Journal of Applied Research in Memory and Cognition 3 (2014) 165–170

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

# Multiple-choice testing as a desirable difficulty in the classroom

Elizabeth Ligon Bjork*, Jeri L. Little[1], Benjamin C. Storm[2]

*University of California, Los Angeles, United States*

ARTICLE INFO

*Article history:*
Received 3 December 2013
Received in revised form 5 March 2014
Accepted 7 March 2014
Available online 26 March 2014

*Keywords:*
Testing effects in the classroom
Desirable difficulties
Multiple-choice quizzing

ABSTRACT

We examined whether the power of tests as learning events, frequently demonstrated in the laboratory, would also occur in a large undergraduate course. Our goals were to determine: if learning of information tested on multiple-choice quizzes administered across the course would be enhanced compared to non-tested control information; and what the effects of quizzing would be for the learning of information conceptually related to the tested information but not itself tested on the quizzes. Given that retrieval practice can have positive (*testing effect*) and negative consequences (*retrieval-induced forgetting*), our concern was that the learning and later retention of non-tested conceptually related information might be impaired by the multiple-choice quizzes. Importantly, learning benefits were found for both types of information on the final exam, indicating that quizzing within a course can enhance not only the learning of specifically tested information, but the learning of non-tested conceptually related information as well.



## 1. Introduction

As instructors, we are faced with a challenging problem: to determine the conditions of instruction that optimize student learning. This is a challenging problem because the main resources that we would use in making this determination—our intuitions, our common sense, and even our observations about what conditions seem to work best at improving our students' performance in the classroom—often turn out to be poor guides for informing our decisions.

This predicament arises because conditions of instruction that make performance improve rapidly often fail to support long-term retention and transfer; whereas, conditions of instruction that create difficulties for the learner—often slowing the rate of *apparent* learning—can actually optimize long-term retention and transfer. It is thus possible for us—as instructors, teachers, and trainers—to be misled as to what are and are not the most effective educational practices and conditions of learning.

The need to distinguish between learning and performance is an old one in psychology. Early investigators were forced to make this distinction when several, now classic, studies examining maze learning (e.g., Blodgett, 1929; Postman & Tuma, 1954) and motor-skills learning (e.g., Adams & Reynolds, 1954; Stelmach, 1969) showed that learning can occur even with no evidence of changes in performance (see Soderstrom & Bjork, in press; Tolman, 1948, for reviews of the performance/learning distinction and latent learning). More recently, a variety of findings suggest that the converse to this old learning/performance distinction is true as well: Namely, just as there can be learning without performance, there can be observable improvements in performance during training with little or no learning, with this lack of learning revealed by, say, a delayed test or a change of context. Thus, as instructors, we are confronted with the following challenge: While we can observe performance, we can only infer learning—our ultimate goal—and the former is an unreliable index of the latter. Consequently, to the extent that we interpret current performance as a valid measure of learning, we become susceptible to preferring poorer conditions of instruction and learning to better conditions of instruction and learning.

The nature of these better, but often challenging, conditions of learning that we, as instructors, should prefer is captured in the framework of *desirable difficulties* (Bjork, 1994). Such conditions include distributed practice (e.g., spacing as opposed to massing study trials), varying the conditions of practice (e.g., studying or practicing in different contexts rather than a constant context), providing contextual interference (e.g., interleaving study trials of

* Corresponding author at: Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA, 90095, United States. Tel.: +1 3104751943.
*E-mail addresses:* elbjork@psych.ucla.edu, bjork@ucla.edu (E.L. Bjork).
[1] Present address: Department of Psychology, Washington University, St. Louis, United States.
[2] Present address: Department of Psychology, University of California, Santa Cruz, United States.

http://dx.doi.org/10.1016/j.jarmac.2014.03.002
2211-3681/

different to-be-learned topics, skills, and/or categories rather than blocking them), and testing (e.g., engaging in retrieval practice of to-be-learned information rather than repeatedly studying it). They are desirable because they support better long-term retention and transfer compared to their counterparts; but because they introduce difficulties that can lower performance during acquisition or training, instructors and students alike are susceptible to perceiving such conditions as ineffective rather than desirable study strategies. As instructors, however, we need to become sensitive to the idea that short-term performance is not a reliable index of long-term learning and that difficult, or challenging, learning conditions often lead to enhanced long-term learning.

A more detailed discussion of these desirable difficulties and how both instructors and students can use them to optimize learning can be found in Bjork and Bjork (2011), and a theoretical account of why conditions that appear to hurt performance can actually help learning can be found in the New Theory of Disuse (Bjork & Bjork, 1992). We turn now, however, to a discussion of testing—the desirable difficulty on which the present study focuses—and its benefits for long-term learning and transfer.

## 2. Testing as a desirable difficulty in the laboratory and the classroom

The effects of testing for long-term retention and transfer have been the focus of much laboratory research in recent years, with many studies providing convincing evidence for the benefits of testing as compared either to no testing or to additional study, and for materials ranging from word lists (e.g., McDaniel & Masson, 1985) to paired associates (e.g., Carrier & Pashler, 1992) to text passages (e.g., Roediger & Karpicke, 2006b; for excellent reviews of the testing effect, see Roediger, Agarwal, Kang, & Marsh, 2010; Roediger & Karpicke, 2006a). Our objective in the present study was to assess the degree to which such effects would also occur in the less controlled environment of the classroom—particularly that of a large multi-sectioned undergraduate course. We had two main goals. One was to determine whether we would see benefits on a final exam for information that had been previously tested on quizzes administered across the course as compared to information that was also presented in the course but not previously tested. Our other main goal—and, in some ways, of most interest to us—was to determine what the effects of such tests might be for the retention of information that was conceptually related to the tested information but that had not itself been tested on any of the quizzes administered during the course.

*Previously tested information*. With respect to the fate of previously tested information in our study, some findings in the literature suggest that we would see a testing benefit in that such benefits have been observed in studies using education-like materials (e.g., Butler and Roediger, 2007; Glover, 1989; Kang, McDermott, & Roediger, 2007; McDaniel & Fisher, 1991; Roediger & Karpicke, 2006b). In most such studies, however, the tests were based on separate and non-related types of verbal materials, such as brief passages, rather than on the type of integrated material taught across an actual course spanning several months. Additionally, studies including comparisons of the testing benefits derived from different types of tests—in particular multiple-choice versus open-ended tests (e.g., short-answer, cued- and/or free-recall tests)—typically find retention of tested information to be better following cued-recall or free-recall testing than following multiple-choice testing (e.g., Carpenter & DeLosh, 2006; Glover, 1989; Kang et al., 2007; Butler and Roediger, 2007), which was the type of testing used in all quizzes of the present study. Exceptions to the use of individual passages and lectures, however, are a study by McDaniel, Anderson, Derbish, and Morrisette (2007), which tested materials across six weeks of a college-level web-based science course, and one by Carpenter, Pashler, and Cepeda (2009), which tested facts at two different times within a semester-long eighth-grade U.S. History course. Although in both of these studies, the information tested was better remembered on a delayed criterion test than was information not tested, the delayed criterion test was not the actual final exam that would be used in calculating the students' final course grades; whereas, in the present classroom study, both quiz performance and our criterion test (i.e., the course final-exam) did factor into the calculation of course grades. Moreover, in the Carpenter et al. study, quizzing was done with factual cued-recall questions, and in the McDaniel et al. study, which used both short-answer and multiple-choice quizzing, the benefits on the final criterion test were not statistically reliable for multiple-choice quizzing.

Recent studies with features closer to the present study are ones conducted by researchers at Washington University in Saint Louis in a public middle school. In both a sixth-grade social studies class (Roediger, Agarwal, McDaniel, & McDermott, 2010) and an eighth-grade science class (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011), benefits on final criterion tests that contributed to the student's course grade were found for the retention of content that had been quizzed versus content that had not been quizzed over the duration of the course. Thus, these latter results, in particular, would seem to imply that similar benefits for previously tested information should also be observed in the present study. Then again, several differences in these studies and ours—in particular, (a) our quizzes not being given immediately after presentation of the tested content (instead, they were given after a delay of 3–4 days); (b) our quizzes being given only once; (c) our quizzes not followed with immediate feedback; and (d) differences in the nature of the multiple-choice questions used (mostly factual in the preceding studies and mostly conceptual or inferential in the present study, as defined by the Bloom, 1956, educational taxonomy)—could mean that a benefit for the retention of quizzed information in the present study would be less likely to be observed. In particular, the timing of feedback could be critical, as at least some prior research indicates that immediate feedback following quizzing leads to better learning than delayed feedback (e.g., Kang et al., 2007; McDaniel et al., 2007), but note that other results indicate that delayed feedback can in some instances be more beneficial for long-term retention (e.g., Butler, Karpicke, & Roediger, 2007; Butler, Mullet, Verdin, von Borries, & Marsh, 2013).

*Related non-tested information*. With respect to what might be the fate of information that is related to previously tested information but that is not itself tested on any of the previous quizzes, the findings of McDaniel et al. (2011), Roediger, Agarwal, Kang, et al. (2010) and Roediger, Agarwal, McDaniel, et al. (2010) do not provide much guidance as questions on the criterion exams used in those studies were the same as those used for the quizzes. We, however, see this question—that is, how testing or giving retrieval practice to some facts or information might affect students' ability to answer questions about related but not tested information—to be critical with respect to whether low-stake quizzing should be used for learning in the classroom owing to the following considerations. First, there can be both positive and negative effects stemming from retrieval practice in that retrieval acts as a memory modifier in two different senses. While information retrieved (or given retrieval practice) becomes more recallable in the future than it would have been otherwise—the *testing effect* (e.g., Bjork, 1975; Carrier & Pashler, 1992; Roediger & Karpicke, 2006a); information in competition with the retrieved information can become less recallable—a phenomenon called *retrieval-induced forgetting* (e.g., Anderson, Bjork, & Bjork, 1994).

Although the theoretical explanation of retrieval-induced forgetting is still being debated, what has become clear, across a