Half of the concepts were tested using a fill-in-the-blank question and the remainder was restudied. Assignment of concept to cued recall testing or restudy, and the question that was used for each concept (i.e., the term needed to be retrieved), was counterbalanced using a Latin square. Six additional test lists, each corresponding to one of the training lists, were used for the criterial test. Each criterial test list contained three blocks of 36 short answer questions each, with each concept assessed once per block and on one of the three terms per block. There were 6 tested-same questions, 12 tested-different questions, and 18 restudied questions per block on the criterial test.

Data coding and analysis. Training and criterial test data were computer scored. Typed responses that exactly matched the correct answer (ignoring capitalization) were scored as correct. In this and all subsequent experiments, we performed two planned orthogonal contrasts that were motivated by prior findings of no transfer: (a) an Analysis of Variance (ANOVA) with factors of Tested-same vs. Not Tested (i.e., the tested-same condition vs. the tested-different and restudied conditions combined) and Block; and (b) an ANOVA limited to data from the tested-different and restudied conditions only, with factors of Tested-different vs. Restudied, and Block. This latter contrast directly tested for transfer relative to restudy. We used a significance criterion of $\alpha = 0.05$ for all statistical analyses.

Results

Initial study and training. Subjects spent an average of 19 min studying the concepts, or about 31 s per concept. During training, they typically retrieved the correct answer to less than one-third of the questions (M = 0.27) prior to viewing feedback (Table 4.2). This relatively low rate of retrieval success (as compared to typical rates in the literature; catalogued in Pan & Rickard, 2018) underscores the difficulty of learning process-based biology concepts despite the presence of diagrams and definitions during initial study.

Criterial test. Results are listed in Table 4.3 and depicted in Figure 4.2, panel a. An important initial observation is that criterial test performance across all blocks in the tested-same condition was substantially higher than that during training (M = 0.50 vs. M = 0.27), confirming the expectation that training yielded substantial learning in that condition. In the first ANOVA contrast with factors as previously described (Table 4.4), there was a main effect of Tested-same vs Not Tested. This is consistent with a retrieval practice effect for previously tested items as is evident in Figure 4.2. There was also a main effect of Block and no interaction. This is consistent with a performance improvement across blocks, an expected pattern given that the answers to criterial test questions in blocks 2 and 3 were viewable in preceding blocks. In the second contrast, there was no main effect of Tested-different vs. Restudied, which is consistent with there being no benefit of cued recall for untested terms relative to restudy. In other words, there was no evidence of positive transfer.

Effect of prior course experience. In this and subsequent experiments, exit surveys revealed that 53-77% of subjects had university-level biology or AP Biology course experience. Those with experience tended to perform better overall, but transfer patterns did not differ substantially between groups (similar to Pan, Gopal, et al., 2015; Pan & Rickard, 2017). Thus, we do not further discuss results as a function of course experience.

Discussion

In this experiment, although there was a benefit of cued recall for tested terms, there was no evidence of transfer of that learning to the tested-different condition. This result replicates the findings of Pan, Gopal, et al. (2016) using non-process-based materials supplemented by diagrams and definitions and with similar training methods. Further, in two unpublished follow-up experiments not detailed here, we observed the same results when a different term had to be retrieved on each of two training blocks, similar to Pan, Wong, et al. (2016; Experiment 2), and when two terms had to be retrieved per trial, similar to Hinze and Wiley (2011; Experiments 1 and 2) but with feedback. Thus, the benefits of cued recall combined with simple correct answer feedback appears to be specific to tested terms for the case of process-based concepts in the context of diagram- and definition-augmented study, extending our prior results.

Experiment 2

In the literature there has been one successful demonstration of transfer from tested to untested biology terms, relative to restudy, using fill-in-the-blank cued recall tests. McDaniel, Anderson, et al. (2007) had students enrolled in a university-level brain and behavior course initially learn target materials via assigned readings and classroom lessons. The students then completed practice quiz questions online over three weeks and in preparation for a later unit test. Question-by-question feedback consisting of the attempted answer and the correct answer was provided only after each quiz was submitted.

That feedback could be viewed repeatedly until the date of the unit test, on which positive transfer was observed relative to restudy (albeit without strict controls for time-on-task during training in the testing and restudy conditions). None of those design features were present in our initial experiment. Thus, to bridge the differences between Experiment 1 and that prior work, and to further improve the level of learning that subjects achieved for each concept with the aim of observing transfer, in Experiment 2 we implemented multiple exposures to each concept during a standalone initial study session (designed to approximate lessons and readings that students may complete before using retrieval practice), had subjects train on those concepts with *delayed and repeated feedback*, and included *whole concept feedback* in which *process timelines* were shown. This design was intended to conceptually replicate McDaniel, Anderson, et al.'s experiment in a laboratory setting.[2]

Method

Subjects. Forty-seven subjects participated for course credit. Data from 4 students were excluded, 2 due to a computer error and 2 due to not following instructions, leaving 43 subjects that were included in the data analysis.

Design and procedure. Uniquely in this experiment, there were three experimental sessions. The first session was entirely devoted to initial study of the concepts. Subjects began that session by cycling through each concept repeatedly for 30 minutes. During that period the concepts and term definitions were presented in paragraph form (see Appendix 4C for an example) with supporting diagrams. Next, subjects cycled through each concept repeatedly for an additional 30 minutes. During this time each concept was presented only in sentence form.

Training methods. The second session occurred 48 hrs after the first. During this session, subjects trained on the concepts using fill-in-the-blank tests or restudy across separate blocks, presented in random order. Each training block lasted 12 min. During the test block, subjects answered one fill-in-the-blank question per tested concept. After all questions were answered, they viewed question-by-question feedback. That feedback consisted of the whole concept in sentence form, their attempted answer, and a process timeline. They were instructed to cycle through that feedback repeatedly at their own pace until time had elapsed.

The restudy block had a similar structure. Subjects first viewed each restudied concept once in sentence form. They next viewed those concepts again in sentence form and with process timelines, similar to the test condition. They were instructed to cycle through those concepts repeatedly at their own pace until time had elapsed. The second session ended after both training blocks were finished. Finally, after a second 48 hr delay, subjects returned for a criterial test that used the same design as that of the prior experiment.

Learning instructions. Uniquely in this experiment, the instructions at the outset of session two informed subjects that they would be training for an exam in session three. Moreover, at the outset of the test block, they were told that the practice questions resembled exam questions, but that any part of a tested concept could be assessed on the exam; thus, they should learn entire concepts. During the restudy block, subjects were told to study.

Materials. The materials consisted of 24 biology concepts drawn from the set used in the preceding experiment, a reduction due to the logistics of experiment scheduling, and with training and criterial lists altered to match. For initial study, a textbook paragraph-style version of each concept, using short sentences and with term definitions embedded within the paragraph, was created. The paragraphs had a mean length of 79 words and a Flesch-Kincaid reading grade level of 11.3. For feedback, a timeline of process steps was created for each concept as previously described. All other materials were identical to those in the prior experiment.

Results and Discussion

Initial study and training. During the first and second halves of the initial study period, respectively, subjects viewed each concept an average of 6 and 4 times (for 21 and 31 s each). During training, subjects cycled through delayed feedback an average of 8 times per tested concept (11 s per concept) and 7 times each per restudied concept (16 s per concept). Subjects' training accuracy was M = 0.62, a substantial improvement over Experiment 1 (Table 4.2). This is especially notable given that training took place 48 hrs after initial study, which allowed more time for forgetting to have taken place, and suggests that the standalone initial study session enabled subjects to attain a higher level of understanding of each concept.

Criterial test. Results are listed in Table 4.3 and depicted in Figure 4.2, panel b; ANOVA results are listed in Table 4.4. As is evident in the figure, criterial test performance was improved over the preceding experiment across all conditions. Moreover, for the tested-same condition, performance is near ceiling. These results likely reflect the more extensive initial study and training methods that were used. However, the now-familiar pattern, namely a retrieval practice effect for tested-same terms and minimal transfer to tested-different terms relative to the restudied condition, was still observed. Thus, in this experiment we were unable to reproduce the findings of McDaniel, Anderson et al. (2007) in a laboratory setting and despite using many of the same training methods (albeit without learning in the context of an actual class, or training on multiple sessions across several weeks – issues that we further address in the General Discussion). Overall, the results of both Experiments 1 and 2 reinforce our conclusion that cued recall in the form of simple fill-in-the blank questions tends to yield highly specific learning. Moreover, the results of Experiment 2 rule out the hypothesis that delayed and repeated feedback that includes more than just correct answers (i.e., to fill-in-the-blank questions) is always sufficient to yield indirect transfer to untested terms.

Experiment 3

The third experiment was the first of two experiments in which we investigated retrieval practice methods that require learners to consider connections between different components of a concept. In the literature, some researchers have suggested that *higher order* questions, including questions that require subjects to retrieve more than just an isolated portion of a fact or concept, yields better overall understanding and may improve certain types of transfer (e.g., Jensen et al., 2014; Karpicke & Aue, 2015). Accordingly, in this experiment each tested concept was trained four times using two question types: *process step* questions, which required retrieval of half of each concept from memory, and *order* questions, which required the determination of the ordinal relationship of two terms from a concept. The use of two question types also introduced variation (i.e., potential encoding variability) in the types of retrieval practice performed for each tested concept.

Method

Subjects. Fifty-four subjects participated for course credit. Data from 4 subjects was excluded due to computer errors and 7 subjects did not return for the second session, leaving 43 subjects' data for analysis.

Design and procedure. The primary difference in this experiment was the training phase design. Subjects completed four successive training blocks in which each concept was tested or restudied once per block for 20 s (and for test trials, 10 s for retrieval and 10 s for feedback). Test and restudy trials were randomly intermixed in each block. With this experiment we also reverted to having both initial study and training occur in an initial session.

Training methods. For process step questions, subjects were given the name of the concept and asked, "*What occurs in the first step? Answer completely with all details.*" (or that question but with *second* instead of *first*). For order questions, subjects were presented with the name of the concept, two terms in alphabetical order, and asked, *"Which of these occurs first in the process?"* (or that question but with *second* instead of *first*). Thus, order questions constituted a form of two-alternative forced choice question, marking the sole deviation from cued recall in the present experiments. Importantly, each tested concept was trained with one process step question and one order question, with each concept tested once per block and on one question type per block. Whole concept feedback was provided immediately after each retrieval attempt. On restudy trials, subjects viewed entire concepts in sentence form.

Learning instructions. To answer test questions, subjects were told to think of the entire concept, including its constituent processes; to recall exact words; and to check feedback word-by-word for complete comprehension. For restudy trials, subjects were told to study.

Materials. To facilitate process step questions, all 36 concepts were reworded into two separate sentences that identified the two main steps for each concept in sequence using the connective words "*first*" and "*second*" or "*as a result*" (e.g., *protein synthesis involves two steps. In the first, DNA is first copied into RNA via transcription. In the second, RNA is coded into protein via translation*.). The sentences had a combined mean length of 17 words and a Flesch-Kincaid reading grade level of 7.6, the reduction in grade level due to the added connective words. Each concept was presented in two-sentence form throughout the experiment.

Two process step and two order questions were created for each concept, one per step, with the assignment of question to concept counterbalanced across four training lists. The names and terms of several concepts were modified to avoid overlapping between cues and responses. Each order question required subjects to select between two terms, one from each step. For each subject, the correct answer to the process step and order questions for each tested concept referred to the same step (e.g., if retrieval of the first step was required for the former question, then identifying a term from the first step was required for the latter question). On the criterial test, which was identical at the trial level to that of the prior experiments, two terms from that step were assessed in the tested-same condition and one term from the other step (i.e., the lure in the order question) was assessed in the tested-different condition (cf. Little, Bjork, Bjork, & Angello, 2012). There were 12 tested-same questions, 6 tested-different questions, and 18 restudied questions per criterial test block.

Data coding. All process step questions were scored by raters blind to condition using a rubric of idea units for each concept (following the method detailed in Pan & Rickard, 2017). All other questions were computer scored in the same manner as in the preceding experiments.

Results and Discussion

Initial study and training. Subjects spent an average of 23 min, or 38 s per concept, during initial study. During training (Table 4.2), mean performance improved from the first to second administrations of each process step and order question (reaching $M = 0.54$ and 0.87, respectively).

Criterial test. Results are listed in Table 4.3 and depicted in Figure 4.2, panel c; ANOVA results are detailed in Table 4.4. The same overall pattern as in the prior experiments was again observed. Mean performance in the tested-different condition reached parity with the tested-same condition by block 3, but the interaction with Block was not significant. Moreover, across blocks 1 and 2, there was little indication of positive transfer, and the first block constitutes the purest measure of learning from the first session. Thus, although process step and order questions required a greater understanding of the relationship between different components of each tested concept than term retrieval questions, and also targeted each concept in different ways (albeit with some overlap in content), using those questions in conjunction with whole concept feedback still yielded specific learning as in the earlier experiments.

Experiment 4

In the fourth experiment, we continued our investigation of higher order questions by having subjects train on concepts using process step questions and *inference questions.* This latter question type required subjects to integrate information about each tested concept in a different way from that in which it was originally presented, which might promote greater conceptual understanding and improve transfer (Jensen et al., 2014).

Method

Subjects. Forty-nine subjects participated for course credit. Data from 5 subjects was excluded due to computer or experimenter error and 1 subject did not return for the second session, leaving 43 subjects' data for analysis.

Design and procedure. The design and procedure were identical to that of the preceding experiment, except that inference questions were used in place of order questions. Directions to think of entire concepts, recall exact words, and completely check feedback were provided for test trials, just as in Experiment 3.

Materials. One of the authors with experience as a biology section instructor created an inference question for each of the 36 concepts. These questions replaced the order questions on the training lists. Depending on the concept, the inference question required integrating multiple pieces of information (e.g., *"what is the purpose of plasmogamy in fungal cells?"*), identifying a critically important component or process (e.g., "*the process of gene flow affects what aspect of a population?"*), or determining a hypothetical state of a conceptual process (e.g., *"if protein synthesis could be reversed, what molecule would you end up with?"*). The answers to these questions did not necessarily refer to any of the essential terms for the tested concept but could be derived from the provided whole concept feedback.

Data coding. All process step questions were scored using same procedures as in the prior experiment, with those procedures adapted for scoring inference questions.

Results and Discussion

Initial study and training. Subjects spent an average of 23 min, or 37 s per concept, during initial study. During training (Table 4.2), mean performance improved from the first to second administrations of each process step and inference question (reaching M = 0.57 and 0.70, respectively).

Criterial test. Results are listed in Table 4.3 and depicted in Figure 4.2, panel d; ANOVA results are detailed in Table 4.4. The same overall pattern as in the prior experiments was yet again observed. Moreover, mean performance in the tested-different condition was lower than the tested-same and restudied conditions in block 3. That apparent negative transfer was evident in the orthogonal contrasts (Table 4.4), but is an anomalous result in the context of all prior results and may thus reflect random variance. Overall, the results of Experiments 3 and 4 are inconsistent with the hypothesis that higher order questions of various types always yield transferrable learning, at least for the case of transfer to untested terms.

Experiment 5

Given the prior failures to observe positive transfer, we next investigated more extensive cued recall and feedback techniques. We hypothesized that subjects (a) had to be made aware of the importance of including multiple exact terms in their retrieval attempts through actual experience and not just through instructions; (b) their attention had to be specifically directed to each essential term in each tested concept; and (c) they had to focus on both while processing feedback. To achieve these goals, we drew inspiration from Rawson and Dunlosky's (2011) use of *retrieval-monitoring-feedback* trials in which subjects had to check each of several main ideas per retrieved definition; without such checking, subjects could erroneously regard incomplete answer attempts as fully correct. We ultimately developed a new training technique, *retrieval-verification-scoring*, which involved recalling nearly entire concepts and then checking terms, scoring terms, and viewing and copying whole concepts. This was a far more time-consuming procedure than in prior experiments, potentially boosting transfer at the cost of efficiency and complexity. Moreover, given the extensive use of feedback on each test trial, any observed transfer could be attributed to indirect effects of retrieval practice.

Uniquely in this experiment, rather than setting a total time limit across the test and restudy conditions (which given the more time-consuming nature of retrieval-verification-scoring trials under self-paced training would have yielded large disparities in the number of trial repetitions per concept), we equated the total trial repetitions per concept across training conditions (i.e., three repetitions each). Further, to equate the amount of word-for-word copying that occurred in both conditions, we implemented copying during both test and restudy trials.

Method

Subjects. Fifty-seven subjects participated. Data from 6 subjects was excluded due to computer or experimenter errors and 5 subjects did not complete the second session, leaving 46 subjects' data for analysis.

Design and procedure. Only the training phase substantially differed from the preceding experiments. Subjects completed three successive test blocks and three successive restudy blocks (with the starting order of test or restudy randomly assigned) and with each block having exactly one trial per concept with no time limit. Thus, subjects completed a total of three test or three restudy trials per concept.

Training methods. Each retrieval-verification-scoring trial involved the following steps (see Figure 4.3 for an example). First, subjects attempted to answer a difficult fill-in-the-blank question which required retrieval of nearly the entire concept except one essential term. After submitting their answer, the feedback screen appeared. On that screen, the answer attempt was displayed along with the three essential terms for that concept, one of which was the cue in the preceding question. Subjects had to verify whether they had correctly typed each of those terms by marking yes or no. Next, they counted how many terms out of 3 that they had correctly recalled and typed that number as their score. Finally, the whole concept was presented for subjects to copy. By comparison, in the restudy condition, subjects simply viewed and copied whole concepts.

Learning instructions. In the test condition, subjects were encouraged to fully learn each concept and to improve their scores over multiple attempts. In the restudy condition, subjects were told to study.

Materials. The materials consisted of 24 biology concepts in one-sentence form as in the first two experiments. Three difficult fill-in-the-blank test questions were created for each concept. In each, all but one essential term and, in many cases, one article word (i.e., *a*, *an*, or *the*) that preceded that term, was replaced with a continuous blank line or several lines interspersed with punctuation marks. On the criterial test, the terms that were to be retrieved during training were assessed in the tested-same condition and the terms that were cues to training questions were assessed in the tested-different condition. There were 8 tested-same, 4 tested-different, and 12 restudied questions per criterial test block. All other materials were largely identical to those used in prior experiments.

Data coding. Difficult fill-in-the-blank questions were computer scored in terms of the number of essential terms correctly recalled per concept.

Results and Discussion

Initial study and training. Subjects spent an average of 14 min, or 34 s per concept, during initial study. During training, mean subject performance improved across each of three test trials per concept (Table 4.2), reaching M = 0.59 by the third iteration. As expected, the more time-consuming nature of retrieval-verification-scoring trials (subjects spent an average of 84 s each vs. 26 s for restudy) resulted in a large time-on-task difference; subjects spent an average of 47 min to complete three repetitions per tested concept, versus only 15 min for restudy.

Criterial test. Results are listed in Table 4.3 and depicted in Figure 4.2, panel e; ANOVA results are detailed in Table 4.4. As is evident in the figure, there was evidence of positive transfer relative to restudy. Performance was best overall in the tested-same condition, second best in the tested-different condition, and lowest in the restudied condition. Thus, the use of retrieval-verification-scoring trials – a method that directed subjects to attend to the essential terms of each concept, highlighted the importance of correctly recalling and verifying exact terms, and encouraged and provided the opportunity to closely examine entire concepts (albeit with much greater time required under self-paced training conditions) – elevated performance for both tested and untested terms above that of conventional restudy.

Experiment 6

Having demonstrated a method of cued recall with extensively processed feedback that successfully yielded transfer relative to restudy, for the final experiment we attempted to replicate that finding under conditions of strict control for time-on-task, fixed-paced training, and equal training phase item repetitions. We also included a modified restudy control that was potentially more competitive with testing.

Method

Subjects. Forty-eight subjects participated. Data from 5 subjects were excluded due to computer or experimenter errors, and 3 subjects were unable to fully comply with instructions, leaving 40 subjects' data for analysis.

Design, procedure, and materials. This experiment incorporated the design of its predecessor with the following changes: there was a fixed time limit of 70 s for both test and restudy trials (and for retrieval-verification-scoring trials, 30 s for testing and 40 s for feedback), and there were two repetitions of either testing or restudy trials for each concept across contiguous blocks (four training blocks in total). Thus, in this design, subjects that had not fully completed all steps of a retrieval-verification-scoring trial by the trial time limit were not permitted to do so and immediately began the next trial. Time limits were chosen based on prior experiment data and experiment timeslot logistics. Additionally, in the restudy condition, the three essential terms were shown alongside the whole concept for one-third of restudied concepts (with the selection of concepts for which terms were shown counterbalanced across training lists). Subjects were informed that those terms, where present, were additional information for them to study. This was intended to further control for the display of terms during retrieval-verification-scoring. Materials were otherwise identical to that of the prior experiment.

Results and Discussion

Initial study and training. Subjects spent an average of 14 min, or 36 s per concept, during initial study. Performance improved over the two test trials per tested concept (Table 4.2), reaching M = 0.41 by the second iteration.

Criterial test. Because criterial test performance in the restudied condition for concepts with essential terms displayed was not better than restudied concepts without terms displayed (M = 0.61 for both; data averaged across criterial test blocks), suggesting that the presence of those terms during some restudy trials was ineffective at improving performance, data from restudied concepts that did or did not have terms separately displayed during training were combined in the analyses. Results are listed in Table 4.3 and depicted in Figure 4.2, panel f; ANOVA results are detailed in Table 4.4. We again observed evidence of positive transfer relative to restudy, replicating the results of the preceding experiment. Thus, the results of the Experiments 5 and 6 indicate that cued recall combined with extensive feedback processing – at a level substantially beyond that used in earlier experiments – can yield transfer from tested to untested terms.

General Discussion

In the forgoing experiments we explored whether a series of cued recall-based training and feedback methods could yield transfer for process-based biology concepts. This research encompassed fill-in-the-blank questions, process step questions, order questions, and inference questions, none of which appeared to be sufficient alone to induce transfer. Additionally, enrichment of initial study using diagrams and definitions, extended initial study periods, immediate or delayed feedback, repeated feedback, or feedback which contained whole concepts or process timelines was insufficient to generate transfer. Rather, criterial test performance in the tested-different condition was in all of those cases (Experiments 1 through 4) nearly equivalent or numerically lower than that in the restudied condition, just as occurred in prior work on this topic (excepting paired associate materials; see Carpenter, Pashler, & Vul, 2006; Rickard & Pan, 2018). Those results underscore the nontrivial challenge of designing a task that produces transfer of cued recall learning on a process-based concept to previously untested terms from that concept. Ultimately, the use of substantially more extensive feedback methods (Experiments 5 and 6) which required subjects to check and verify different components of each concept, followed by copying that concept, was necessary to generate transfer.

When Indirect Effects of Testing May Yield Transfer

The absence of transfer in Experiments 2 through 4 is particularly notable given that the post-retrieval elaborated feedback methods used in those experiments resembled that used in prior studies in which indirect transfer of various types has been observed (e.g., Butler et al., 2013; McDaniel, Anderson et al., 2007; cf. McDaniel, Wildman, & Anderson, 2012). Those contrasting results invite further consideration of the conditions under which various feedback types are effective at yielding transfer.

The efficacy of different feedback types. Regarding the delayed, extended, and in some cases repeated feedback of McDaniel, Anderson, et al. and Experiment 2 of the present work, two general differences between that study and this experiment may account for the contrasting transfer results, although neither possibility can at present be clearly linked to a psychological mechanism. First, the authentic educational context and longer learning interval (i.e., up to three weeks instead of 48 hrs) in that study may have been pivotal. During that time, students likely took advantage of repeated and spaced feedback learning opportunities, plus studied relevant course materials. Post-retrieval study of materials (i.e., reading text passages or chapters) has also been employed in several other retrieval practice studies that observed positive transfer to application and inference questions (e.g., Karpicke & Blunt, 2011; Lechuga, Ortega-Tudela, & Gómez-Ariza, 2015; McDaniel, Bugg, Liu, & Brick, 2015; McDaniel, Howard, & Einstein, 2009). Moreover, the online quizzing and feedback platform used in McDaniel, Anderson, et al. has since been used to successfully generate transfer between factual and application questions (Thomas, Weywadt, Anderson, Martinez-Papponi, & McDaniel, 2017; cf. McDaniel, Wildman, et al., 2012, described further next). Second, transfer may be easier to achieve for the neuroscience facts used in that study than for the process-based concepts of the present work.

Further, transfer of different types (i.e., to application and inference questions and to untested materials seen during the initial study phase, respectively) than that pursued in the present experiments was the focus of Butler et al. (2013) and McDaniel, Wildman, et al. (2012), both of which used explanatory feedback (which was comparable to the whole concept feedback and process timelines used in Experiments 2 through 4). A recent meta-analytic review found that the effects of retrieval practice on transfer can vary substantially by transfer type and tend to be more robust for application and inference questions than to untested materials seen during initial study or across terms from a fact or other multi-element stimulus (Pan & Rickard, 2018). Additionally, explanatory feedback has not always yielded strong positive transfer relative to restudy in prior work (e.g., it was observed in Butler et al. but was not significant in McDaniel, Wildman, et al., although there were differences in the amount of explanatory content provided). Given the current results, it follows that indirect transfer is not an assured consequence of even deliberately implemented post-retrieval elaboration in the form of either explanatory feedback or many of the other aforementioned feedback types.

Retrieval-verification-scoring and transfer. When we did observe transfer, it required extensive post-retrieval study-like activities. For Experiment 5, we interpreted the observed transfer as reflecting indirect effects of retrieval practice, although that inference was tentative due to a potential time-on-task confound. That confound was eliminated in Experiment 6, and more complete (albeit smaller in magnitude) transfer was observed. Thus, we conclude that it is possible to achieve indirect transfer of test-based learning for process-based concepts, but that such transfer may require any or all of the post-retrieval activities used in the retrieval-verification-scoring procedure.

Retrieval-verification-scoring differed from the feedback presentation and processing methods used in prior experiments in four major ways. First, the verification step directed subjects' attention to each of the essential terms per concept, including the term that was later assessed in the tested-different condition on the criterial test. Second, that step provided subjects with firsthand knowledge of the importance of fully and exactly recalling key components of each concept. Third, the scoring step furnished subjects with a numerical measure of their mastery of each concept. This may have corrected inaccurate judgments of learning and motivated greater learning on subsequent trials. Finally, all of these feedback processing steps may have enhanced the efficacy of the final copying step. Overall, a prerequisite for transfer from tested to untested terms appears to be training procedures that deliberately focus learners' attention to those terms during feedback and not just instructional prompts, the provision of more explanatory details, or other types of elaborative information.

Limitations, Educational Implications, and Future Work

The reader might suspect that the lack of transfer in Experiments 1 through 4 reflects the fact that subjects did not undergo deep conceptual learning during either the initial study or training phases. We are open to that possibility, even in the case of extensive initial study in Experiment 2. If true, however, it does not necessarily constitute a weakness of this work from the applied perspective. We are probably not alone in our strong intuition that the onset of deep conceptual understanding in STEM fields lags the acquisition of facts and of more tenuous, and often piecewise, conceptual understanding.

Further, progress through those levels may be necessary for achieving deeper understanding (Anderson & Krathwohl, 2001). In that light, the current insights regarding transfer may be especially applicable to classroom-based cued recall tests with feedback, and for novice and intermediate students. It is also important to recognize that the new manipulations that improved transfer in Experiments 5 and 6 (i.e., retrieval-verification-scoring) are unlikely to have yielded a substantial leap in deep understanding relative to Experiments 2 through 4. Rather, that transfer likely reflects the greater efficacy relative to restudy of those manipulations in focusing subjects' attention on the entire stimulus presented on practice test trials compared to restudy. Future work that further explores the critical properties of activities that produce indirect transfer, and especially feedback methods such as the retrieval-verification-scoring method employed here, is warranted.

Our demonstration of retrieval-verification-scoring also represents a proof-of-concept for more complex (but potentially more effective, contingent on learning objective) forms of cued recall practice that may not only be useful for improving transfer across terms for process-based concepts and other stimuli (e.g., multi-term facts), but also for computer-based training programs more generally. Such methods could be refined (i.e., made more efficient) via the use of automated scoring procedures, adaptive algorithms that tailor training schedules to performance levels, or even gamification that incorporates the goal of score improvement over training trials. As retrieval practice research increasingly addresses issues of practical application, more sophisticated implementations of retrieval practice may emerge as a highly competitive alternative to conventional cued recall methods.

Acknowledgments

This chapter, in full, has been submitted for publication as Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., and Rickard, T. C., In search of transfer following cued recall practice: the case of biology concepts. The dissertation author was the primary investigator and author of this paper.

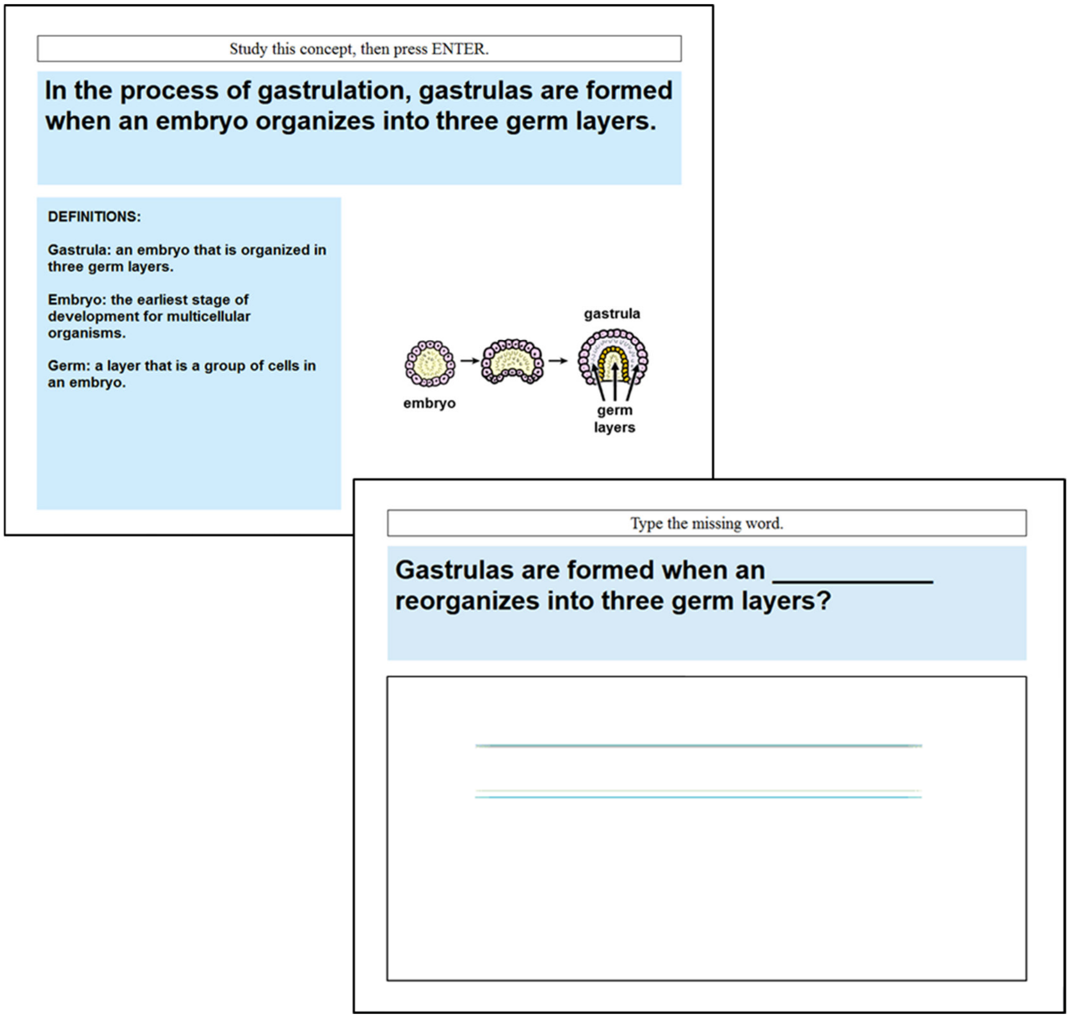


Figure 4.1. Example initial study and cued recall training test trials (top and bottom panels, respectively). During initial study, concepts are shown with diagrams and definitions. During the training phase, concepts are trained using cued recall (fill-in-the-blank example from Experiments 1-2 displayed) or restudied (not shown).

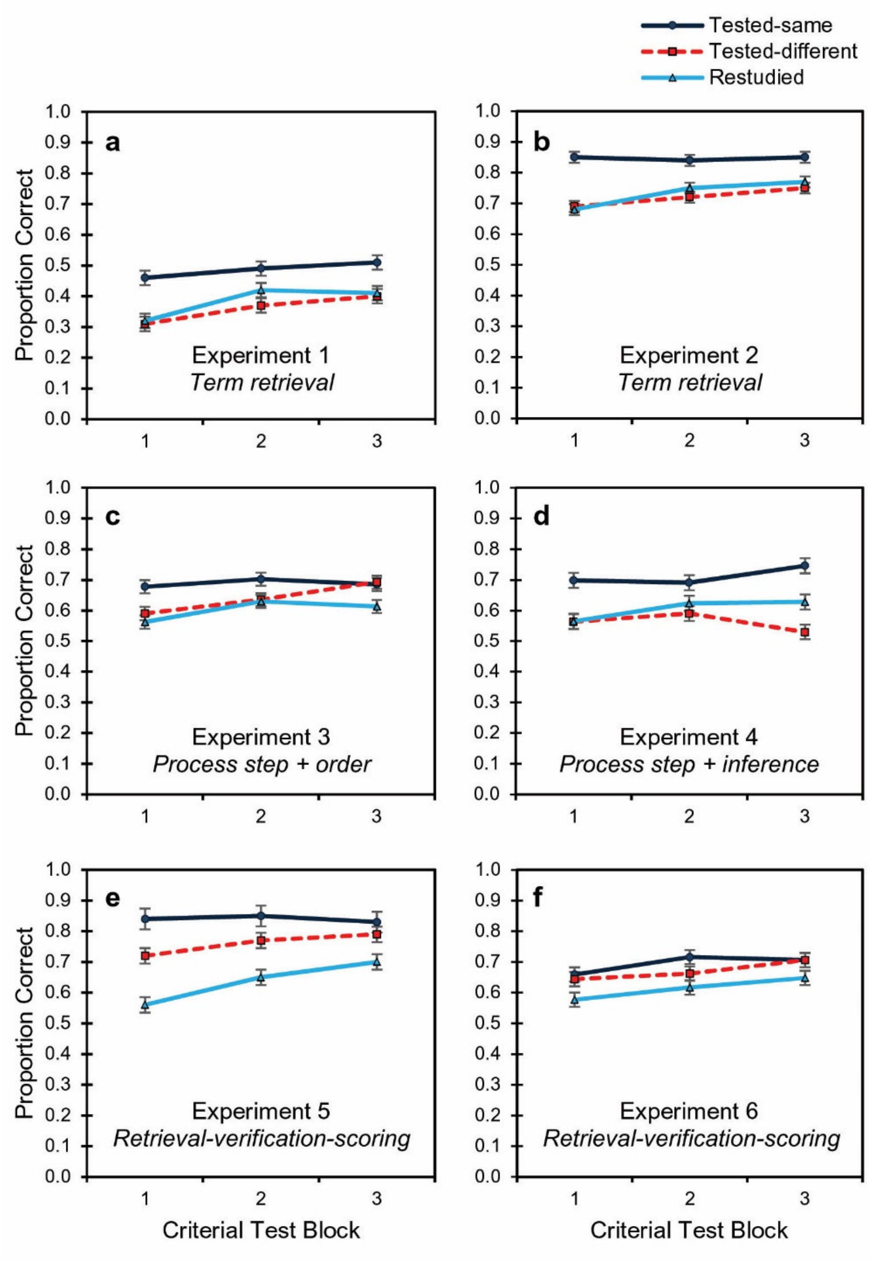


Figure 4.2. Criterial test results of Experiments 1-6. Error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject-level mean accuracy scores (based on Loftus & Masson, 1994).

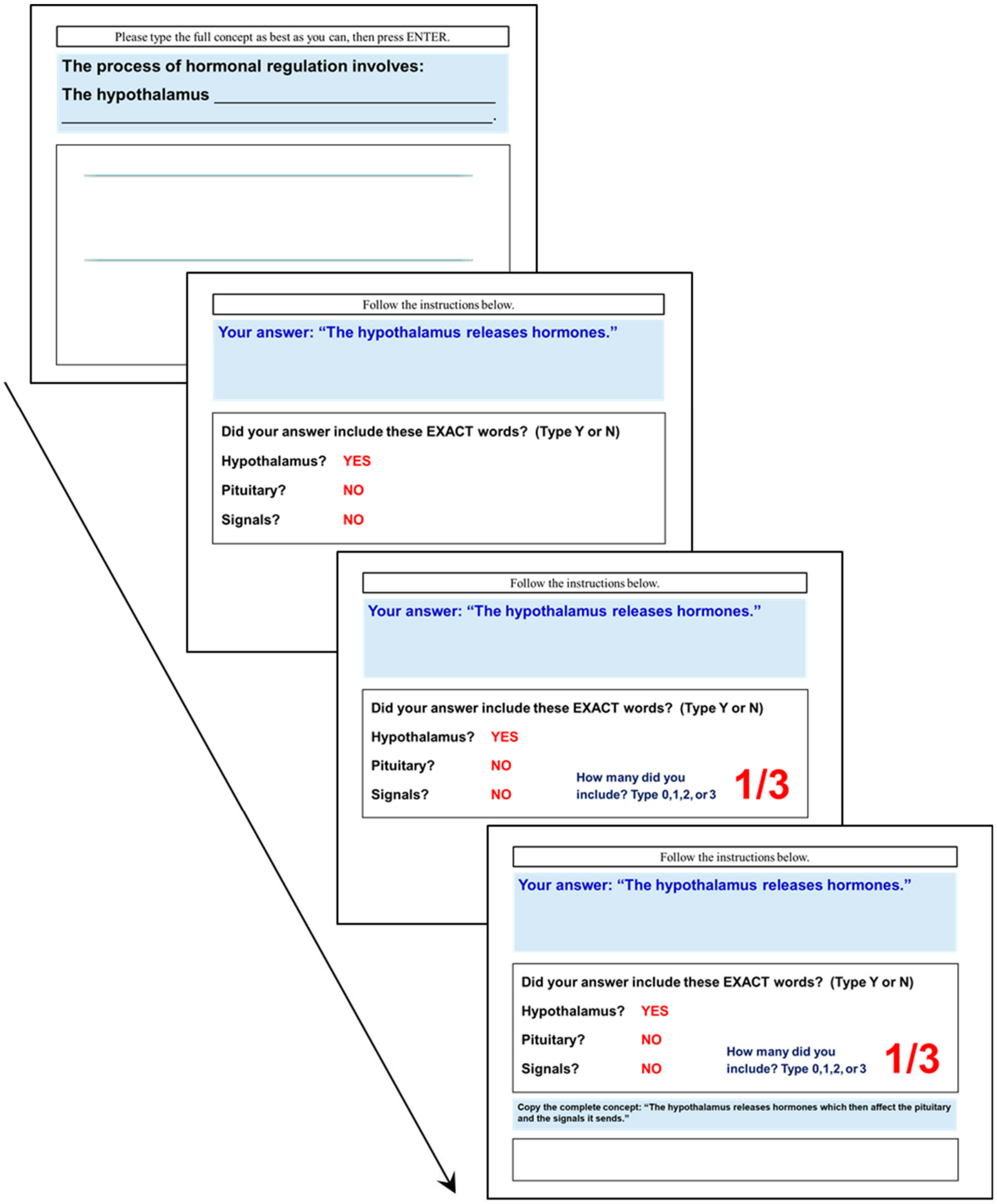


Figure 4.3. Example *retrieval-verification-scoring* trial used in Experiments 5-6. First panel: difficult fill-in-the-blank question. Second panel: subjects check their answer for essential terms. Third panel: numerical scoring. Fourth panel: whole concept presented for word-for-word copying.

Table 4.1. Summary of training methods used.

Exp.	Question type	Feedback condition	Restudy condition
Term retrieval			
1	Fill-in-the-blank	Correct answer only	Whole concept
2	Fill-in-the-blank	Whole concept, process timeline	Whole concept, process timeline
Higher order questions			
3	Process step + order	Whole concept	Whole concept
4	Process step + inference	Whole concept	Whole concept
Retrieval-verification-scoring			
5	Difficult fill-in-the-blank	Terms scored, whole concept copied	Whole concept copied
6	Difficult fill-in-the-blank	Terms scored, whole concept copied	Whole concept copied, with or without terms listed for study

Note: Exp. = Experiment.

Table 4.2. Training phase mean (SE) accuracy data.

			Initial test proportion correct		
	Exp.	Question type	Training block 1	Training block 2	Training block 3
Term retrieval					
	1	Fill-in-the-blank	0.27 (0.028)	–	–
	2	Fill-in-the-blank	0.62 (0.032)	–	–
Higher order questions					
	3	Process step	0.36 (0.021)	0.54 (0.028)	–
		Order	0.78 (0.031)	0.87 (0.027)	–
	4	Process step	0.38 (0.023)	0.57 (0.028)	–
		Inference	0.51 (0.031)	0.70 (0.033)	–
Retrieval-verification-scoring					
	5	Difficult fill-in-the-blank	0.33 (0.024)	0.51 (0.021)	0.59 (0.015)
	6	Difficult fill-in-the-blank	0.21 (0.056)	0.41 (0.065)	–

Note: Exp. = Experiment.

Table 4.3. Criterial test mean (SE) accuracy data.

	Exp.	Condition	Block 1	Block 2	Block 3
Term retrieval					
	1	Tested-same	0.46 (0.037)	0.49 (0.041)	0.51 (0.040)
		Tested-different	0.31 (0.027)	0.37 (0.032)	0.40 (0.030)
		Restudied	0.32 (0.027)	0.42 (0.030)	0.41 (0.028)
	2	Tested-same	0.85 (0.031)	0.84 (0.027)	0.85 (0.028)
		Tested-different	0.69 (0.036)	0.72 (0.036)	0.75 (0.035)
		Restudied	0.68 (0.032)	0.75 (0.028)	0.77 (0.027)
Higher order questions					
	3	Tested-same	0.68 (0.031)	0.70 (0.031)	0.69 (0.031)
		Tested-different	0.59 (0.036)	0.64 (0.041)	0.69 (0.036)
		Restudied	0.56 (0.031)	0.63 (0.031)	0.61 (0.029)
	4	Tested-same	0.70 (0.029)	0.69 (0.033)	0.75 (0.030)
		Tested-different	0.56 (0.039)	0.59 (0.037)	0.53 (0.037)
		Restudied	0.57 (0.026)	0.62 (0.025)	0.63 (0.026)
Retrieval-verification-scoring					
	5	Tested-same	0.84 (0.029)	0.85 (0.027)	0.83 (0.028)
		Tested-different	0.72 (0.038)	0.77 (0.038)	0.79 (0.039)
		Restudied	0.56 (0.036)	0.65 (0.033)	0.70 (0.033)
	6	Tested-same	0.66 (0.037)	0.72 (0.041)	0.71 (0.037)
		Tested-different	0.64 (0.049)	0.66 (0.044)	0.71 (0.040)
		Restudied	0.58 (0.036)	0.62 (0.031)	0.65 (0.038)

Note: Exp. = Experiment.

Table 4.4. Criterial test analysis of variance (ANOVA) results.

Exp.	ANOVA type	Factor	*df*	*F*	*MSE*	*p*	η_p^2
Term retrieval							
1	Tested-same vs Not Tested						
		Main effect of Tested-same vs. Not Tested	1,56	34.14	1.55	<.0001***	0.38
		Main effect of Block	2,112	9.096	0.29	.00022***	0.14
		Interaction	2,112	0.60	0.015	.55	0.011
	Tested-different vs. Restudied						
		Main effect of Tested-different vs. Restudied	1,56	2.43	0.04	.13	0.042
		Main effect of Block	2,112	16.23	0.26	<.0001***	0.22
		Interaction	2,112	0.89	0.018	.46	0.019
2	Tested-same vs Not Tested						
		Main effect of Tested-same vs. Not Tested	1,42	50.4	1.26	<.0001***	0.55
		Main effect of Block	2,84	3.11	0.078	.050	0.069
		Interaction	2,84	2.37	0.056	.099	0.053
	Tested-different vs. Restudied						
		Main effect of Tested-different vs. Restudied	1,42	0.88	0.013	.35	0.021
		Main effect of Block	2,84	7.044	0.13	.0015**	0.14
		Interaction	2,84	0.83	0.010	.44	0.019
Higher order questions							
3	Tested-same vs Not Tested						
		Main effect of Tested-same vs. Not Tested	1,43	14.17	0.40		0.25
		Main effect of Block	2,86	5.78	0.11	.0044**	0.12
		Interaction	2,86	2.046	0.35	.14	0.045
	Tested-different vs. Restudied						
		Main effect of Tested-different vs. Restudied	1,43	3.14	0.095	.083	0.068
		Main effect of Block	2,86	5.48	0.14	.0058**	0.11
		Interaction	2,86	1.21	0.031	.30	0.027
4	Tested-same vs Not Tested						
		Main effect of Tested-same vs. Not Tested	1,43	46.79	1.45	<.0001***	0.52
		Main effect of Block	2,86	1.15	0.029	.32	0.026
		Interaction	2,86	2.77	0.051	.068	0.061
	Tested-different vs. Restudied						
		Main effect of Tested-different vs. Restudied	1,43	9.21	0.13	.0041***	0.18
		Main effect of Block	2,86	1.42	0.041	.25	0.032
		Interaction	2,86	1.76	0.054	.18	0.040

(*table continues*)

Table 4.4. Criterial test analysis of variance (ANOVA) results (continued).

Exp.	ANOVA type	Factor	*df*	*F*	*MSE*	*p*	η_p^2
Retrieval-verification-scoring							
5	Tested-same vs Not Tested						
		Main effect of Tested-same vs. Not Tested	1,44	80.00	1.77	<.0001***	0.65
		Main effect of Block	2,88	5.95	0.14	.0038**	0.12
		Interaction	2,88	6.13	0.11	.0032**	0.12
	Tested-different vs. Restudied						
		Main effect of Tested-different vs. Restudied	1,44	20.77	1.029	<.0001***	0.32
		Main effect of Block	2,88	8.55	0.25	.00040***	0.13
		Interaction	2,88	1.085	0.024	.34	0.024
6	Tested-same vs Not Tested						
		Main effect of Tested-same vs. Not Tested	1,39	8.40	0.22	.0061**	0.18
		Main effect of Block	2,78	4.029	0.11	.022*	0.094
		Interaction	2,78	0.60	0.015	.55	0.015
	Tested-different vs. Restudied						
		Main effect of Tested-different vs. Restudied	1,39	7.54	0.19	.0091**	0.16
		Main effect of Block	2,78	3.18	0.089	.047*	0.075
		Interaction	2,78	0.078	0.0022	.93	0.0020

Note: Exp. = Experiment.

Appendix 4A

List of Process-Based Biology Concepts

Concept no.	Concept name	Example essential terms
1	Active transport	antiporter, hydrogen, sodium
2	Audioception	cochlea, hair, pinna
3	Baleen filtration	filter, plankton, whale
4	Blood flow	capillaries, oxygen, vein
5	Endosymbiotic theory	endosymbiosis, mitochondria, protist
6	Enzyme catalysis	enzyme, product, substrate
7	Exocytosis	plasma, vesicle, waste
8	Gastrulation	embryo, gastrula, germ
9	Hormonal signaling	hypothalamus, pituitary, signals
10	Immune response	clone, infection, lymphocyte
11	Membrane formation	bilayer, phospholipid, tails
12	Movement initiation	CNS, effector, receptor
13	Muscle contraction	actin, discs, myosin
14	Neuronal signaling	axon, dendrite, soma
15	Opthalmoception	bipolar, ganglion, light
16	Paracrine signaling	cell, paracrine, target
17	Passive transport	concentration, diffuse, equilibrium
18	Photosynthesis	chloroplast, light, thylakoid
19	Phylogenetic representation	ancestor, branch, node
20	Protein synthesis	RNA, transcription, translation
21	Reflex arc	interneuron, motor, sensory
22	Sexual reproduction	gametes, meiosis, organism
23	Upwelling	coastline, current, wind
24	Viral reproduction	generation, host, virus
25	Allopatric speciation	evolution, isolation, species
26	Cellular respiration	ATP, phosphate, sugar
27	Condensation reaction	condensation, covalent, hydroxyl
28	Crossing over	chiasma, chromatid, genetic
29	Dehydration reaction	bond, dehydration, monomer
30	Gene flow	allele, flow, population
31	Hydrolysis	hydrolysis, polymer, water
32	Lipid digestion	bile, fat, lipase
33	Membrane potential	electrochemical, ion, membrane
34	Metabolism	bacteria, carbohydrate, pyruvate
35	Mitosis	centromere, chromosome, mitosis
36	Plasmogamy	fertilization, fungi, plasmogamy

Note. For Experiments 1, 3, and 4, concept numbers 25-36 were also used. As described in the text, some concept names and terms were modified (i.e., replaced with synonyms) where necessary for certain question types. No. = number.

Appendix 4B

Training Phase Question Subtypes and Feedback Examples

Exp	Question type	Example	Feedback
Term retrieval			
1	Fill-in-the-blank	*The hypothalamus releases hormones which then affect the _______ and the signals it sends.*	*pituitary*
2	Fill-in-the-blank	*The hypothalamus releases hormones which then affect the _______ and the signals it sends.*	*Correct answer: pituitary* *The hypothalamus releases hormones; those hormones affect the pituitary and the signals that it sends.* *Process: hypothalamus – hormones – pituitary – signals*
Higher order questions			
3, 4	Process step	*Hormonal signaling involves two main steps. What occurs in the first step? Answer completely with all details.*	*First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the signals it sends.*
3	Order	*In hormonal signaling, what is involved first? The hypothalamus or the pituitary?*	*First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the signals it sends.*
4	Inference	*What serves as the ultimate "control center" that is in charge of the entire process?*	*First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the signals it sends.*
Retrieval-verification-scoring			
5, 6	Difficult fill-in-the-blank	*The process of hormonal regulation involves:* *The hypothalamus ____________ __________________________ _________________.*	*Did your answer include these exact words (yes or no)?* *Hypothalamus? Pituitary? Signals?* *How many out of those three did you include?* *Concept: the hypothalamus releases hormones which then affect the pituitary and the signals it sends.*

Note: Exp. = Experiment.

Appendix 4C

Example Concept Paragraph Used During Initial Study in Experiment 2

Protein synthesis is the process by which a cell makes protein. It involves DNA, a molecule that contains genetic information, and RNA, a modified version of DNA. The process is as follows: first, DNA is copied to RNA via transcription. Second, RNA is coded into protein via translation. The process therefore involves two steps: transcription and then translation. The order of the components is: DNA to RNA to protein.

Endnotes

1. Whole concept feedback arguably could be described as *explanatory feedback* given that it includes an explanation of the entire process-based concept.
2. In McDaniel, Anderson et al. (2007), cued recall and multiple-choice quiz formats were used across different weeks, with assignment of materials to each quiz counterbalanced. We focused on the cued recall results here as they are most relevant.

References

Anderson, L., & Krathwohl, D. A. (2001) *Taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290-298. doi:http://dx.doi.org/10.1037/a0031026

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

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:http://dx.doi.org/10.1177/1529100612453266

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215-228. doi:http://dx.doi.org/10.1007/s10648-016-9386-y

Freeman, S., Quillin, K., & Allison, L. (2014). *Biological science*. San Francisco, CA: Benjamin-Cummings Publishing Company.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304. doi:http://dx.doi.org/10.1002/acp.3032

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test ... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2), 307-329. doi:http://dx.doi.org/10.1007/s10648-013-9248-9

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558. doi:http://dx.doi.org/10.1080/09541440601056620

Kang, S. H., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation?. *Journal of Applied Research in Memory and Cognition*, *3*(3), 183-188. doi:http://dx.doi.org/10.1016/j.jarmac.2014.05.006

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317-326. doi:http://dx.doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition, 3*(3), 198-206. doi:http://dx.doi.org/10.1016/j.jml.2009.11.010

Lechuga, M. T., Ortega-Tudela, J., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction, 40*, 61-68. doi:http://dx.doi.org/10.1016/j.learninstruc.2015.08.002

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337-1344. doi:http://dx.doi.org/10.1177/0956797612443370

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476-490.

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:http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied, 21*(4), 370-382. doi:http://dx.doi.org/10.1037/xap0000063

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:http://dx.doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., & Little, J. L. (in press). Multiple-choice and short-answer quizzing on equal footing in the classroom: potential indirect effects of testing. In J. Dunlosky & K. Rawson (Eds.), *Handbook of Cognition and Education*, Cambridge University Press.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*(1), 18-26.

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology 107*(4). doi:http://dx.doi.org/10.1037/edu0000074

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language 83*, 53-61. doi:http://dx.doi.org/10.1016/j.jml.2015.04.001

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3), 278-292. doi:http://dx.doi.org/10.1037/xap0000124

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychological Bulletin*. doi:http://dx.doi.org/10.1037/bul0000151

Pan, S. C., Rubin, B. R., & Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied*, *21*(4), 356-369. doi:http://dx.doi.org/ 10.1037/xap0000062

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition 44*(1), 24-36. doi:http://dx.doi.org/ 10.3758/s13421-015-0547-x

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283-302. doi:http://dx.doi.org/10.1037/a0023956

Rickard, T. C., & Pan, S. C. *Test-enhanced learning of pairs, triplets, and facts: when and why does transfer occur?* Manuscript under review.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2017). Testing encourages transfer between factual and application questions in an online learning environment. *Journal of Applied Research in Memory and Cognition.*

GENERAL DISCUSSION

This dissertation research investigated retrieval practice and transfer to rearranged stimulus-response elements. Its constituent studies encompassed a wide variety of stimulus materials, subject domains, implementations of retrieval practice, feedback types, and criterial test questions. Throughout those studies, which included 953 unique subjects across 18 experiments, a consistent theme emerged: retrieval practice commonly yields *highly specific learning*. That is, (a) retrieval practice typically yields robust learning benefits relative to restudy when the correct answer is identical on both practice and criterial tests (i.e., the non-transfer questions in all experiments), replicating the "conventional" retrieval practice effect or testing effect; but (b) any enhancement is usually minimal or absent when the correct answer on the criterial test is different from that which was retrieved, or supposed to be retrieved, during training (i.e., questions involving rearranged stimulus-response elements). In other words, retrieval practice commonly yields "*potent, but piecewise, learning*" (Chapter 2). It should be noted, however, that the present results do not rule out the occurrence of any transfer entirely, but rather suggest that any transfer that did occur was not markedly greater than performance in the restudy condition (for prior discussion see Chapter 1). The most substantial exception to the overall pattern of results was observed in the final two experiments of Chapter 4, in which (c) retrieval practice coupled with extensive post-retrieval study-like activities substantially improved transfer. Finally, results from the last experiment in each of Chapters 1 and 2 suggest that (d) the benefits of retrieval practice do not diminish when all or most stimulus-response combinations are trained.

Summary and Comparison of Study Results

In Chapter 1, retrieval practice on *triple associates*, which took the form of cued recall of single missing terms followed by correct answer feedback, yielded specific and nontransferable learning regardless of whether one or two terms were retrieved during training. In Chapter 2, retrieval practice on *multi-term facts* drawn from AP History and AP Biology courses, which took the form of fill-in-the-blank or multiple-choice questions targeting single key terms followed by brief correct answer feedback, again yielded specific learning regardless of whether a key term was retrieved once or twice during training, and whether one or two key terms were retrieved during training. In Chapter 3, retrieval practice on *term-definition* neuroscience facts, which took the form of multiple-choice questions targeting a key term or its definition, yielded the same general pattern regardless of whether retrieval occurred one or three times per fact during training. Finally, in Chapter 4, retrieval practice on *process-based concepts* drawn from university-level biology courses, which incorporated concept diagrams and glossaries of key term definitions, plus featured training methods ranging from single key term recall to inference questions coupled with various forms of feedback, generally yielded specific learning. An exception to that pattern involved *retrieval-verification-scoring* trials, a new practice testing and feedback processing method that was implemented in the final two experiments of that study.

The fact that highly consistent patterns were observed across all four studies is all the more striking given the various differences between them. This includes the underlying associative structure of the materials (cf. Ross & Bower, 1980). For example, it is likely that learning a word triplet (which theoretically can be broken down into $A \leftrightarrow B \leftrightarrow C \leftrightarrow A$ associations) yields a different memory structure than a term-definition fact (which would appear to have an $A \leftrightarrow B$ associative structure but may actually involve multiple associations from a given key term to the elements of its definition, plus associations between elements of the definition; for further discussion see Chapter 3). It also includes the semantic relatedness and richness of the materials, which ranged from simple word triplets to highly interconnected, complex, and process-based biology concepts. The studies also varied in the types of retrieval practice that were used, ranging from cued recall of single terms (all experiments in Chapters 1 and 2, plus several in Chapter 4) to multiple-choice questions (Chapter 2, Experiment 3; all experiments in Chapter 3) and finally inference questions (Chapter 4). This is particularly noteworthy given prior meta-analytic and experimental evidence of the moderating effects of test format on retrieval practice effects (e.g., Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Rowland, 2014), as well as prior theorizing on the contrasting benefits of different types of practice test and quiz questions for learning and transfer (e.g., Jensen, McDaniel, Woodard, & Kummer, 2014; Karpicke & Aue, 2015). Finally, it includes 24 hr (Chapter 2), 48 hr (Chapters 2-4), and one-week retention intervals (Chapter 1).

All studies save Chapter 3 involved a three-block criterial test design, and in nearly each case the pattern of results, as evident upon visual inspection, was very similar: conventional retrieval practice effects and minimal transfer (cf. Chapter 4, Experiments 5-6). Only in two experiments in Chapter 3 did a retrieval practice effect fail to manifest for the case of definition-response questions, but that was remedied in a subsequent experiment using additional training and a change in criterial test format. Thus, the lack of transfer throughout nearly all of this dissertation research cannot be attributed to a failure to obtain a retrieval practice benefit, but rather the specificity of that benefit to tested stimulus-response combinations only.

Overall, given the consistency of the observed learning patterns across nearly the entire set of studies, it follows that the specific and often nontransferable nature of learning through retrieval practice (for materials that were presented but not retrieved during initial study and training) is likely a general property of retrieval practice itself.

Major Theoretical and Practical Implications

Of the many different candidate theories of retrieval practice effects, the present results readily align with a central assumption of the quantitative *dual memory* theory (Rickard & Pan, 2017), which is that the act of retrieval via cued recall generates a *new and separate memory* of the test event (i.e., comprised of the test stimulus and its correct answer) that is distinct from the memory that was formed during initial study (i.e., during which an intact stimulus was studied without separate question and answer elements). According to that theory, it is the second *test memory*, in addition to the original *study memory*, that is the basis of the retrieval practice effect (i.e., items which are tested are represented by two memories, whereas items that are restudied are only represented by a single study memory). When applied to the present results, the following scenario is plausible: (a) when the criterial test does not involve transfer (i.e., same test stimulus and correct answer as during training), both test and study memory can contribute to responding, but (b) when the criterial test involves transfer to rearranged stimulus-response elements, test memory cannot or only weakly contributes; thus, there is no enhancement (for related discussion involving the identical elements model, see Rickard, 2005; Rickard & Bajic, 2006; Rickard, Healy, & Bourne, 1994). Other theories of retrieval practice effects might also be adapted to account for the present results (as discussed in Chapter 1), but the dual memory theory incorporates the most directly relevant mechanism to date. Beyond the retrieval practice literature, the present results are also generally consistent with the assumptions of learning specificity proffered by Thorndike (1906), Tulving (1984), among others.

From the perspective of the broader retrieval practice and transfer literature, the present results are broadly, but not entirely, accommodated by the *three-factor framework* proposed by Pan and Rickard (2018) in their meta-analysis of that literature. In that framework, the extent of transfer resulting from retrieval practice is moderated by (a) the level of *initial test performance*, (b) whether there is *response congruency* between the practice and criterial tests, and (c) whether *elaborated retrieval practice* was used or not. Regarding (a), the level of initial test performance might moderate transfer to rearranged stimulus-response elements, at least in the sense that high practice test performance indicates greater learning of target materials and greater potential for recall of non-trained responses. However, Pan and Rickard's meta-analysis did not find significant effects for that moderator at the category level, possibly due to insufficient data. Regarding (b), by definition there is minimal response congruency in any study of rearranged stimulus-response elements, which is copacetic with the finding of no transfer in nearly all cases in that category. The finding of positive transfer for paired associates (e.g., Carpenter, Pashler, & Vul, 2006; Rickard & Pan, 2018) is the sole exception. Regarding (c), few studies in this category have used elaborated retrieval practice, but there are some indications that those methods can indeed yield transfer of this type (e.g., McDaniel, Bugg, Liu & Brick, 2015, which involved training using term-retrieval and application questions; McDaniel et al., 2007, which involved repeatedly viewed feedback; and Experiments 5-6 of Chapter 4 in this dissertation, in which retrieval-verification-scoring trials were used).

From the practical standpoint, the results of all four studies in this dissertation provide insights into effective uses of retrieval practice in classroom and other training contexts. Specifically, a general rule is that the target of practice testing (i.e., the correct answer) derives the most benefit from the technique. Thus, if the objective is to extend the benefits of retrieval to a greater proportion of to-be-learned materials, then multiple stimulus-response combinations ideally should be practiced. Alternatively, specialized feedback processing methods such as the retrieval-verification-scoring method introduced in Chapter 4 may combine the "best of both worlds" with regards to retrieval practice and subsequent learning from answer feedback.

Directions for Future Research

Followup work to this dissertation includes an already-completed retrieval practice study, Rickard and Pan (2018), in which the basis for contrasting transfer results between paired associates and triple associates was explored. That work contributes a further developed theory of the conditions under which transfer to rearranged stimulus-response items does and does not occur. Additionally, another completed study involving the pretesting paradigm, which is analogous to the retrieval practice paradigm except that there is no initial study phase and hence practice test performance is low (e.g., Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009), and in which transfer was repeatedly observed for the case of triple associate word stimuli, adds further insights into the learning patterns involved with rearranged stimulus-response elements (Pan, Lovelett, Stoeckenius, & Rickard, 2018).

Building on the findings for the retrieval-verification-scoring method introduced in Chapter 4, future research may uncover other forms of elaborated retrieval practice that can also overcome the learning specificity barrier. Additionally, future work may investigate transfer to rearranged stimulus-response elements in more naturalistic (e.g., classroom) settings. These efforts stand to further the ultimate objective of helping to refine uses of retrieval practice for the enhancement of learning, education, and training.

References

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review,* 13(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test ... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307-329. doi:http://dx.doi.org/10.1007/s10648-013-9248-9

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology,* 19(4-5), 528-558. doi:http://dx.doi.org/10.1080/09541440601056620

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, 27(2), 317-326. doi:http://dx.doi.org/10.1007/s10648-015-9309-3

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:http://dx.doi.org/0.1037/a0015729

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:http://dx.doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied,* 21(4), 370-382. doi:http://dx.doi.org/10.1037/xap0000063

Pan, S. C., Lovelett, J., Stoeckenius, D., & Rickard, T. C. (2018). *Specificity of learning following cued recall does not extend to pretesting.* Manuscript under review.

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:http://dx.doi.org/10.1037/a0016496

Rickard, T. C. (2005). A revised identical elements model of arithmetic fact representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 250-257. doi:http://dx.doi.org/10.1037/0278-7393.31.2.250

Rickard, T. C., & Bajic, D. (2006). Cued recall from image and sentence memory: A shift from episodic to identical elements representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 734-748. doi:http://dx.doi.org/10.1037/0278-7393.32.4.734

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1139-1153. doi:http://dx.doi.org/10.1037/0278-7393.20.5.1139

Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review 25*(3), 847-869. doi: http://dx.doi.org/10.3758/s13423-017-1298-4

Rickard, T. C., & Pan, S. C. (2018). *Retrieval practice on pairs, triplets, and facts: when and why does transfer occur?* Manuscript under review.

Ross, B. H., & Bower, G. H. (1981). Comparisons of models of associative recall. *Memory & Cognition*, *9*(1), 1-16.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

Thorndike, E. L. (1906). *The elements of psychology*. New York: A.G. Seiler.

Tulving, E. (1984). Précis of elements of episodic memory. *Behavioral and Brain Sciences, 7*(2), 223-268.