**BAR-ILAN UNIVERSITY**

# Increasing the Effectiveness of Learning by Testing:

# The Relationship Between Practice Test Difficulty and Criterion Test Difficulty

Itay Weiss

Submitted in partial fulfillment of the requirements for the Master's Degree in the School of Education, Bar-Ilan University

Ramat Gan, Israel 2018

# Abstract

The current study focuses on the effectiveness of tests as a learning tool. Tests are widely used in school, but usually, they are used to evaluate students' performance, and only seldomly tests are used as a learning tool. However, numerous studies found that practicing learned material through practice tests (that involve retrieval of the learned material from memory) is effective in improving subsequent retention of the studied material more than alternative learning strategies, such as restudying the material. This advantage is called the 'Testing Effect' (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Butler, 2011; Rowland, 2014). This study joins the long line of research that examined the use of tests as a tool for learning verbal material and examines how to maximize their effectiveness.

The current study is based on a theoretical model which explains the process that occur during practice through testing (as opposed to practice through restudying), the Distribution-Based Bifurcation Model of Testing (Halamish & Bjork, 2011; Kornell, Bjork, Garcia, 2011). According to this model, practice carried out through restudying strengthen the memory traces of all the restudied items. However, practice carried out through testing strengthen only the memory traces of the items that are successfully retrieved during the test, whereas non-retrieved items are not strengthened at all. Nevertheless, the strengthening of items that were retrieved during a test is greater than the strengthening from restudying. In addition, the level of strengthening of items from testing is related to the effort involved in retrieving them from memory (Bjork & Bjork, 1992). The more difficult the practice test is, the fewer items are retrieved and strengthened, but their strengthening is greater.

Based on this model, the current study suggests that the effectiveness of learning via testing is, among other things, a function of the match between the difficulty of the practice test and the difficulty of the final test (hereinafter: the criterion test). An easy practice test will result in retrieval and strengthening of many items but the degree of strengthening will be low, whereas a difficult practice test will result in retrieval and strengthening of fewer items, but the degree of strengthening will be high. The resulting prediction is that if the criterion test is easy, it is best to prepare for it with an easy practice test, whereas if the criterion test is difficult, it is better to prepare for it with a difficult practice test. Accordingly, the research hypothesis was that there would be an interaction between the difficulty of the practice test and the difficulty of the criterion test, that will affect the performance of the criterion test as follows: Performance on an easy criterion test would be better following an easy practice test than following a difficult practice test. On the other hand, performance on a difficult criterion test would be better following a difficult practice test than following an easy practice test.

Two experiments were conducted to examine this hypothesis. Participants were 124 and 97 university students in the first and second experiments, respectively. In both experiments, materials were lists of weakly-related paired associates (e.g., lottery-luck, pocket-coat). The two experiments included three phases: a) initial study of the word pairs; b) practice test; c) criterion test. Participants performed filler tasks between phases. In both experiments, the difficulty levels of the practice test and the criterion test were independently manipulated, in order to examine the study hypothesis. Specifically, both experiments had two levels of practice test difficulty - a practice test shortly after the initial study (assumed to be relatively easy) or a practice test longer after the initial study (assumed to be relatively difficult), and two levels of criterion test

difficulty – a criterion test in a cued recall format (assumed to be relatively easy) or a criterion test in a free recall format (assumed to be relatively difficult).

In the first experiment, practice test difficulty was manipulated within participants. Each participant learned two lists of word pairs one after the other (with a filler task between them), and then was tested on both lists intermixed. The practice test was relatively delayed for the first studied list (hereinafter: delayed practice test) and relatively immediate for the second studied list (hereinafter: immediate practice test). This procedure was based on the hypothesis that immediate practice test would be easier (i.e., would results in better performance on the practice test) than delayed practice test. Contrary to what was expected, results of the first experiment showed that performance was better in the delayed practice test than in the immediate practice test. As for the criterion test, the findings supported the hypothesis that performance would be better in a the cued than in the free recall criterion test. The hypothesized interaction between practice test difficulty and criterion test difficulty was not obtained. This finding might be attributed to the limitation of manipulating practice test difficulty, as will be detailed below. However, the interaction hypothesis was examined using additional analyzes. One of them was an idiosyncratic analysis, in which practice test difficulty was idiosyncratically determined for each participant according to his or her actual performance on that test (i.e., if one's performance was better on the immediate than on the delayed practice test, then the immediate practice test was defined as the easy practice test, and vice versa). In this analysis, an interaction was obtained that was partially in line with the direction of the hypothesis. The obtained interaction indicated a larger benefit of an easy practice test than of a difficult practice test for both criterion test formats, but the advantage of the easy practice test was larger for the cued recall

criterion test (easy criterion test) than for the free recall criterion test (difficult criterion test).

A number of follow-up analyses were conducted in order to better understand the findings regarding the difficulty level of the practice test in the first experiment. The analyses indicated two limitations in the design of the experiment. One was that although the two study lists were composed to be equivalent, performance differed between the lists. In particular, it was found that for word list 2, the practice test difficulty manipulation did not work at all, as there was no difference between the immediate and delayed practice tests, whereas for word list 1, the practice test difficulty manipulation worked in the opposite direction. The second limitation that emerged from the analyses was that the participants might have remembered the first list better on the practice test because the second list that was studied after it suffered from proactive interference. It was concluded that these limitations could be overcome in a following experiment consisting of a single list of word pairs in which practice test difficulty would be manipulated between-participants, so that there would be no interference between lists. Indeed, such an experiment was carried out (experiment 2).

In the second experiment, the participants learned only one list of word pairs. Half of them took an immediate practice test (2 minutes after completing the studying phase), and the other half took a delayed practice test (10 minutes later). Criterion test difficulty was operationalized as in the first experiment. Criterion test was either in a cued recall format (easy criterion test) or in a free recall format (difficult criterion test).

The findings of Experiment 2 indicated that the manipulations of practice test difficulty and criterion test difficulty worked as hypothesized. An immediate practice test was found to be easier (that is, resulted in better performance on this test) than a

delayed practice test. In addition, cued recall criterion test was easier than free recall criterion test. Additionally, and importantly, there was also an interaction, of borderline significance, between practice test difficulty and criterion test difficulty, which was partially consistent with the hypothesis. For a cued recall (easy) criterion test, participants who were tested immediately (easy practice test) performed better than participants who were tested after a delay (difficult practice test). On the other hand, for a free recall (difficult) criterion test, no difference was found between participants who were tested immediatly versus after a delay, contrary to the hypothesis that in this criterion test participants would perforn vetter following a delay (difficult) practice test versus an immediate (easy) practice test. Nonetheless, the obtained interaction suggests that a match between practice test difficulty and criterion test criterion test is effective in increasing the contribution of the practice test. More extreme levels of test difficulty might result in findings that will better support the hypothesis, and further research, both in the laboratory and in the field, could examine this.

The present study promotes the understanding of how tests should be integrated as a learning tool, and its findings are expected to have implications for curriculum development. The second experiment suggested, as predicted, that if the criterion test is relatively easy, it is best to practice with an easy practice test. For a difficult criterion test, practicing via easy or a difficult test resulted in similar results, but follow-up experiments might provide findings that would be more consistent with the hypothesis. These results indicate that the difficulty of the practice test should be adapted to the difficulty of the criterion test. Additionally, the present study reinforces the need to seek answers to the many questions regarding the integration of tests as a learning tool. Many studies demonstrated the effectiveness of tests as a learning tool,

but many questions about the informed use of testing remain open for further exploration.