Journal of Educational and Developmental Psychology; Vol. 3, No. 1; 2013
ISSN 1927-0526 E-ISSN 1927-0534
Published by Canadian Center of Science and Education

Effect of Repeated Testing on the Development of Secondary Language Proficiency

Jari Metsämuuronen[1]

[1] Faculty of Behavioral Sciences, Helsinki University, Finland

Correspondence: Jari Metsämuuronen, Faculty of Behavioral Sciences, Helsinki University. Siltavuorenpenger 5, 00014, Helsinki University, Finland. Tel: 358-400-579-848. E-mail: jari.metsamuuronen@gmail.com

Received: August 7, 2012 Accepted: August 29, 2012 Online Published: January 7, 2013

doi:10.5539/jedp.v3n1p10 URL: http://dx.doi.org/10.5539/jedp.v3n1p10

Abstract

Karpicke & Roediger (2008) showed that delayed recall is optimized, not with repeated studying sessions, but with repeated testing sessions. The result was very soon re-interpreted by Lasry, Levy and Tremblay (2008), who hypothesized that repeated testing may lead to multiple traces in one's memory, which facilitates recall. The aim of the study was, first, to test whether the results of the Karpicke & Roediger study of repeating similar tests are applicable in real life with dissimilar, though equated, tests of a gradually increasing level of difficulty in Biblical Hebrew language with no connection to the use of the spoken language in everyday life. Second, the aim was also to reveal the profiles of the students' proficiency levels of during the study process. Randomized, matched-pairs (N = 7 + 7) of students with the same teacher, the same lessons and the same routines participated in the experiment. The intervention for the experimental group consisted of a set of short, ten-minute tests during each three-hour study session held twice weekly. IRT modeling with linking items served to equate the test scores. ANOVA was used to analyze the gain score between the pre-test and post-test. The experimental group improved statistically significantly more than did the control group. The learning curve was a nonlinear, wide U or J shape after the first elementary period of learning the letters and basic vocabulary. The repeated testing sessions helped the students raise their language proficiency level more than without the repeated testing sessions. Although the groups were small, the effect was high (Cohen's $d > 1.0$ or $d > 1.5$ depending on the indicator). The result supports the idea that Karpicke & Roediker's results may also be applicable when the tests measure different topics with the different in each test.

Keywords: language testing, secondary language, Biblical Hebrew, IRT modeling, experiment, matched-pairs

1. Introduction

For more than 40 years since Neisser (1967; see also Eysenck & Keane, 2005), cognitive psychology has attempted to solve the universal computational laws of the human brains. Cognitive psychology works intensively with, for example, the working memory (e.g., Baddeley, 2000; Engle, 2002; Conway *et al.,* 2003; Conway *et al.,* 2005) and the development of language (e.g., Carrol, 1994; Whitney, 1998; Harley, 2001; Jay, 2003). The basic theories of the human mind claim that humans have two different memory modes: semantic memory and episodic (or narrative) memory (see Tulving, 1983; Bruner, 1986; 1990). According to Tulving (1983, 9), semantic memory handles knowledge concerning the world; it is independent of the identity of the person and of personal history, whereas the episodic memory consists of a store of personal events, actions and memories. The content of semantic memory is something the individual *knows,* whereas the content of episodic memory is something the individual *remembers*. The units of semantic memory are facts and concepts, whereas the units of the episodic memory are events and episodes. Semantic memory is organized according to concepts, whereas episodic memory is organized according to time.

Cognitive models also assume that retention and retrieval can be explained by co-operation between working memory and long-term memory. Long-term memory can be divided into two components: declarative memory which includes episodic and semantic memory, and procedural memory, which applies to skills. According to Tulving and Schacter (1990), declarative memories are best established by using active recall combined with mnemonic tools and spaced repetition. Thus, a basic doctrine of human learning and memory research is that the repetition of material improves its retention. A rather interesting recent experiment in language learning (Karpicke & Roediger, 2008; also in Roediger & Karpicke, 2006a; 2006b) challenged this tenet.The experiment

showed that delayed recall is optimized, not with repeated studying sessions, but with repeated testing sessions. The result was soon re-interpreted by Lasry, Levy and Tremblay (2008), who hypothesized that repeated testing (say, the retrieval of the memory) may lead to multiple traces of the memory, which facilitates recall. Lasry *et al.* suggest that the new interpretation could lead to a new framework for explaining the effectiveness of frequent in-class assessments in pedagogies such as Peer Instruction. Which interpretation is more correct less important that the fact they lead to the same conclusions: we need to re-evaluate what are considered the most effective ways to learn languages.

2. A Brief Literature Review of the Testing Effect

2.1 The Testing Effect in the Laboratory Setting

The Testing effect, that is, the phenomenon of improved performance from taking a test, is a rather old idea. The years since Gates (1917) and Spitzer (1939) have seen many experiments in laboratories concerning the effects of testing. According to Roediger and Karpicke (2006a), the surprising and counterintuitive results of Tulving (1967) created a boom in the field. Before Tulving's experiment, the common assumption was that learning occurs only (or best) during study sessions. According to Tulving's results, however, the proportion of recalled words in the tests given to the in standard Study-Test-Study-Test group, the Repeated study (Study-Study-Study-Test) group, and the Repeated testing (Study-Test-Test-Test) group were the same although the Repeated study group had studied the words six times more than had the Repeated testing group. Their learning curves were also identical. These results were soon replicated with minor variations. Karpicke and Roediger (Roediger & Karpicke, 2006a; 2006b; Karpicke & Roediger, 2008) replicated Tulving's design with 40 pairs of English words and their Swahili equivalents and noted the same effect as that noted earlier. However, the new finding suggested that though the students in all the groups learned the same number of words, one week later the Repeated testing group recalled the words better than did the other groups. Hence, Karpicke and Roediger inferred that repeated testing optimized delayed recall, that is, retrieval from memory. These results seem to have again motivated researchers to work on the topic (see Carpenter, 2009; Chan, 2009; Chan & McDermott, 2007; Chan, McDermott & Roediger, 2006; Kang, McDermott & Roediger, 2007; Karpicke, 2009; Karpicke, Butler & Roediger, 2009; Karpicke & Roediger, 2007; Karpicke & Roediger, 2010; Kester & Tabbers, 2008). One recent finding of these studies is that while expanding retrieval practice, that is, the idea of gradually increasing the spacing between repeated tests, is usually regarded as a superior technique for promoting long-term retention (see Landauer & Bjork, 1978), the better technique for long-term retention seems to be *equally* spaced retrieval, that is, when the repeated tests are taken at equally spaced intervals (Karpicke & Roediger, 2007; 2010). Another interesting result is that although Anderson, Bjork & Bjork (1994) argue that repeated testing may impair one's later recall of untested material, that is, the long-term remembering of some information may can cause one to forget other information, recent results have shown that repeated testing can sometimes also improve one's later recall of the untested material (Chan, 2009; Chan, McDermott & Roediger, 2006). Thus, at least the results appear to hold in laboratory situations, where such tests usually include the same repeated set of material (and thus, the design actually tests the effect of repeatedly testing the same test material), and the amount of information retained is trivially small in comparison with the real life.

2.2 The Testing Effect in the Classroom Setting

A challenge in the classroom setting is that arranging a genuine experimental situation is seldom easy. Therefore, doe to the uncontrolled studying of the students, students' interest in the course material, and students' motivation to learn, the inferences are difficult to conduct as directly as in the laboratory setting (Roediger & Karpicke, 2006a). However, Bangert-Drowns, Kulik and Kulik (1991) conducted a meta-analysis of 35 real-life studies that manipulated the number of (dissimilar) tests given to the students. Most of the studies were carried out in college classrooms; 29 of the 35 studies failed to randomly assign students to the repeated testing group and control group and were thus mainly, at the best, quasi-experiments. Still, it is noteworthy that 83% of the studies found a positive effect of frequent testing. According to Bangert-Drowns, Kulik and Kulik's model the effect size was, on average, closely related to the number of tests taken during the semester; the greater the number of tests, the greater the difference between the groups. Several recent studies of the testing effect have also found a positive connection between the testing and later test results in various university courses (see Leeming, 2002, "Introductory psychology"; McDaniel *et al.,* 2007, the "Brain and Behavior" course; Cranney *et al.,* 2009, "Psychobiology"; Vojdanoska, Cranney & Newell, 2009, "Psychology"; Johnson & Mayer, 2009, "Multimedia learning"). Karpicke, Butler and Roediger (2009) noted that relatively few students engage in self-testing while studying although they repeatedly read their notes or textbooks. It may also be worth noting Gurung and Daniel's (2006) result, which showed that supervised tests were associated with better examination performance, and unsupervised settings with poorer performance. The reason may be that in unsupervised

settings, students feel they are testing themselves, but their behavior may be less effective, as Karpicke, Butler and Roediger (2009) suggested. Although the bulk of the research on the testing effect is convincing, it is surprising – as McDaniel and his colleagues (2007) note – that the educational community has virtually ignored the testing effect on literature for educational practice.

2.3 The Testing Effect and the Tests

Testing procedures can be divided into static testing and dynamic testing (see Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2001; 2002; Roediger & Karpicke, 2006a). Static tests involve an examiner, who gives the test to the test-taker without providing feedback about his or her performance on the test (or providing feedback at the very end of the testing process). Static testing is used when it is important for the correct answers to the test items to remain unknown; items such as those in IQ or SAT tests or in some National learning outcome programs that link items to keep the test scores from different years comparable, must not be released. Releasing linked items would allow students to learn test items, would render impossible the comparison of student results from subsequent years. It is worth noting that the use of static testing makes it is possible to test the pure testing effect.

Dynamic tests involve giving feedback to the test-takers in order to help them improve their scores for the next test. Dynamic tests are used when one is willing to teach the topic through testing; incorrect answers are corrected, which reveals the learning potential or underlying capacities of the test-takers. The test results of the final test appear to improve when feedback on the performance is forthcoming during the process rather than when it is not (see Butler, Karpicke & Roediger, 2007; 2008; Butler & Roediger, 2008; Metcalfe, Kornell & Finn, 2009; Vojdanoska, Cranney & Newell, 2009). It is worth noting that by using dynamic testing, the examiner can test the effect of learning through testing and teaching the correct answers rather than the testing effect alone.

3. Aim of the Study

The aim of the study was two-fold. Firstly, the aim was to test whether the laboratory results of the Karpicke & Roediger study about repeating the same set of materialsare applicable in a real-life situation with a dissimilar, though equated, set of tests of a gradually increasing level of difficulty, in Biblical Hebrew language with no use for the spoken language on everyday life and without releasing the correct answers after the test (or, by static testing). The last characteristic means that the test design tests only the testing effect, not teaching through testing. Second, the further aim was to reveal the profiles of the students' proficiency levels during the study process with a matrix design of repeated, linked tests and Item Response Theory (IRT) modeling.

4. Methods

4.1 Sample

The subjects of the study are those students of theology at Helsinki University who participated in Biblical Hebrew lessons in the autumn of 2009. The training for Biblical Hebrew at Helsinki University aims at knowledge of the most frequent vocabulary; analysis of the nominal structures; verb morphology and syntax; and it mainly aims, thus, understanding about the written language with the use of devices, such as dictionaries. The faculty has three teachers and three study groups and the experiment was carried out in one of those groups. The students were assigned to one of the language lesson groups and thus should regularly follow the course program for their specific group. However, although the language program is compulsory for most of the students to graduate, many students drop out and begin the courses again, some of them several times. Some students also turn to individual studying instead of opting for classroom teaching. Thus the experiment began after the first of three periods; the number of students who followed the course was somewhat fixed after the first seven weeks (that is, after the first of three periods). Altogether 30 students began the period in the group. These students were randomized into two matched groups (N = 15+15) on the basis of their language proficiency level. However, two students from the experiment group and eight students from the control group dropped during the process. Thus, only seven matched pairs remained in the experiment and are reported in this article although 13 students remained in the EG.

Several other students were also tested in addition to those in the experimental and control groups. All those students who began the course in any of three study groups were tested three times to reach stable item difficulty parameters for the tests later in the experiment. In the very first sessions, 96 students took the test before starting to teach Biblical Hebrew; three weeks later, 66 students took the second test. After the first full period (7 weeks), 51 students ended the first period with classroom teaching and were then tested the third time (Table 1). In the last testing session after the experimental phase, that is, in the post-test, only 33 students took the test. This decreasing trend in student numbers indicates that many students prepared for the second official test at home at

the end of the second study period. This trend may also indicate the student custom of working during the one-month Christmas break, which precluded them from attending the last lessons.

4.2 Design & Hypotheses

The design itself is a classic pre-test/post-test design supplemented with two preliminary tests prior to the intervention (Figure 1).

R	Experimental group	O	O	O	X	O
R	Control group	O	O	O		O

Figure 1. The study design

The students were randomly assigned into two groups. The assignment was completed on the basis of the last of three common tests for all students; as closely matched-pairs as possible were selected: one for the EG and one for the CG. Both groups participated in the same courses of Biblical Hebrew by the same teacher at the same time and in the same manner. The control group (CG) attended lessons according to normal routines. The experimental group (EG) also attended lessons according to the same normal routines but repeatedly took a short, ten-minute test during each three-hour session twice a week without receiving any feedback on the test results: thus, the testing was based on static tests. During the test, the CG studied the course book. Before the random assignment, both groups (with all other students) were pre-tested three times 1) to establish the baseline of the study prior to participating in any Biblical Hebrew lessons, 2) to reduce the (negative) testing effect caused by, for example, unknown item formats, 3) to test suitable item formats for the later experimental testing, 4) to gain stable item parameters for linked items, and 5) to obtain the information on the proficiency level of the students to suitably match the pairs.

It is worth noting that although the students were randomly assigned, the real-life situation affects that it is not possible to control how much the students study, their level of interest in the course material, or their motivation to learn, as Roediger and Karpicke (2006a) noted.

On the basis of previous results (e.g., Cranney *et al.,* 2009; Johnson & Mayer, 2009; Karpicke & Roediger,2008; Leeming, 2002; McDaniel *et al.,* 2007; Roediger & Karpicke, 2006a; 2006b; Vojdanoska, Cranney & Newell, 2009), the statistical hypothesis is one-sided: the alternative hypothesis is H_1: *The gain score is higher in the EG than in the CG*.

4.3 Missing Values

Because of the nature of the longitudinal experiment in real-life tertiary education, forcing student attendance in the study sessions every time was impossible. However, 13 students in the EG studied in the classroom learning manner and thus participated nearly all the study sessions (Table 1). The students were eager to participate in the testing process, so when they were unable to attend a lesson, they were willing to take the test the very next day on their own time or during the next study session.

Table 1. Number of test-takers in the different phases of the experiment

	Test 1. Zero-line	Test 2.	Test 3. Pre-test	Tests 4. – 12. Intervention	Test 13. Post-test
Experimental group	N = 15	N = 15	N = 15	N = 13	N = 13
Control group	N = 15	N = 15	N = 15	-	N = 7
All students	N = 96	N = 66	N = 51	-	N = 33

Some students had missing values in their long sequence of test scores although replacements for the missing values were usually easy to model (see Fig. 2.). Several strategies were used. Non-linear modeling, as in Figure 2, was used to model the obvious, unexpected peaks in the trend. The mean score for two tests (the ones immediately before and after the missing test score) was used and when the trend was obviously linear and the trends of the other test-takers were similar. In one case, the missing value of a weaker student was replaced by modeling the downward trend based on the mean of corresponding change in the smaller group (the weakest

students). Altogether 12 missing values out of 162 (= 13*13) observations (7%) were replaced in this manner.

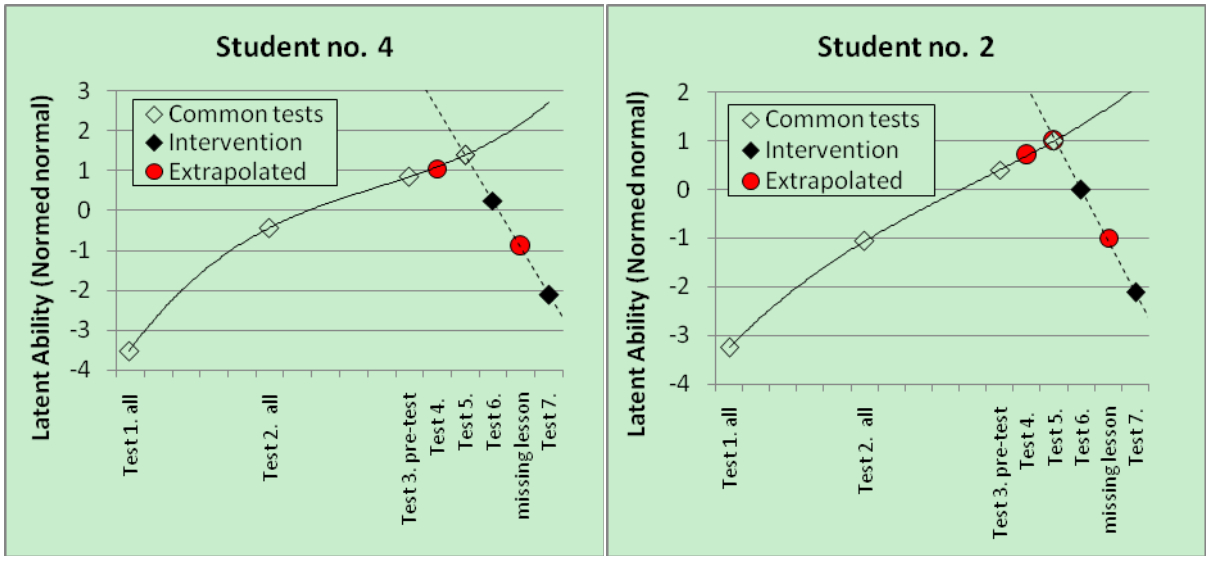

Figure 2. Examples of modeling the missing values

As noted above, several students dropped out of the CG. Unfortunately, most of them came from the two extremes of the proficiency scale: either the lowest or the highest. Thus, the remaining part of the CG (N = 7) were clustered mainly in the middle range of the proficiency scale. New matches were made so that the means of the groups would be as close as possible. This change resulted in lower though statistically insignificant (p = 0.125) variance than in EG.

Table 2. Basic statistics of the EG and CG in the pre-test

Group		pre-test
Experimental group (N = 7)	Mean	1.63
	Std. Deviation	0.94
Control group (N = 7)	Mean	1.76
	Std. Deviation	0.59
Total	Mean	1.70
	Std. Deviation	0.76

4.4 Tests and Items

Altogether 239 items were written for the tests, but only 218 were used in the last calibration. The 21 omitted items were either too difficult (no one knew the correct answer) or too easy (all knew the correct answer) for the test-takers. The items covered a wide range of Biblical Hebrew skills beginning with Recognition of the Hebrew letters and Recognition of transliterated Hebrew words to Verb morphology, including the recognition of the Perfect, Imperfect, Jussive, Imperative, Cohortative, Infinitive and Participle forms as well as the personal and objective Suffixes (Table 3). The wide range of topics in the tests serves as a face validity indicator that the tests were valid for testing Biblical Hebrew. All the tests (except the very first one) were constructed as a review of the earlier study sessions; no new material was tested in the tests. Thus, the tests reflected the expected increase in the proficiency level of the students and were, in principle, a set of the tests of increasing difficulty.

The number of items on the tests varied from 13 to 44 (during the intervention, from 16 to 32), reliabilities ranged from 0.53 to 0.97 (during the intervention, from 0.79 to 0.94), and the mean of the item-total correlations

varied with few exceptions from 0.40 to 0.60 (during the intervention, from 0.43 to 0.60), the last indicating sufficiently or highly discriminating items. The high reliabilities in the EG during the intervention also indicate that the experimental group included a wide range of proficiency levels and, thus, high discrimination.

Table 3 shows that 366 items were used in different test versions, 218 of which were unique. Thus one can infer that on average 40% (148/366) of the items overlapped, that is, on average 40% of the items on every test were linked to the previous tests. Some items served as links several times. The number of linked items is sufficient for equating.

With regard to the administration of the test four important notes merit emphasis. First, of 366 administered items, 255 (71%) were objective, multiple choice questions (MCQ) with three alternatives, 105 (29%) were objective, short answer items (Translation of a short expression or Translation of a verb), and 6 were translations of a longer sentence. Second, the time spent on the tests on the intervention was set at ten minutes even though the length of the tests on the intervention varied from 16 items to 33 items. It is worth noting that even the weakest students managed to answer all 33 items, probably because the MCQs are quickly answered. The last test, with its 44 items, was longer than the others and took 15 minutes. Third, the tests were administered in the middle of the three-hour study sessions so that the students would be "warmed up", but not exhausted. Fourth, the students did not receive the correct answers after each test. Rather, the tests were collected ten-minutes later and the study sessions followed without reference to the tests or test items. Thus, no test items were handled in the study sessions and the experiment therefore measures the effect of pure testing, not of testing and teaching through testing.

Table 3. Characteristics of the tests administered during the process

Test version	No. of items	Alpha reliability	Mean of the Item-total correlations	Content covered in the test
Test 1 A & B	27	A: 0.88 B: 0.89	A: 0.49 B: 0.50	Recognition of transliterated Hebrew words (k = 10); Recognition of short transliterated phrases (k = 3); Recognition of the Hebrew letters (Hebrew - Latin) (k = 6); Recognition of Hebrew words (Hebrew - Latin) (k = 4); Meaning of frequent words (Hebrew - Pictures) (k = 4)
Test 2 A & B	19	A: 0.73 B: 0.66	A: 0.41 B: 0.37	Recognition of the Hebrew letters (Hebrew - Latin) (k = 6); Alphabetical order of the words (Hebrew) (k = 2); Meaning of frequent words (Hebrew - Pictures) (k = 5); Transcription of Hebrew words (k = 6)
Test 3 A & B	13	A: 0.64 B: 0.53	A: 0.44 B: 0.39	Alphabetical order of the words (Hebrew) (k = 3); Meaning of frequent words (Hebrew - Pictures) (k = 3); Construction of a nominal clause (Hebrew) (k = 3); Structures (k = 4)
Test 4	19	0.80	0.47	Construction of a nominal clause (Hebrew) (k = 3); Translation of a short sentence (k = 2); Structures (k = 6); Verb morphology (Perfect) (k = 4); Translation of verbs (k = 4)
Test 5	16	0.81	0.51	Verb morphology (Perfect) (k = 3); Translation of verbs (k = 10); Meaning of infrequent words (Hebrew - Pictures) (k = 3)
Test 6	22	0.79	0.43	Structures (k = 6); Translation of a short expression (k = 4); Verb morphology (Perfect) (k = 4); Translation of verbs (k = 5);

Test 7	29	0.92	0.55	Meaning of frequent words (Hebrew - Pictures) (k = 3) Structures (k = 6); Translation of a short expression (k = 7); Verb morphology (Perfect - Imperfect) (k = 6); Translation of verbs (k = 6);
Test 8	20	0.84	0.49	Meaning of frequent words (Hebrew - Pictures) (k = 4) Structures (k = 5); Translation of a short expression (k = 6); Translation of a sentence (k = 1); Translation of verbs (k = 3);
Test 9	21	0.90	0.56	Verb morphology (Perfect - Imperfect - Suffixes) (k = 5) Structures (k = 4); Translation of a short expression (k = 6); Translation of a sentence (k = 1); Translation of verbs (k = 5);
Test 10	24	0.90	0.54	Verb morphology (Imperfect - suffixes) (k = 5) Structures (k = 5); Translation of a short expression (k = 6); Translation of a sentence (k = 2); Translation of verbs (k = 2); Meaning of frequent verbs (Hebrew - Finnish) (k = 5);
Test 11	32	0.94	0.60	Verb morphology (Perfect - Imperfect - Suffixes) (k = 4) Structures (k = 4); Translation of a short expression (k = 6); Translation of a sentence (k = 1); Translation of verbs (k = 11); Meaning of frequent verbs (Hebrew - Finnish) (k = 4);
Test 12	23	0.84	0.49	Verb morphology (Juss.-Imperat.-Coh.) (k = 6) Structures (k = 7); Translation of a short expression (k = 5); Verb morphology (Juss.-Imperat.-Coh.-Infinit.) (k = 5); Translation of verbs (k = 6)
Test 13 Experiment	44	0.89	0.46	Meaning of frequent words (Hebrew - Pictures) (k = 8); Structures (k = 5); Translation of a short expression (k = 7);
Test 13 Control	44	0.73	0.32	Translation of a sentence (k = 1); Meaning of frequent verbs (Hebrew - Finnish) (k = 3); Recognition of the root of a verb (k = 4);
Test 13 others	44	0.96	0.60	Verb morphology (Perf.-Impf.-Juss.-Imperat.-Coh.) (k = 8); Translation of verbs (k = 8)

4.5 Equating and IRT

The intervention consists of short tests where a certain number (on average 40%) of items were linked to previous tests in order to estimate the sample-free item parameter for item difficulty and to estimate the latent proficiency level of the students in the experimental group. Thus all the separate short tests were linked to each other with identical linked items to the previous tests and ultimately to the first pre-test. Thus, it is possible to equate the test scores with IRT modeling (i.e., Rasch, 1960; Lord & Novick, 1968; Hambleton, 1993; Béguin, 2000) and finally to acquire the latent ability of each student. IRT models are widely used in language testing (see Takala, 2009; Kaftandjieva, 2004; Verhelst, 2004). The specific characteristic of all IRT models – the sample-free feature of the analysis (see Wright, 1968) – had led to 40 triumphant years for IRT modeling in relation to Classical Test Theory (Gulliksen, 1987). Laboratories and universities all over the world have contributed to the development of IRT modeling since the days of Lord (1952), Rasch (1960), Birnbaum (1968), and Lord and Novick (1968). Recently, much effort has focused on multidimensional IRT (MIRT), Differential Item Functioning (DIF), Computer Assisted Testing (CAT) and Test equating (see Stout, 2002). IRT modeling is also the very tool used to equate test scores in the well-known international comparisons of PISA studies and Trends in Mathematics and Science Studies (TIMSS). Rasch modeling (Rasch, 1960), that is, one parametric logistic IRT model was used in the modeling in this study.

The need for equating arises because it is unwise to use identical tests with expert-and novice level students; obviously, a test focusing on more an expert-level testee should be more difficult than one measuring a novice level test-taker. Also, test lengths may differ (i.e., comparing shorter tests with the final). Prior to making the comparison of test scores meaningful, the scores must be made comparable. This is accomplished by vertical equation, that is, by equating the test scores from different tests of different lengths administered in different time frames.

Vertical equating was administered according to IRT modeling with the following principles and practices: the scores were fitted to the same scale on the basis of characteristics of IRT models, which assume that a learner's latent level of ability (θ) and the difficulty level of an item (β) are identical, when certain preconditions are met (see Wright, 1968). The latent ability of each learner can be determined in the same scale for every test as long as linked items connect the test versions. Because of the small number of students in the experimental group (N = 13), the only recommendable model for estimating latent ability was the one-parameter model (that is, the Rasch model). The estimation was carried out using the OPLM program (Verhelst *et al.,* 1995). A brief technical description of the equation process appears below (for more detailed, see Béguin, 2000, 17–36):

1) Define the structure of the linked items. Because the values of "difficulty parameter" for the linked items are exactly the same in each version, the difficulty levels of all other items are calibrated to the same scale as the linked items are.
2) Use the *Conditional Maximum Likelihood* (CML) procedure to estimate each item's difficulty level (β parameter).
3) Use the *Marginal Maximum Likelihood* (MML) procedure to estimate the distribution of means and standard deviations for each student's latent ability (θ parameter) in each version.
4) Estimate the θ parameter of the scores for each version using the means and deviations of the distributions of β and θ. This results in a unique latent value, measured on a common scale, for each observed value of the scores in all versions.

The success of the equating depends on three things. First, the linked items should reflect the proficiency level of the test-takers. They should represent a sufficient range of ability; items that are too easy and too difficult should, however, be avoided. In the intervention, the linked items for the next test were selected on the basis of the previous tests; those sufficiently discriminating items were selected which were neither too easy nor too difficult. Second, the linked items should represent a short test within the test; the items should cover the different content areas. In the intervention, the linked items were selected so that they would represent as widely as possible different content areas. Third, the stable parameters in the equation process depend on the sample; the better the sample represents the target population, the better the calibration corresponds with the population parameter. In the intervention, all the students were tested at the beginning to obtain as large population as possible in order to acquire stable item parameters. However, from the viewpoint of the population, the parameters for items measured only in the EG are unstable. Also, due to the small population in the intervention, the values for item "difficulty parameter" depended considerably on those students who participated in the test. Thus it was important to get all the possible test papers from the test-takers – even day after the test. Although the item parameters are somewhat vague, the results are much more accurate than if only classical metrics (the proportion of correct answers) were used in comparison. The missing values for the latent ability (θ) were modeled as described in Section 4.3.

In the Results section, the original θ values (Normed normally distributed scores) for latent ability are used when comparing the groups. An average student in the whole population – combining all tests – would score $\theta = 0$, and the higher the proficiency level is, the higher is the θ value. The values usually range from –4 to +4 in the Normed normally distribution. As a result of the procedure, each student acquires equated scores, that is, scores that are on the same scale for each version of the tests. Thus it is possible, first, to test whether the proficiency level of the students in the EG is higher than the level of those in the CG (comparing the pre-test- and post-test scores without administering parallel tests), and second, to create the profiles of the students (learning curves based on the equated scores for the test-takers) during the process.

4.6 Analysis

Standard methods are used to analyze the pre-test/post-test design. Two different practices for analysis are used. Analysis of Covariance (ANCOVA) is usually used with randomized experiments (see Cribbie & Jamieson, 2004; Miller & Chapman, 2001). However, because this method presumes reasonably large sample size and an identical pre-test distribution for all groups, it is therefore a less desirable choice in situations where the sample

size is small and the variance within the groups differ due to the small sample size. Another possibility is to analyze the gain score with the standard Analysis of Variance (ANOVA) procedure. Because of the small number of observations and the lower variance within the control group, the gain score served as a basis for inferences. The gain score was analyzed with the t-test and the results were confirmed with a non-parametric Mann-Whitney U test because of the notably small sample size. Cohen's *d*, calculated on the basis of the t-value, was used to determine the effect size. Also, an alternative formula for*d* for experimental studies (see different options in, e.g., Morris, 2008) is used for referential purposes:

$$d = \frac{\left(\overline{x}_{e.post} - \overline{x}_{e.pre}\right) - \left(\overline{x}_{c.post} - \overline{x}_{c.pre}\right)}{\sigma_{pooled.pre}}$$

where *e* refers to the experimental group, *c* refers to the control group, *post* refers to the post-test, *pre* refers to the pre-test, and $\overline{x}$ and σ refer to the mean and standard deviation in the groups.

5. Results

The main result is that the experimental group improved statistically significantly more than did the control group (Table 4 and Figure 3) during the intervention ($t_{(12)}$= –1.637, $p_{(one\text{-}sided)}$= 0.064*;*Mann-Whitney U: *Exact p* = 0.036). The difference between the gain scores is higher than the variances within the groups; the size of the effect is high (Cohen's *d = 0.95,* and when using the formula for the experimental design, *d* = 1.11*)*. On the basis of η^2= 0.18, the experiment explains 18% of all variance in the data, which is also high.

Table 4. Group statistics and ANOVA table for gain scores

Group	N	Gain Score Mean	Std. Deviation	Pre-test Mean	Post-test Mean	Pre-test Std. Deviation	Post-test Std. Deviation
Experimental group	7	0.77	1.03	1.68	2.41	0.94	0.77
Control group	7	–0.07	0.90	1.76	1.63	0.59	0.87

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig. (one-sided)	Eta Squared
Gain Score	Between Groups	(Combined)	2.508	1	2.508	2.679	0.064	0.182
		Within Groups	11.233	12	0.936			
		Total	13.740	13				

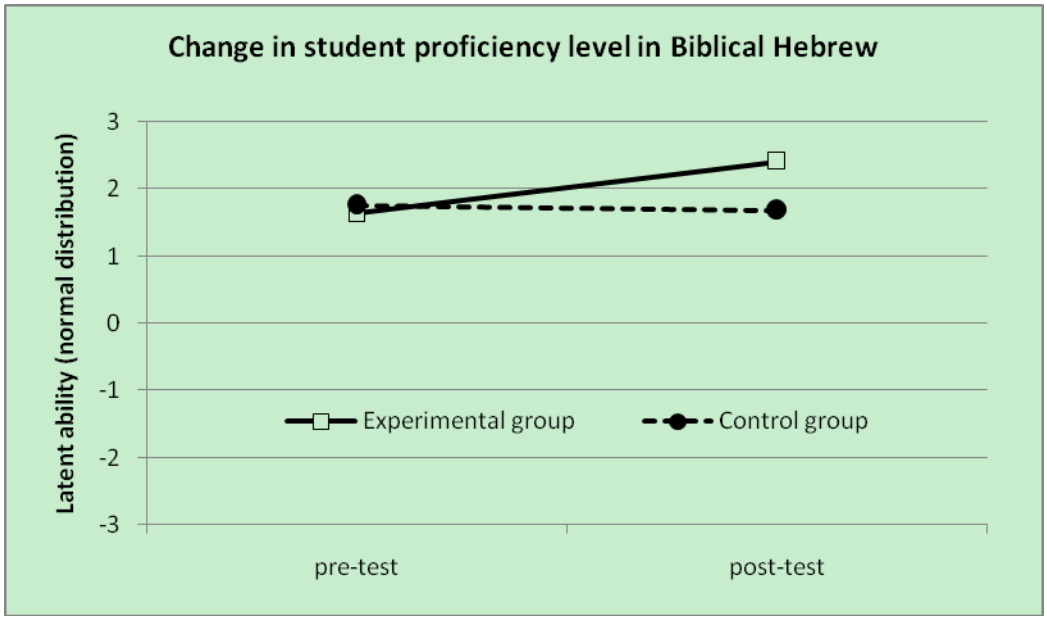

Figure 3. Change in student proficiency level: comparison of the groups at pre-test and post-test

In practice, the result means that starting from the slightly lower mean in the pre-test, the EG improves by 0.77 standard units when at the same time, with the same teacher, with the same lessons, but without the intervention effect, the CG, paradoxically, lost proficiency, that is, the proficiency level *decreased* (–0.07 standard units).

The lower proficiency level in the CG can be attributed to the interesting profile of language proficiency development in the experimental group, provided that the process was the same in the CG as in the EG (see Figure 4). In Figure 4, the profile of the experimental group is fulfilled by intra polation in order to avoid the misleading time perspective between the preliminary tests (three weeks) and the intervention tests (two tests per week).

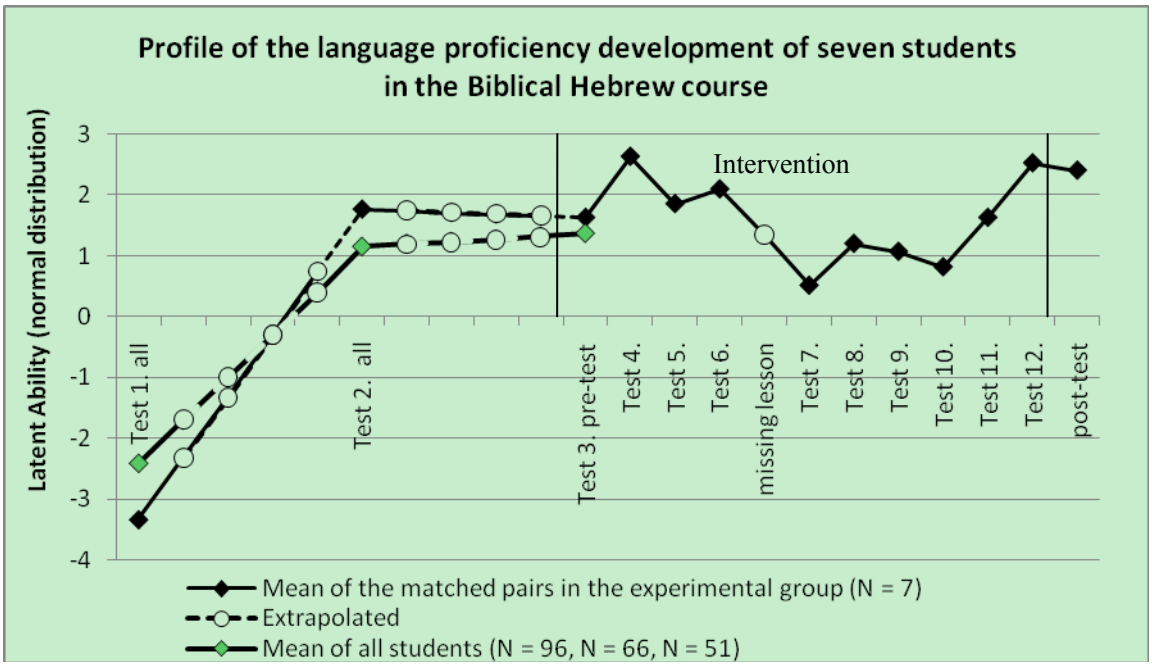

Figure 4. Change in student proficiency level: longitudinal profile in the experimental group

Figure 4 shows that the main average shift in proficiency level occurs during the first three weeks after learning the Hebrew letters and after concentrating on the most frequent and simple words, basic concepts, and preliminary structures, such as nominal clauses. The mean of the latent ability for all students increases until the third test, and in the experimental group, it seems to increase until the first intervention test (Test 4.). After this peak, the latent ability declined gradually until the seventh test. It is important to remember that the latent ability measured with the equated tests is not a "sense of knowing", but one's true latent ability, which seems to decline after a certain period of studying. A simple explanation for this is that after the first period of studying the alphabet and elementary characteristics of Biblical Hebrew, the second period consists mainly of verb morphology and many new – and infrequent – words, and thus much more to remember than in the earlier period. The exhaustive number of new words, structures, verb morphologies, suffixes and so on, may confuse the students for some weeks. It is therefore not a surprise that the students forgot such familiar words, and structures, and verb morphologies learned in the first period and which were insufficiently practiced during the second period. Thus, the decline in ability level is understandable.

It is noteworthy that in the middle of the intervention, the trend starts to rise and in the EG, it rises higher than in the CG. The repeated testing sessions seem to have helped the EG to regain the peak level– and to surpass it. If the learning profile in the CG was of a similar pattern to that in the EG, it would seem that *without* the repeated testing, and thus with lack of practice in the structures and verb morphology in novel situations, the CG remained at the lower proficiency level than in the pre-test – though most probably with an increasing profile, as the EG. However, most probably the increasing was less steep in comparison with EG.

On the basis of Figure 4, one can also speculate about the role of one canceled lesson between tests 6 and 7 – a sudden illness of the teacher – in the decline in proficiency level. This value in Figure 4 is extrapolated on the basis of an obvious declining trend for all 13 students in the EG. Could it be that without the canceled lesson, the ability level would be higher than it now was? The question can also be put as follows: how rapidly the thin mastery of the Biblical Hebrew language can dissipate when it is not actively reinforced? These questions remain open. However, on the basis of the data, in the middle of the process of language learning, one's ability level can, in some cases, sink within *two* weeks from a standard point of +1.4 to –2.1 (Fig. 5; see also Fig. 2), and still rise up again. Such a decrease in ability level is quite challenging to model mathematically; in practice, two independent models appear to explain the profile. Only an interview of students with such profiles may shed light on possible causes for the curves. The case in Figure 5 is not the only one in the data; of course, these are extreme cases. The trend, however, is obvious to most students.

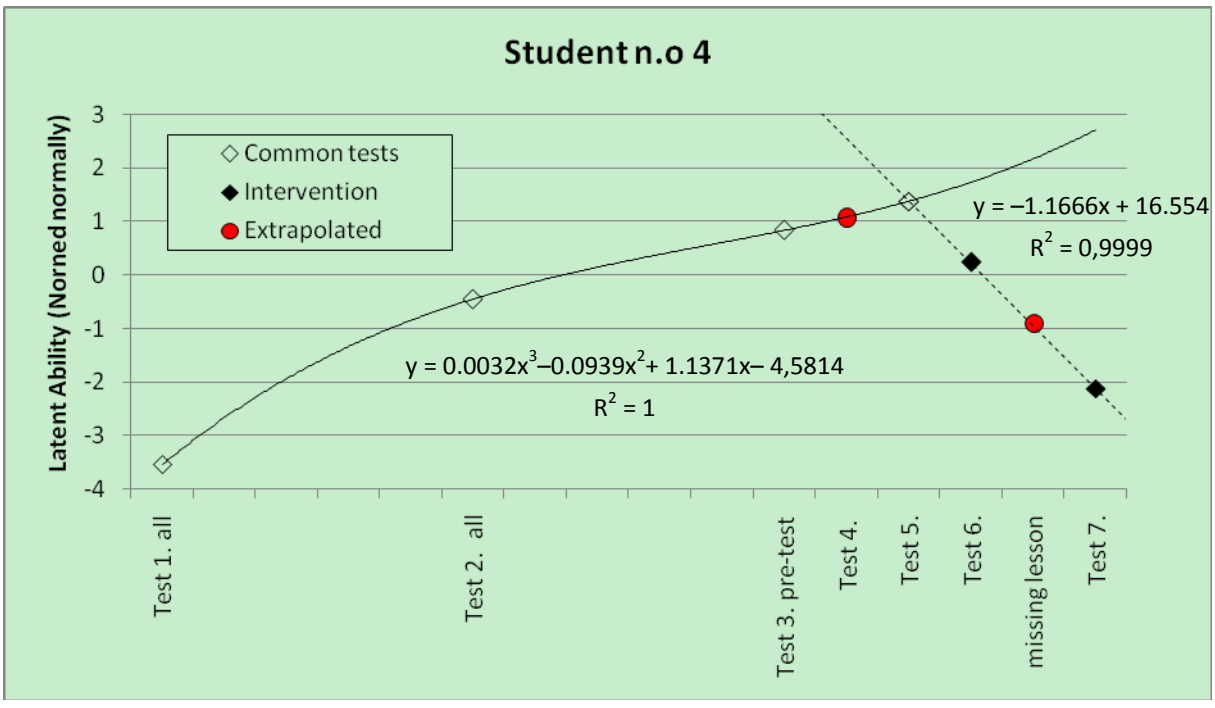

Figure 5. Change in proficiency level: mathematical modeling of proficiency of a single student

6. Discussion

6.1 Main Results, Their Meaning and Limitations

On the basis of the experimental design to measure student proficiency in Biblical Hebrew with static tests, and with no connection to speaking the language in everyday life, repeated testing sessions seem to have caused the students to improve their language proficiency more than without the repeated testing sessions. Although the groups were small, the effect was high ($d > 0.90$ or $d > 1.0$). The result is in line with earlier results where the final test results are enhanced with repeated testing sessions. An important result is that the learning curve in the data does not steadily increase but merely traces a nonlinear, wide U or J shape after the initial phase of the studies.

Lasry, Levy and Tremblay's (2008) interpretation to Karpicke and Roediker's (2008) result – that repeated testing may lead to multiple traces in one's memory, which facilitate recall – may be correct from the viewpoint of cognitive psychology. From this viewpoint, repeated testing sessions enhance semantic memory (Tulving 1983) and that repetition of the subjects learned enhances active recall (Tulving & Schacter, 1990). However, a more constructive psychological narrative story – in the Brunerian sense (see Bruner, 1986, 11, 50; 1990, 710; 1996, 39, 130) – of the development of language proficiency and the mastery of a language envisions the metaphor of an eagle taking to the air. The longitudinal profile of the development of latent ability in one's experimental group hints that there may be three mechanisms for taking to the air and for raising the ability level. First, the learning curve in the control group may sink deeper due to the lack of intensive testing thus resulting in fewer multiple traces in one's memory. This lack would push students deeper into confusion about new structures, perplexing prefixes and suffixes, similar-looking new words, and lists of verb morphologies. The intensive testing sessions, however, promise to help students keep their ability level higher than that of their fellow students, thus enabling them to take flight sooner. Second, the learning curve in the experimental group may increase flight much more steeply than that of the control group. This would mean that in their depths, both groups are in the same confusing mess of a new language and its idio syncrasies, but students with repeated testing sessions take the flight into the open air much faster than those do without the testing sessions. Also, the third possibility – combining the two previous ones –is possible: the testing sessions prevent the students from falling to the depths of their deepest valley and even helps them to rise steeply into more open air. The decision not to test the control group allows little more than speculation when analyzing the underlying mechanisms of the phenomenon.

The study carries some obvious limitations. One is that the original experimental group is small and thus even one outlier in the data may have caused radical change in the output. Another limitation is that the study is conducted in a real-life situation and hence it is not as rigorous as the laboratory experiment would have been. Third, naturally the results may be limited only to the subject of Biblical Hebrew. Yet one relevant challenge in the generalization is the random missing values in the data.

6.2 Implications for Practice and Further Possibilities

The result of better performance due to repeated testing gives support to a bulk of previous literature on the testing effect (see Bangert-Drowns, Kulik & Kulik, 1991; Cranney *et al.,* 2009; Johnson & Mayer, 2009; Leeming, 2002; McDaniel *et al.,* 2007; Vojdanoska, Cranney & Newell, 2009; see also Glover 1989). Although the routine of repeated testing cannot be the philosopher's stone which promises to turn the dust into gold, it may nevertheless enrich the arsenal of the wise teacher willing to raise the standard in the classroom. For the teacher of secondary language, the longitudinal profile of the learning and the profiles of individual cases with the steep decline of the language proficiency may enhance the intuitively clear fact that language learning is a process – a long one. Secondary language learning takes time and efforts – and without a continuous fostering one loses the superficial new language surprisingly effortless. It can vanish in two weeks.

This experiment was administered with a small number of students in one teacher's Biblical Hebrew class. The results, however, are promising for a larger scale testing. The results, despite being a curiosity in the ordinary language teaching society, are potentially of considerable interest from an international viewpoint because Biblical Hebrew is studied all over the world by students in the faculties of theology learning the basic skills needed to understand the original language of the Hebrew Bible. Thus, the results may be widely applicable not only in Biblical Hebrew, but also in Biblical Greek and Biblical Latin courses. Obviously, the previous results support the idea that the repeated testing would enhance the learning in several school subjects.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in

long-term memory. *J Exp Psych: Learn Mem Cogn, 20*(5), 1063-1087. http://dx.doi.org/10.1037/0278-7393.20.5.1063

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. http://dx.doi.org/10.1016/S1364-6613(00)01538-2

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85,* 89–99.

Béguin, A. (2000). *Robustness of Equating High-Stake Tests*. Enschede: Febodruk B.V.

Birnbaum, A. (1968). Estimation of ability. In F. M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison–Wesley Publishing Company.

Bruner, J. S. (1986). *Actual Minds, Possible Worlds*. Cambridge, Massachusetts: Harvard University Press.

Bruner, J. S. (1990a). Culture and Human Development: A New Look. *Human Development, 33*(6), 344–355. http://dx.doi.org/10.1159/000276535

Bruner, J. S. (1996). *The Culture of Education*. Cambridge, London: Harvard University Press.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *J Exp Psychol Appl, 13*(4), 273–281. http://dx.doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *J Exp Psychol Learn Mem Cogn., 34*(4), 918–928. http://dx.doi.org/10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. 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. *J Exp Psychol Learn Mem Cogn.*, *35*(6), 1563–1569. http://dx.doi.org/10.1037/a0017021

Carroll, D. (1994). *Psychology of Language* (2nd ed). Brooks/Cole.

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. http://dx.doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K., & McDermott, K B. (2007). The Testing Effect in Recognition Memory: A Dual Process Account. *J Exp Psychol Learn Mem Cogn.*, *33*(2), 431–437. http://dx.doi.org/10.1037/0278-7393.33.2.431

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially non tested material can benefit from prior testing of related material. *J Exp Psych: General*, *135*, 553–571. http://dx.doi.org/10.1037/0096-3445.135.4.553

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychol. Bull & Review*, *12*, 769–786. http://dx.doi.org/10.3758/BF03196772

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working Memory Capacity and Its Relation to General Intelligence. *Trends Cogn. Sci., 7,* 547–552. http://dx.doi.org/10.1016/j.tics.2003.10.005

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology, 21*(6), 919–940. http://dx.doi.org/10.1080/09541440802413505

Cribbie, R. A., & Jamieson, J. (2004). Decreases in Posttest Variance and the measurement of Change. *Methods of Psychological Research Online*, *9*(1), 37–55. http://dx.doi.org/10.1177/0146621608329889

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Dir. in Psychol. Sci*., *11*, 19–23. http://dx.doi.org/10.1111/1467-8721.00160

Eysenck, M. W., & Keane, M. T. (2005). *Cognitive Psychology: A Student's Handbook* (5th ed). Psychology Press.

Gates, A. I. (1917). *Recitation as a factor in memorizing*. *Archives of Psychology, 6*(40). New York: The Science press. Retrieved from http://www.archive.org/details/recitationasfact00gaterich.

Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly forgotten. *Journal of Educational*

Psychology, 81, 329–399. Retrieved from http://psycnet.apa.org/doi/10.1037/0022-0663.81.3.392

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124,* 75–111. http://dx.doi.org/10.1037/0033-2909.124.1.75

Gulliksen, H. (1987). *Theory of Mental Tests*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Gurung, R. A. R., & Daniel, D. (2006). Evidence Based pedagogy. Do text-based pedagogical features enhance students learning? In D. S. Dunn & S. L. Chew (Eds), *Best practices for teaching introduction to psychology* (pp. 41–55). Mahwah, N.J.: Erlbaum.

Hambleton, R. K. (1982). *Item response theory: The three-parameter logistic model.* Centre for the Study of Evaluation Report No. 220. LA: University of California.

Hambleton, R. K. (1993). Principles and selected Applications of Item Response Theory. In R. N. Linn (Ed.), *Educational Measurement* (3rd ed). American Council of Education. Series of Higher Education. Oryx Press.

Harley, T. (2001). *The Psychology of Language: From Data to Theory* (2nd ed). Psychology Press. http://dx.doi.org/10.4324/9780203345979

Jay, T. (2003). *The Psychology of Language.* New York: Prentice-Hall.

Johnson, C. I., & Mayer, R. E. (2009). A Testing Effect with Multimedia Learning. *Journal of Educational Psychology*, *101*(3), 621–629. http://dx.doi.org/10.1037/a0015183

Kaftandjieva, F. (2004). Standard Setting. In S. Takala (Ed.), *Manual for relating Language Examinations to the Common European Framefork of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Language Policy Division, Strasbourg. Reference Supplement. Section B. Retrieved from http://www.coe.int/T/DG4/linguistic/CEF-refsupp-SectionB.pdf

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. http://dx.doi.org/10.1080/09541440601056620

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *J. Exp. Psychol Gen., 138*(4), 469–486. http://dx.doi.org/10.1037/a0017341

Karpicke, J. D., Butler A. C., & Roediger, H. L. (2009). Metacognitive trategies in students learning: Do students practise retrieval when they study on their own. *Memory, 17*(4), 471–479. http://dx.doi.org/10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promote short-term retention, but equally spaced retrieval enhances long-term retrieval. *J Exp Psychol Learn Mem Cogn., 33*(4), 704–719. http://dx.doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*, 966–968. http://dx.doi.org/10.1126/science.1152408

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Mem Cognit., 38*(1), 116–124. http://dx.doi.org/10.3758/MC.38.1.116

Kester, L., & Tabbers, H. (2008). The Effect of Intervening Tests on Text Retention. In J. Zumbach, N. Schwartz, T. Seufert & L.Kester, (Eds.), *Beyond Knowledge: The Legacy of Competence* (pp. 183–187). Springer Netherlands. http://dx.doi.org/10.1007/978-1-4020-8827-8_25

Lasry, N., Levy, E., & Tremblay, J. (2008). Making Memories, Again. *Science, 320*. http://dx.doi.org/10.1126/science.320.5884.1720a

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212. Retrieved from http://psycnet.apa.org/doi/10.1207/S15328023TOP2903_06

Lord, F. M. (1952). *A theory of test scores.* Psychological Monographs. Psychometric society.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Mass: Addison-Wesley Publishing Company.

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.

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

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

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*, 40–48. http://dx.doi.org/10.1037/0021-843X.110.1.40

Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, *11*(2), 364–386. http://dx.doi.org/10.1177/1094428106291059

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century Crofts.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

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. 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. *Psychol Sci, 17*(3), 249–255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656. http://dx.doi.org/10.1037/h0063404

Sternberg, R. J., & Grigorenko, E. L. (2001). All testing is dynamic testing. *Issues in Education, 7*(2), 137–170.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing*. New York: Cambridge University Press.

Stout, W. (2002). Psychometrics: From Practice to Theory and back. 15 Years of Nonparametric Multidimensional IRT, DIF/Test Equity, and Skills Diagnostic Assessment. *Psychometrika, 67*(4), 485–518. http://dx.doi.org/10.1007/BF02295128

Takala, S. (2009). *Manual for relating Language Examinations to the Common European Framefork of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Language Policy Division, Strasbourg. Retrieved from http://www.coe.int/T/DG4/linguistic/Source/Manual%20Revision%20-%20proofread%20-%20FINAL.doc

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6,* 175–184. http://dx.doi.org/10.1016/S0022-5371(67)80092-6

Tulving, E. (1983). *Elements of Episodic Memory*. New York: Oxford University Press.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247,* 301–306. http://dx.doi.org/10.1126/science.2296719

Verhelst, N. D. (2004). Item Response Theory. In S. Takala (Ed.), *Manual for relating Language Examinations to the Common European Framefork of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.* Language Policy Division, Strasbourg. Reference Supplement. Section G. Retrieved from http://www.coe.int/T/DG4/linguistic//CEF-ref-supp-SectionG.pdf

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One Parametric Logistic Model OPLM.* Arnhem: CITO.

Vojdanoska, M., Cranney, J., & Newell, B. R. (2009). The Testing Effect: The role of Feedback and Collaboration in a Tertiary Classroom Setting. *Applied Cognitive Psychology, 24*(8), 1183–1195. http://dx.doi.org/10.1002/acp.1630

Whitney, P. (1998). *The Psychology of Language*. Houghton Mifflin.

Wright, B. D. (1968). Sample free calibration of items. In B. S. Bloom (Chair), *Invitational Conference on Testing Problems* (pp. 84–101). Princeton NJ: Educational Testing Service.