Testing is Learning:  
Re-assessing “Assessment” in Higher Education Today

Stephen Wee Hun LIM

Department of Psychology, National University of Singapore
Singapore 117570

Author’s E-mail: psylimwh@nus.edu.sg

Keywords: Assessment; Educational Practice; Retrieval-based Learning; Testing Effect.

Extended Abstract

Introduction
Students typically learn new educational materials by studying them repeatedly, while tests are normally employed as tools to assess learning (Roediger & Karpicke, 2006a). These tests refer to standardised tests, such as multiple-choice and short-answer tests, that assess students’ knowledge, understanding, and aptitude. These tests are, moreover, summative because they not only measure prior learning, but are also used to summarise performance, assign grades, and rank students. These assessment tools have been applied in education systems all over the world, and from preschool (e.g., spelling and vocabulary tests) through to graduate school (e.g., midterm and final exams using multiple-choice or other objective questions).

In contemporary educational settings, the concept of testing has a dubious reputation, however. Many modern educators believe that testing is overly emphasised in today’s schools. Some educators hold the belief that testing should be minimised, so that valuable time will not be taken away from, but instead be devoted towards classroom instruction (refer to Roediger & Karpicke, 2006b for a fuller discussion). This belief thrives even in universities where there is generally less testing. University courses commonly entail only a midterm test and a final exam. Both students and teachers tend to have a negative perception towards tests, because testing can cause high anxiety among students or bring on burdens for professors who would have to grade papers (Foong, 2008; Wittmaier, 1972).

Such a perception on testing in educational settings is unfortunate and actually misguided; testing actually constitutes a potent technique to conduce learning benefits. Intuitively, if teachers were to administer tests on a regular basis, students would have distributed their studying throughout the academic semester, instead of cramming prior to the final exams (Bangert-Drowns, Kulik, & Kulik, 1991; Leeming, 2002). Notwithstanding the traditional perspective that testing is merely a tool to assess how much knowledge a student has acquired, there is a growing body of research postulating that testing effectively promotes long-term learning. This phenomenon is known as the testing effect.
The testing effect has also been referred to as test-enhanced learning (Roediger & Karpicke, 2006a) or more recently, retrieval-based learning (Karpicke, 2012).

The Present Study
In the standard retrieval-based learning paradigm, learners either studied educational materials repeatedly, or studied and then tested themselves by retrieving the materials before taking a final test to assess their learning. Here, I report retrieval-based learning outcomes that were assessed in a series of experiments involving (a) educational materials relating to psychological research and statistical methods, (b) online learning platforms, and (c) non-verbatim tests, although only methodology and data associating with (b) are detailed in this extended abstract, owing to writing space constraints. There were two main hypotheses. First, repeated studying—relative to repeated testing—would improve learning performances when a final test was immediately administered. Second and more important, even though learners who studied repeatedly typically imagined that they would perform well on delayed tests (after a week) as compared to those who tested themselves repeatedly, repeated testing would actually produce superior performances at the final test than would repeated studying.

Methodology
Participants
One hundred and thirty-one students from the National University of Singapore, aged 18 to 26, participated either voluntarily to fulfill course requirements or with monetary compensation. All participants reported normal or corrected-to-normal vision, with no history of hearing impairment.

Materials
Two online lectures were used. Each lecture covered a single topic (“Music History” and “Brain Matter” respectively), which was divided into 30 idea units for scoring purposes. The “Music History” and “Brain Matter” videos spanned 2 minutes, 40 seconds and 2 minutes, 52 seconds respectively; both videos were transcribed into text passages that were 496 and 465 words in length respectively. In both online lectures, the respective lecturers remained visible to the viewers.

Design
The experiment used a 3 × 2 fully between-subjects design. The two independent variables were

1. Learning condition (“S” denotes study; “T” denotes test): (a) repeated study (SSSS), (b) single test (SSST), and (c) repeated test (STTT), and
2. Retention interval: (a) 5-minutes retention interval, and (b) 1-week retention interval.

The dependent variable was the mean proportion of idea units recalled.

Procedure
Participants underwent two sessions (Phase 1 and Phase 2). Phase 1 comprised of four consecutive periods. In Phase 1, participants were instructed in English regarding what they would experience during the four consecutive periods. Participants in the SSSS condition studied the online lecture for four 6-minute periods. Participants in the SSST condition studied the same lecture for three 6-minute study periods and then took one recall test during the fourth 6-minute period. Participants in the STTT
condition studied the lecture for one 6-minute study period and then took three consecutive recall tests during the next three 6-minute periods. Participants in the SSST and STTT conditions, who were required to take one or three recall test(s) respectively, were instructed to recall as much of the online lecture material as they could during each (successive) recall test.

At the end of Phase 1 and before solving multiplication problems for 5 minutes, participants were given a questionnaire that sought their responses on six questions using a 7-point Likert scale. Participants were asked to indicate (1) how interesting they thought the online lecture was (1=“very boring”; 7=“very interesting”), (2) how understandable the content of the online lecture was (1=“very difficult to understand”; 7=“very easy to understand”), (3) how understandable the lecturer’s accent was (1=“very difficult to understand”; 7=“very easy to understand”), (4) how well they thought they would remember the online lecture in five minutes or in one week’s time, depending on which condition the participants were in (1=“not very well”; 7=“very well”), (5) whether they have watched the online lecture before (circled “Yes” or “No”) and (6) how well they knew the subject matter covered prior to viewing the online lecture (1=“not very well”; 7=“very well”).

Phase 2 ensued after either a 5-minute or 1-week retention interval. Accordingly, participants in the 5-minute retention interval conditions stayed on for the experiment after completing Phase 1, whereas participants in the 1-week retention interval conditions were allowed to leave and returned exactly a week later. During Phase 2, participants were instructed to freely recall the online lecture material which they had previously studied in Phase 1; the recall instructions were identical to those given in Phase 1. The retention test lasted for 10 minutes. A schematic representation for Phase 1 and 2 appears in Figure 1.

Figure 1. Procedure outline for the SSST conditions (for both the 5-minute and 1-week retention intervals). Labels for the learning conditions (“S” denotes study; “T” denotes test).
Results

Only the main analyses and data outcomes are reported in this extended abstract. A $3 \times 2$ between-subjects analysis of variance (ANOVA) was performed, including the specified learning condition (SSSS, SSST, or STTT) and retention interval (5-min or 1-week) as the independent variables. A graphical representation of the present data appears in Figure 2.

![Graphical representation of the present data](image)

**Figure 2.** Mean proportion of idea units recalled on the final free recall test for both the 5-minute and 1-week retention intervals, as a function of the type of learning conditions (SSSS, SSST, or STTT). Labels for the learning conditions (“S” denotes study; “T” denotes test). The error bars represent standard errors.

Results revealed a significant main effect for retention interval, $F(1, 111) = 38.70, MSe = 21.26, p < .001, \eta^2_p = .26$. The main effect of learning condition did not reach significance, $F(2, 111) = 1.42, MSe = 21.26, p = .24$. There was no significant effect of the type of online lecture, $F(1, 111) = 3.15, MSe = 21.26, p = .08$. This indicated that there were no significant differences in terms of the type of content that the two different online lectures contained. Most important, results were qualified by a significant retention interval $\times$ type of learning condition interaction, $F(2, 111) = 9.88, MSe = 21.26, p < .001, \eta^2_p = .15$. This interaction suggests that repeated studying produced short-term benefits, whereas repeated testing produced retention benefits.

To ascertain the present observations directly, post-hoc analyses were performed. For the 5-minute retention interval, participants in the SSSS condition ($M = 19.81, SD = 4.26$) recalled significantly more than did participants in the STTT condition ($M = 14.77, SD = 5.31$), $t(41) = 3.42, p < .005, d = 1.06$, albeit not significantly more than did participants in the SSST condition ($M = 19.09, SD = 2.89$), $t(35.02) = .645, p = .52$. Participants in the SSST condition recalled more than did the participants in the STTT condition, $t(32.42) = 3.35, p < .005, d = 1.18$. However, this pattern was reversed for the 1-week retention interval. Participants in the STTT condition ($M = 14.29, SD = 4.49$) recalled more than did the participants in the SSSS condition ($M = 10.35, SD = 4.66$), $t(37) = 2.69, p < .05, d = .88$, albeit not significantly more than did the participants in the SSST condition ($M = 13.37, SD = 5.51$), $t(36) =$
.57, p = .57. In addition, participants in the SSST condition recalled more than did the participants in the SSSS condition, although this difference was just approaching significance, t(37) = 1.85, p = .07, d = .61. This suggests that a single test can conduce a substantial mnemonic benefit for long-term performance. More important, the results show that repeated testing led to retention benefits.

**Conclusion**

This study reports novel evidence of facilitatory retrieval-based learning effects which emerged across a series of new contexts involving, e.g., an online learning platform. Retrieval-based learning is more than merely a passive assessment of learning. Rather, the process of retrieval itself is central to active learning and represents a powerful tool for improving learning. The present data suggest the need for reassessing the concept of “assessment”, due to the fact that assessments and testing can, when optimally designed and leveraged on by educators, promote active learning and long-term retention of knowledge.
Acknowledgement

The Author gives credit to two National University of Singapore Faculty of Arts and Social Sciences Heads and Deanery Research Support Grants (R-581-000-150-133 and R-581-000-160-101) which supported this work.

References


