Journal of Educational and Developmental Psychology; Vol. 3, No. 2; 2013
ISSN 1927-0526 E-ISSN 1927-0534
Published by Canadian Center of Science and Education

Effect of Repeated Testing to the Development of Vocabulary, Nominal Structures and Verbal Morphology

Jari Metsämuuronen[1] & Markus Mattsson[1]

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

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

Received: June 16, 2013 Accepted: July 24, 2013 Online Published: September 24, 2013
doi:10.5539/jedp.v3n2p89 URL: http://dx.doi.org/10.5539/jedp.v3n2p89

Abstract

The repeating testing has shown to increase the general proficiency level of the students. Metsämuuronen (2013) showed with an experimental study that the overall achievement level in a secondary language enhanced statistically significantly whit repeated testing design. Previously, Tuvling (1967) and Karpicke & Roediger (2008) showed with a laboratory experiment that remembering the material studied is the most efficient with repeated testing sessions rather than with repeated studying sessions. An explanation for this, given by Lasry, Levy and Tremblay (2008), is that the repeated testing leads to multiple traces to the memory, which optimizes recall. This study concentrates on the increase in the proficiency level in Vocabulary, Nominal structures and Verbal Morphology after the set of exhaustive testing sessions. It also reviles change in the proficiency levels of the students in these areas during the study process. The experimental group gained more than the control group in all areas though the difference is statistically significant only in the content of Vocabulary. The effect sizes are high (Cohen's $d > 1.0$). In all areas of interest, the learning curve was of wide U-shape after the elementary period of studies.

Keywords: language testing, L2; IRT modeling, Experimental Study, matched-pairs, longitudinal research

1. Introduction

Almost 60 years, from Miller (1956), Broadbent (1958) and Neisser (1967), the computational principles of human brains have interested the scientists. Intense research efforts have been targeted at revealing the organization of human memory systems (see Squire, 2009; Conway *et al.,* 2005; Conway *et al.,* 2003; Engle, 2002; Poldrack & Packard, 2003; Baddeley; 2000) and development of language (see Jay, 2003; Harley, 2001; Whitney, 1998; Carrol, 1994).

According to the basic theories of human mind (see Squire, 2009), the human long-term memory comes in two flavours: declarative and procedural (or non-declarative) memory. As the name suggests, declarative memory refers to things and facts that can be declared and explicitly stated by brought them to mind whereas the contents of the procedural memory cannot be put into words; this includes the motor- and cognitive skills and habits (Squire, 2009; Ullman, 2004; Poldrack & Packard, 2002). The declarative memory can be divided into semantic and episodic (or narrative) memory (see Bruner, 1986; 1990; 1996; Tulving, 1983). The semantic memory is thought to be independent of the personal history and the identity of the person whereas the episodic memory consists of a store of personal actions, memories, and happenings (Tulving, 1983). Further, an influential view claims that there are separate subsystems for phonological and visual-spatial short-term memory and that these form a part of the working memory system, which is needed for selecting, from the environment and long-term memory, information appropriate for fulfilling the person's current goal (Baddeley, 1998; for the theories of working memory, see Miyake & Shah, 1999).

Cognitive psychologists have found several ways that the nature of memory encoding affects the later retention of memories. It has been noted that if the to-be-learned material is connected with previous knowledge of the topic and elaborated with imagery and stories rather than processed in a superficial way (for instance, concentrating on the letters of the words to be remembered), it will be retained better (Levels-of-processing view, Lockhart & Craik, 1990; Craik & Lockhart, 1972). These effects are, however, mediated by the sameness of the type of processing and encoding – practically speaking, if an association between words is encoded by rhyming the words, recall will be better following the rhyming task than a semantic task – and vice versa (transfer appropriate processing, Morris, Bransford & Franks*,* 1977). It has been shown that, at the time of memory encoding, the context, that is, surroundings (Godden & Baddeley, 1975), physiological state (Eich, 1980), and mood (Eich & Metcalfe, 1989) affect the later recall.

In spite of all these advances in memory research, a tacit assumption shared by all the views remains: learning is something that happens during the encoding phase while the tests at the recall phase are a passive way of probing what was learned earlier. Indeed, a basic doctrine of human learning and memory research is that suitably spaced repetition of material improves its retention (Cepeda *et al.*, 2006). A quite interesting set of experiments of language learning (see the original study of Tulving, 1967 and a later replication of Roediger & Karpicke, 2006a; 2006b; see also Karpicke & Roediger, 2008) showed with a laboratory experiment that remembering the material studied is the most efficient with repeated *testing* sessions rather than with repeated *studying* sessions. An explanation for this, given by Lasry, Levy and Tremblay (2008), is that the repeated testing leads to multiple traces to the memory, which optimizes recall. According to Lasry and colleagues their interpretation may lead to frequent in-class assessments in pedagogies such as Peer Instruction. While the idea is intriguing, it may be based on over interpreting the results underlying the multiple trace theory (Moscovitch & Nadel, 1998). Moscovitch & Nadel (1998) ground their theory in sound neuroscientific research but it should be noted that the results concern autobiographical (that is, episodic) memory and it is not at all clear whether they generalize to other forms of declarative memory.

2. Testing Effect in Literature

Metsämuuronen (2013) has intensively reviewed the literature of testing effect. The discussion is condensed here. In Section 2.1, the literature of testing effect related to laboratory settings is reviewed. In Section 2.2, the literature of testing effect related to classroom settings is focused.

2.1 Testing Effect Related to the Laboratory Settings

The phenomenon of improving performance by taking a test, that is, the testing effect, was studied already at the beginning of the 20th Century by Gates (1917) and Spitzer (1939). Since these pioneering studies, many laboratories have conducted experiments concerning the effect of testing. The basic tenet in the field was that learning occurs the most efficient way by using intensive study sessions. However, Tulving (1967) showed something radically new: the proportion of recalled words and the learning curves in different test groups were identical although the study group with repeated studying had studied six times more than the group of repeated testing. Later, Karpicke and Roediger (Roediger & Karpicke, 2006a; 2006b; Karpicke & Roediger, 2008) replicates the Tulving's design and noted that the group with repeated testing recalled the words better than the other groups. Thus, they inferred that repeated testing optimized the retrieval from the memory. The latter result radically boosted the research on the topic (see Karpicke & Roediger 2010; Karpicke, Butler & Roediger, 2009; Karpicke, 2009; Chan, 2009; Carpenter, 2009; Kester & Tabbers, 2008; Chan & McDermott, 2007; Kang, McDermott & Roediger, 2007; Karpicke & Roediger, 2007; Chan, McDermott & Roediger, 2006). It has been showed that when the repeated tests are taken equally spaced, the long-term retention is promoted more than by using the gradually increasing spacing (Karpicke and Roediker, 2007; 2010; see opposite in Landauer & Bjork, 1978). Sometimes the repeated testing can improve the later recall of even the non-tested material (Chan, McDermott & Roediger, 2006; Chan, 2009; see opposite in Anderson, Bjork, & Bjork, 1994).

2.2 Testing Effect Related to the Classroom Settings

Roediger and Karpicke (2006a) pointed out the challenge in the testing in the classroom settings: because of non-controlled motivation to learn, interest in course material, or amount of studying, it is not easy to conduct the inferences of the datasets. Nevertheless, many studies in various university courses have found a positive connection between the testing and external test results (see Metsämuuronen, 2013; Vojdanoska, Cranney & Newell, 2009; Cranney *et al.,* 2009; Johnson & Mayer, 2009; McDaniel *et al.,* 2007; Leeming, 2002). Bangert-Drowns, Kulik, and Kulik (1991) found, in their meta-analysis, that 83% of the studies showed a positive effect of frequent testing. More, the higher the number of tests the higher was the more difference between the groups. Gurung and Daniel (2006) showed that the supervised tests were related to better examination performance. Although the research literature on the testing effect is convincing, McDaniel and his colleagues (2007, see also Glover, 1989) lamented that the literature has been virtually ignored by the educational community.

2.3 Static and Dynamic Tests

Testing procedures are usually divided into two: static- and dynamic testing (see Sternberg & Grigorenko, 2001; 2002; Grigorenko & Sternberg, 1998). In the static tests, the test-taker is not provided feedback about the performance in the test. This strategy is used when the correct answers of the test items are important to kept unknown – like in IQ-tests or SAT type of tests. It is noteworthy that this procedure makes it possible to study pure testing effect. In the procedure of the dynamic tests, the test-takers are given the feedback so that they can make improvement in their latter test score. The procedure of the dynamic test is more usable when willing to teach the topic through testing; the incorrect answers are corrected, and the learning potential of the test-taker can be reviled. The test results in the final test are, naturally, better with dynamic testing than with static testing (see Vojdanoska, Cranney & Newell, 2009; Metcalfe, Kornell & Finn, 2009; Butler & Roediger, 2008; Butler, Karpicke & Roediger, 2008; 2007). It is noteworthy that this procedure does not make it possible to study pure testing effect.

2.4 Views to the Second Language (L2) Acquisition from the Cognitive Psychology Viewpoint

Learning a language is, in itself, a multifaceted phenomenon. The learner needs to acquire, on the one hand, new mappings between sound and meaning and, on the other hand, rules that govern combining these mappings. Ullman (2001) proposes a model according to which the sound-meaning mappings (mental lexicon) are stored in declarative memory while rules operating on this material are grounded in procedural memory. This model is called the declarative/procedural model of language processing.

While the neural basis of second language (L2) acquisition is too broad a topic to be covered here with much detail, let it be mentioned that the declarative/procedural model can be used as a tool in understanding the relevant phenomena (Ullman 2005). Several testable hypotheses concerning second language acquisition can be derived from the model. For instance, learning L2 grammar or structures as an adult may be more difficult than the first language (L1) structures because the procedural system develops through certain critical periods that have already been passed. This forces the learner to initially store complex structural forms in declarative memory, while being later able to use procedural memory for grammatical processing. Even if learning the grammatical rules in L2 often proceeds rather slowly, it has been demonstrated that people learn the syntax of an artificial language rather quickly and that the brain signatures of artificial language syntactic violations closely resemble those of a natural language (Friederici, Steinhauer & Pfeifer, 2002).

Another interesting, while similarly speculative, an option for interpreting the results of Karpicke & Roediger (2008) is to consider processes involved in modifying motor engrams and in turning a previously consolidated memory back into a labile state requiring later reconsolidation (Walker *et al.*, 2003). In the study, the subjects learned the series of finger tappings, governed by a simple “grammar”. When a previously learned series was brought back to mind immediately prior to learning a new series, motor knowledge concerning the first series was seriously impaired when tested the following day. The authors propose an adaptive function for the phenomenon enabling the fine-tuning of previously learned motor sequences. They propose that “*similar mechanisms may also contribute to the integration of episodic memories and the revision of semantic knowledge based on newly acquired information*” (Walker *et al.*, 2003, 619). What makes this observation especially interesting is the fact that the neural basis of memory consolidation and reconsolidation has been intensely investigated (see, for instance, McCaugh, 2000).

Results such as these may have a role in grounding educational practices on the foundation provided by basic neuroscientific research. It may not matter which of these proposed interpretations for the results of Karpicke & Roediger (2008) is the correct one. The practical conclusion shared by all seems to be that we need to re-evaluate what would be the most effective ways to learn languages.

3. Aim of the Study

The main aim of the study is to reanalyse the dataset of Metsämuuronen (2013) from the viewpoint of three content areas of L2 learning: Vocabulary and Words (or, shorter, ‘Vocabulary’), Nominal structures, and Verbal morphology and to reveal which of the areas benefit most of the repeated testing. Another aim is to revile the profiles of proficiency levels of the students in Vocabulary, Nominal structures, and Verbal Morphology during the study process.

4. Methods

The same data is used here as in Metsämuuronen (2013). Hence, the same methodological choices are made in order to produce comparative results.

4.1 Sample and Drop-Out

Altogether 30 students of Biblical Hebrew in Helsinki University participated in the experiment at the second phase of their language studies. The students were randomized into two matched-pairs groups (n = 15+15) on the basis of their attitudes and proficiency levels in the pre-test. During the experiment of six weeks, some students dropped from the experiment: two from the experiment group (EG) and eight from the control group (CG), because it was not possible to force the students to continue their studies. Except two drop-outs, the students in the EG were motivated of the testing process. Hence, when they were not able to come to the lesson they took the test on their own time or – in some cases – during the next study session. In CG, in contrast, there were several drop-outs of which the most, unfortunately, came either from the lowest- or the highest extreme of the proficiency scale. Thus, the remaining part of the CG (n = 7) were mainly in the middle range of the proficiency scale. Hence, seven matched pairs are reported.

4.2 Design and Hypotheses

The study design follows the classical procedure of pre-post-test design. An additional feature was two pre-tests during the first phase of the studies: one before any studies and one at the middle of the first period of studies (after three weeks). Another character of the study was the longitudinal approach to the learning and testing; the students were tested at every lesson.

Both the CG and EG had the lessons the same way. The EG was tested with a ten-minute-test in the mid of each three-hour study session while the CG studied the course book. The testing was of a static type: no feedback was given to the students. Because of convincing previous results (see Metsämuuronen, 2013; Vojdanoska, Cranney & Newell, 2009; Johnson & Mayer, 2009; Cranney *et al.,* 2009; Karpicke & Roediger, 2008; McDaniel *et al.,* 2007; Roediger & Karpicke, 2006a; 2006b; Leeming, 2002) the alternative hypothesis is kept one-sided: *the gain score in the EG is higher than in the CG.*

4.3 Replacing the Missing Values

In the long sequence of test scores, 12 missing values were replaced by using either linear- or non-linear modelling (see Fig. 3, for example). In most cases, the missing values were usually easy to model as the mean score for two tests (the one immediately before and after the missing test score).

4.4 Items and Tests

Altogether 218 items were used in the item calibration and test equation (see Section 4.5). The items covered the recognition of the transliterated Hebrew words and Hebrew letters (the elementary basics of the language) to the Verb's morphology (see in detail Metsämuuronen, 2013, Table 3). The tests were constructed so that they were, practically speaking, in an order of increasing difficulty level. During the intervention, the number of items ranged from 16 to 32, reliabilities of the test scores ranged from 0.79 to 0.94, and the item-total correlations ranged from 0.43 to 0.60.

4.5 Linking of the Tests, IRT Modelling and Equating

All the tests were linked with each other by a set of linking items from the previous tests. The test scores were equated by using Item Response Theory (IRT) modelling (Rasch, 1960; Lord & Novick, 1968; Birnbaum 1968; Lord, 1980; Hambleton, 1982; 1993; of equating, see Béguin, 2000). IRT modelling is widely used in the large scale student assessment (such as in Trends in Mathematics and Science Study, TIMSS and Programme of International Student Assessment, PISA) and especially in the settings of language testing (see Verhelst, 2004; Kaftandjieva, 2004; Takala, 2009). Rasch modelling (Rasch, 1960) with OPLM software (Verhelst, Glas & Verstralen, 1995) was used in estimation. By using the IRT modeling, one estimates the latent ability of each student; the latent ability is symbolized by the Greek Theta (θ) and it follows the Standardized Normal distribution ranging usually from -4 to +4. An average student gains $\theta = 0$ and the lower the proficiency level the lower below zero the value of Theta is. Resulting from the procedure, the scores in each test are in the same scale and, hence, they are comparable.

4.6 Analysis Methods

There are two standard approaches to analyzing pre-post-test design: the procedure of Analysis of Covariance (ANCOVA) usually used with randomized experiments (see Miller & Chapman, 2001; Cribbie & Jamieson, 2004) and the procedure of the Analysis of Variance (ANOVA) or *t*-test to analyze the gain score. Because of the reduced variance in CG, ANOVA approach (or practically, the *t*-test) was selected. However, because of small sample size, the main analysis tool was the non-parametric alternative for t-test, Mann-Whitney U test. The effect size was calculated two ways: primarily, Cohen's *d* on the basis of t-value and the *d* for experimental studies (see Morris, 2008) as comparison to Cohen's *d*:

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

where *c* refers to the CG, *e* refers to the EG, *post* refers to the post-test, *pre* refers to the pre-test, and $\bar{x}$ and σ refer to the mean and standard deviation in the groups.

5. Results

5.1 Differences in Gain Scores

During the intervention, in all areas of interest in the study, EG gained notably more than CG (Fig. 1). The gain score is the highest in the sub-area of Words & Vocabulary (Mann-Whitney U: *p* = 0.064; $t_{(12)} = 1.93$; $p_{(one\text{-}tailed)} = 0.037$; *d* = 1.01 or *d* = 1.12; Tables 1 and 2). On the basis of Eta Squared ($\eta^2 = 0.24$), the experiment explains 24% of the difference in Words & Vocabulary, which is quite a high value. EG gains 0.89 standard units while, with the same lessons and with the same teacher though without the continuous testing, CG "gained" -0.54 standard units. The latter means that the proficiency level in CG got, paradoxically, *lower* during the experiment. The reasons are discussed in what follows. As known also from Metsämuuronen (2013), the language proficiency level as a whole was statistically significantly higher in the EG than in the CG (Mann-Whitney U: *p* = 0.037, $t_{(12)} = 1.64$; $p_{(one\text{-}tailed)} = 0.064$; *d* = 0.95 or *d* = 1.11).

It may be worth noting that the difference between the groups is not statistically significant in the sub-areas of Nominal structures and Verbal morphology. Thus, it seems that repeated testing has its main effect enhancing the retention of learning L2 vocabulary, but not necessarily of learning L2 structures. Naturally, with larger sample sizes the difference would have been statistically significant.

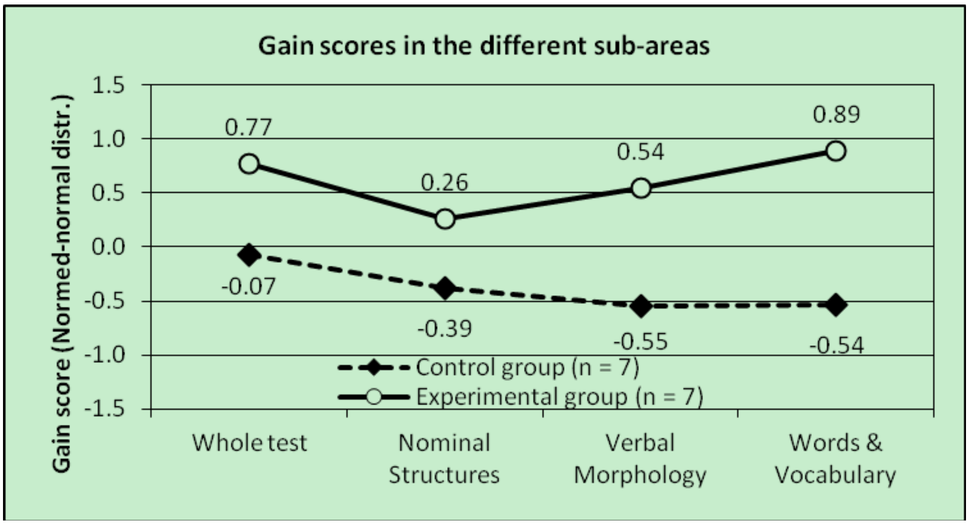

Figure 1. Change in the proficiency level – Gain scores in the different sub-areas

Table 1. Group statistics

			Whole test		Nominal Structures		Verbal Morphology		Words & Vocabulary	
	Group	N	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Pretest	CG	7	1.8	0.59	2.1	1.55	2.0	1.33	2.2	1.94
	EG	7	1.6	0.94	1.6	1.09	1.0	1.03	1.6	0.89
Posttest	CG	7	1.7	0.87	1.7	1.23	1.5	0.79	1.6	1.29
	EG	7	2.4	0.77	1.9	1.06	1.5	0.92	2.5	0.43

Table 2. Test statistics for difference in the gain scores

	Mann-Whitney U	Exact Sig. (1-tailed)	t(12)	Sig. (1-tailed)	d based on t	d based on exp. formula	Eta Squared
Whole test	10	0.037	1.64	0.064	0.95	1.11	0.18
Nominal Structures	18	n.s.	0.79	n.s.	0.46	0.49	0.05
Verbal Morphology	15	n.s.	1.20	n.s.	0.69	0.93	0.11
Words & Vocabulary	12	0.064	1.93	0.039	1.12	1.01	0.24

5.2 Longitudinal Profile of the Development of Proficiency

The longitudinal change in the different subareas of language proficiency in the EG is shown in Figure 2. The graph may also explain the negative gaining in the proficiency in the CG – this is discussed below the graph. In Figure 2, the missing values in the profile at the beginning phase of the study are extrapolated in order to avoid the misleading time perspective between the preliminary tests (3 weeks) and the intervention tests (two tests per week). The curves of Nominal structures and Verbal morphology at the pre-test phase are estimated on the basis of the known curriculum in the language course. The curve of Words & Vocabulary followed mainly the curve of the Total mean because at the beginning phase the tests concentrated on this content area.

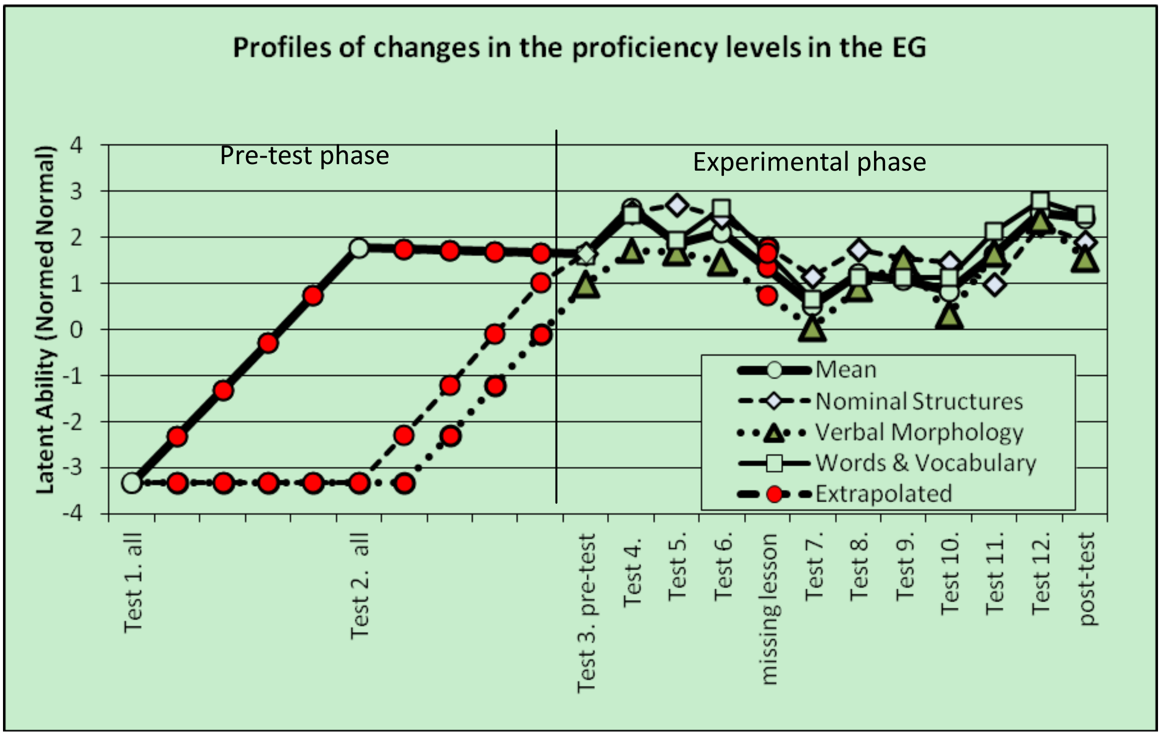

Figure 2. Change of the proficiency level – profiles of the means in the experimental group

Three points are raised from the longitudinal profiles of the students. First, Figure 2 demonstrates how the proficiency level raises dramatically during the first three weeks in the language course. During this time, the unfamiliar letters, the most frequent words, and basic structures of the language are learnt. During the next three weeks, the proficiency level does not raise much – if at all – though it stays at the level reached. The experiment started at the second period of the course. The peak point of the achievement seems to be at the beginning of the period (Tests 4. and 5.). After this, the proficiency level, in all content areas, declines mildly though gradually until 7th test. This can be explained by the fact that the experimental phase consisted of lots of verb morphology and, hence, lots of new words. Hence, there were more specifics to remember compared with the earlier period.

It is easy to understand that the structures, verb morphologies, suffixes, and words, confuses the students at this phase. It seems evident and understandable that the students forgot some of the words, structures and verb morphologies learnt in the first period. Hence, the decline in proficiency.

Second, the proficiency levels of all content areas start to incline in the middle of the intervention. In the EG it appears to rise higher than in the CG. It seems that the repeated testing sessions helped the EG to gain the peak level again – and to go further. Assumingly, the profiles of the proficiency in the CG followed similar patterns as they were in the EG. If so, it seems that *without* the intervention (repeated testing), and thus lack of drilling the Vocabulary, Nominal structures and Verb morphology, the CG gained negatively: the proficiency level, most probably, in the CG increased as it did in the EG but *not as steep* or as *fast*. It is somewhat interesting that, in all the tests, the proficiency level of Nominal structures is practically higher than Vocabulary or Verbal morphology. This is somewhat more interesting when knowing that in the course the Vocabulary is learnt longer than Nominal structures. However, at the end of the intervention, the proficiency level of Vocabulary supersedes both Nominal structures and Verbal morphology.

Third, on the basis of the U-shape curve of the learning, one can speculate of the decline of the proficiency level. On the basis of the data, it is obvious that in some cases, during the process of learning a language, the ability level can sink within *two* weeks from the standard point +2.8 to –4.0, that is, practically no correct answers in the test, (see Fig. 3) and still rise up again. It is worth noting that, in the case in Figure 3, the proficiency level in Verbal structures seems to sink the most and in the Nominal structures the least. This can be explained by the fact the Nominal structures were drilled through the previous lessons but the Verbal structures were just started when the experiment began. Hence the proficiency in Verbal structures was very thin in any case.

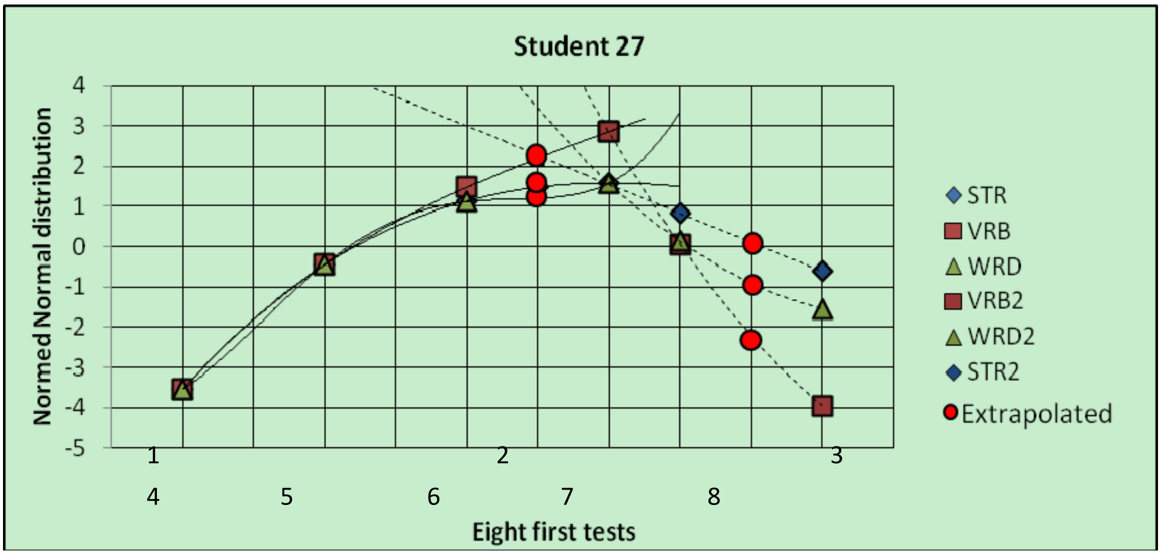

Figure 3. Change of the proficiency level – profiles of the scores of an example case

6. Discussion

On the basis of the results, it seems evident that the repeated testing sessions affected the students to raise their language proficiency level, in all areas studied, more than without the repeated testing sessions. Though the groups in the intervention are small, the effect sizes are high (in all areas, $d > 1.0$ or $d > 1.5$). The enhanced performance in the real life study settings, because of repeated testing, was supported by the experiment as suggested also by Metsämuuronen (2013), Vojdanoska, Cranney and Newell (2009), Cranney *et al.* (2009), Johnson & Mayer (2009), McDaniel *et al.* (2007), Leeming (2002), and Bangert-Drowns, Kulik and Kulik (1991). The routine of repeated testing may raise the proficiency level in L2 performance, especially the proficiency in words and vocabulary. This may be valuable information because without vocabulary it is difficult to think that there would be much reading, writing or oral skills. The obvious reason may be that the words and vocabulary are strictly related also to the structures of the language; whenever testing the structures of the language related vocabulary is used. Hence, the Vocabulary is repeated more than the structural matters within the testing process. Maybe, if the experiment was longer than six weeks, the results might have been seen also in other areas, too.

Another result, also interesting, is that the learning curves in all the areas in the study follow wide U shapes, after the very elementary phase of the studies. The U-shape is well-known, though not very widely discussed, phenomenon in the learning process. It is known in learning as general (e.g., Gershkoff-Stowe & Thelen, 2004), the development of intuition by level of expertise (e.g., Baylor 2001), in mathematics learning (McNeil 2007), arts learning and symbolization (e.g., Davis 1997a; 1997b; Haanstra, Damen & van Hoorn, 2011), and specifically, in the language learning from the cognition and cognitive science viewpoint (e.g., Plunkett & Marchman, 1991; 1993; Marcus, 1995; Taatgen & Anderson, 2002; Ramscar & Yarlet, 2007). The ultimate explanations for the phenomenon are not discussed here. However, it seems, on the basis of Metsämuuronen (2013) and the results of this study, that the repeated testing lowers the deepness of the U-curve.

As discussed by Metsämuuronen (2013), the longitudinal U-profile of the development of the latent abilities in the EG hints that the repeated testing may include three mechanisms in raising the proficiency level. First, it is possible that the learning curves in the control group sink deeper because of the lack of intensive testing and thus, of the lack of multiple traces to the memory (see Lasry, Levy and Tremblay, 2008; Karpicke & Roediger, 2008). This can mean that the tested students, confused with new prefixes and suffixes, similarly appearing new words, and endless lists of verb morphologies, *keep the level of ability higher* than their fellow students because of the intensive testing sessions. Another option is that both the groups sink similarly low, but the proficiencies in the experimental group *increase radically faster* than those of the control group. This means that, at the lowest point of the U-curve, both groups are in the same confuse of the specialties of the language, but the students with repeated testing sessions find their way up steeper than those students without the intensive testing sessions. The design does not allow the much deeper analysis of the mechanisms than speculations only. Third option is that both these work simultaneously: the students with intensive testing do not sink as low and they rise faster than their peer students.

Third results, confirming the results of Metsämuuronen (2013), is that for some cases, the thin mastery of a language may sink within two weeks to the starting level even after several weeks of practice. Especially steep seems to be the sinking in Verbal structures where, in some cases, the reduction of proficiency may be from +3 standard units to -4 standard units within two weeks. Practically speaking, some students were able to solve all the problems related to Verbal morphology at the beginning of the experiment and after two weeks they were not able to solve even one of those.

The study carries the same limitations as Metsämuuronen study (2013): the small study group where even one case has an impact on the output and the real life setting where the study behavior or motivation to learn, for example, cannot be controlled.

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. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. Retrieved from http://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(00)01538-2

Baddeley, A. D. (1998). Recent developments in working memory. *Current Opinion in Neurobiology*, *8*, 234-238. http://dx.doi.org/10.1016/S0959-4388(98)80145-1

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85,* 89–99. http://dx.doi.org/10.1080/00220671.1991.10702818

Baylor, A. L. (2001). A U-shaped model for the development of intuition by level of expertise. *New Ideas in Psychology, 19*(3), 237-244. http://dx.doi.org/10.1016/S0732-118X(01)00005-8

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 (Eds.), *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison–Wesley Publishing Company.

Broadbent, D. (1958). *Perception and Communication*. London: Pergamon. http://dx.doi.org/10.1037/10037-000

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

Bruner, J. S. (1990) 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.

Cepeda, N. J., Vul, E., Rohrer, D, Wixted, J. T., & Pashler, H. (2006). Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention. *Psychological Science*, *19*(11), 1095–1102. http://dx.doi.org/10.1111/j.1467-9280.2008.02209.x

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 nontested 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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal behavior, 11*(6), 671–684. http://dx.doi.org/10.1016%2FS0022-5371%2872%2980001-X

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

Davis, J. H. (1997). Drawing's Demise: U-Shaped Development in Graphic Symbolization. *Studies in Art Education, 38*(3), 137–152.

Davis, J. H. (1997). The What and the whether of the U: Cultural Implications of Understanding Development in Graphic Symbolization. *Human Development, 40,* 145–154. http://dx.doi.org/10.1159/000278717

Eich, J. E. (1980). The Cue-Dependent Nature of State-Dependent Retrieval. *Memory & Cognition, 8*(2), 157–173. http://dx.doi.org/10.3758/BF03213419

Eich, E., & Metcalfe, J. (1989). Mood Dependent Memory for Internal Versus External Events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*(3), 443–455. Retrieved from http://psycnet.apa.org/doi/10.1037/0278-7393.15.3.443

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

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing:

Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 529–534. http://dx.doi.org/10.1073/pnas.012611199

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

Gershkoff-Stowe, L., & Thelen, E. (2004). U-Shaped Changes in Behavior: A Dynamic Systems Perspective. *Journal of Cognition and Development, 5*(1), 11–36. http://dx.doi.org/10.1207/s15327647jcd0501_2

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 329–399. http://dx.doi.org/10.1037/0022-0663.81.3.392

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. http://dx.doi.org/10.1111/j.2044-8295.1975.tb01468.x

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

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.

Haanstra, F., Damen,M.-L. & Van Hoorn, M. (2011). The U-Shaped Curve in the Low Countries: A Replication Study. *Visual arts research, 37*(72), 16–29. http://dx.doi.org/10.5406/visuartsrese.37.1.0016

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

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Lasry, N., Levy, E., & Tremblay, J. (2008). Making Memories, Again. *Science, 320*, 1720. http://dx.doi.org/10.1126/science.1152408

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

Lockhart, R. S., & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology, 44*, 87–112. http://dx.doi.org/10.1037/h0084237

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.

Marcus, G. F. (1995). Children's overregularization of English plurals: A quantitative analysis. *Journal of Child Language*, *22*(2), 447-459. http://dx.doi.org/10.1017/S0305000900009879

McGaugh, J. L. (2000.) Memory: A century of consolidation. *Science, 287*, 248–251. http://dx.doi.org/10.1126/science.287.5451.248

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

McNeil, N. M. (2007). U-Shaped Development in Math: 7-Year-Olds Outperform 9-Year-Olds on. *Developmental Psychology, 43*(3), 687–695. http://dx.doi.org/10.1037/0012-1649.43.3.687

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

Metsämuuronen, J. (2013). Effect of repeated testing to the development of Biblical Hebrew language proficiency. *Journal of Educational and Psychological Development, 3*(1). http://dx.doi.org/10.5539/jedp.v3n1p10

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review, 63*, 81-97. http://dx.doi.org/10.1037/h0043158

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

Miyake, A., & Shah, P. (Eds.) (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139174909

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533. http://dx.doi.org/10.1016/S0022-5371(77)80016-9

Moscovitch, M., & Nadel, L. (1998) Consolidation and the hippocampal complex revisited: In defense of the multiple-trace model—discussion point. *Current Opinion in Neurobiology, 8*(2), 297–300. http://dx.doi.org/10.1016/S0959-4388(98)80155-4

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

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition, 38*(1), 43–102. http://dx.doi.org/10.1016/0010-0277(91)90022-V.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, *48*(1), 21–69. http://dx.doi.org/10.1016/0010-0277(93)90057-3

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251. http://dx.doi.org/10.1016/S0028-3932(02)00157-4

Ramscar, M., & Yarlett, D. (2007). Linguistic Self-Correction in the Absence of Feedback: A New Approach to the Logical Problem of Language Acquisition. *Cognitive Science*, *31*(6), 927-960. http://dx.doi.org/10.1080/03640210701703576

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. Retrieved from http://psycnet.apa.org/doi/10.1037/h0063404

Squire, L. R. (2009). Memory and brain systems: 1969–2009. *The Journal of Neuroscience*, *29*(41), 12711–12716. http://dx.doi.org/10.1523/JNEUROSCI.3575-09.2009

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

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "Broke"? A model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155. http://dx.doi.org/10.1016/S0010-0277(02)00176-2

Takala, S. (Ed.) (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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2296719

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*, 717–726. http://dx.doi.org/10.1038/35094573

Ullman, M. T. (2004) Contributions of memory circuits to language: The declarative/procedural model. *Cognition*. *92*, 231–270. http://dx.doi.org/10.1016/j.cognition.2003.10.008

Ullman, M. T. (2005). A Cognitive Neuroscience Perspective on Second Language Acquisition: The Declarative/Procedural Model. In C. Sanz (Ed.), *Mind and Context in Adult Second Language Acquisition: Methods, Theory, and Practice* (pp. 141-178). Washington, DC: Georgetown University Press.

Walker, M. P, Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature, 425*(6958), 616-620. http://dx.doi.org/10.1038/nature01930

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.

Copyrights