TEST-ENHANCED LEARNING, WORKING MEMORY, AND DIFFICULTY OF MATERIAL

Dennis Nordstrand

Master’s Thesis in Cognitive Science, 30 hp/ECTS
Master’s Programme in Cognitive Science, 120 hp/ECTS
Spring semester 2018
Supervisor: Patrik Hansson, PhD
Abstract

It is well established that repeated testing is more beneficial for durable learning than repeated studying of the same material, a phenomenon known as the testing effect. This study sought to investigate the role of working memory capacity (WMC) in relation to the learning process and the difficulty of the material to be learned when using a test-enhanced learning method. As between subject manipulation, participants \( n = 99, M = 25.62 \) years of age were divided into two groups, one using repeated studying and one using alternated testing and studying. A material of two difficulty levels, as well as immediate and delayed retention tests, was used in each condition as within subject manipulation. Further, an \( n \)-back task was used to measure WMC. Results from mixed model ANOVAs showed no significant impact of WMC on either the learning process or retention in relation to the difficulty of the material. The testing condition performed significantly higher than the studying condition on the retention tests. The testing effect is further cemented as a promising method for practical application in the educational sector regardless of both WMC and difficulty level.

Test-Enhanced Learning, Working Memory, and Difficulty of Material

The science of memory and learning as one of the prime fields of cognitive science has many proposed practical applications, but use within education seems undeniably to be one of the most promising and advantageous (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Gluck, Mercado and Myers, 2008; McDaniel, Roediger, & McDermott, 2007). Both students and teachers have an increasingly limited amount of time at their disposal to construct a substantial, durable base of knowledge (Dunlosky et al., 2013). While building durable knowledge is theoretically trivial, doing so efficiently within the time constraints of prevailing field conditions is decidedly non-trivial (McDaniel, Roediger, & McDermott, 2007; Dunlosky et al., 2013). As an example, repeated reading of a material with appropriate spacing between repetitions has long been known to enhance the rate of learning a material as a function of repetitions (Capeda, Vul, Rohrer, Wixted, & Pashler, 2008; Rothkopf, 1968). This approach is, however, difficult to implement formally due to its time-consuming nature as it would require a complex, intricate schedule of a multitude of topics over many repetitions. It has therefore been argued that the true challenge is to promote durable learning with efficiency as a primary priority (Rawson & Dunlosky, 2011).

Despite a long-standing tradition of the educational system using testing for assessment of already existent knowledge on a given topic, it is well known that the very act of testing can itself work as a highly effective method of learning (Dunlosky et al., 2013; Karpicke & Aue, 2015; Karpicke & Roediger, 2008; McDaniel et al., 2007; Roediger & Karpicke, 2006a, 2006b). In point of fact, a wealth of evidence shows that the mechanisms of information retrieval enhances the ability to later retrieve the same information again – indeed, to a greater degree than a mere repetition of exposure to stimuli (e.g. Dunlosky et al., 2013; Karlsson-Wirebring et al., 2015). Initially recognized as early as a century ago (e.g. Ebbinghaus, 2013), this phenomenon, commonly referred to as the testing effect, or test-enhanced learning, has seen a revival during the last decade with the attention of Karpicke and Roediger (2008; Roediger & Karpicke, 2006a, 2006b).

The testing effect is typically investigated by one of two means (see Figure 1). In the first, subjects are exposed to a to-be-learned material and subsequently divided into two subgroups. The first group will then receive repeated exposure to the material (a study condition), while the second group will be repeatedly tested on the material (a test condition). This is the between-group design variant. The other common method is to have one group of subjects repeatedly study one material and repeatedly be tested on another. This is the within-group variant. Both of these methods are usually followed by an initial test on the relevant material immediately after the study or test condition (i.e. the learning phase), after which learning is again gauged in a delayed test which is usually administered a number of days or weeks after the first session (e.g. Rowland, 2014; see Figure 1). The tests typically show a clear advantage in performance for the repeated testing conditions for within- as well as between-group comparisons, particularly for any delayed tests, indicating that learning is significantly more durable over time when the material is tested rather than restudied (Rowland, 2014).

A vast variety of materials, such as word pairs, functions, semantic facts, and text content have shown to be subject to the testing effect (e.g. Karpicke & Roediger, 2010; Carpenter, Pashler, & Cepeda, 2009; Kang, McDaniel, & Pashler et al., 2011; Karpicke, & Aue, 2015; Karpicke & Blunt, 2011; Pashler & Cepeda, 2009; Wiklund-Hörnqvist, Jonsson, & Nyberg, 2013). Further, test-enhanced learning has been compared and contrasted with many alternative methods and stands as one of the prime methods for learning (Dunlosky et al., 2013; Karpicke & Blunt, 2011; Rawson & Dunlosky, 2011; Stenlund, Jönsson, & Jonsson, 2016). Potent though test-enhanced learning may be in practice regarding materials
and circumstances, however, some critique has been levied against the field as to a lack of theoretical explanatory models (Rowland, 2014). In response, a number of models have been proposed as scrutiny has shifted toward understanding not only the power of the paradigm, practically, but also its underpinnings (Lehmann, Smith, & Karpicke, 2014).

![Figure 1. The typical design for a testing effect study. After the initial exposure, the subjects or material are divided into a testing or a study condition, where the material is repeatedly tested or studied respectively. After this learning phase one test is administered immediately, and one or more delayed.](image)

The desirable difficulty theory was first put forward by Bjork (1994) and posits that tasks that are more difficult will take longer to learn at first, but will be more beneficial in the long term, as long as the task is not too difficult. Similarly, the retrieval effort hypothesis states that encoding for durable learning is stronger the more effort is required to retrieve the item, but only once retrieval is successful (Pyc & Rawson, 2009). Support for this model had been found by manipulating the inter-stimulus interval, to investigate whether more spaced or clustered testing would cause learning to be more durable over time (Pyc & Rawson, 2009). It was shown that instances of successful retrieval led to significantly more durable learning if the interval was longer (Pyc & Rawson, 2009).

According to Bjork and Bjork’s (1992) theory of disuse, memory representation can be defined as two modes of memory strength: storage strength and retrieval strength. The strength of storage is the degree to which a memory maintains fidelity over time when not in use, and is the main determinant for long-term retention - the higher the storage strength of a representation after encoding, the higher the prospect of retrieval. Retrieval strength, meanwhile, is concerned with momentary access to stored information in the short term. This theory of disuse postulates that events of encoding as well as of retrieval have the facility to reinforce both storage and retrieval strength, however, that retrieval events are more potent in facilitating such reinforcement. Thus, by the continuum of memory strength in this proposed memory configuration, a stored memory can retain an extent of storage as well as retrieval strength, and yet not be successfully recalled (Bjork & Bjork, 1992). That is, memory strength is not binary (i.e. recalled or not recalled), but rather continuous. In light of this view, Kornell, Bjork, and Garcia (2011) introduced the bifurcation model of successful recollection. As noted, benefits from successful retrieval are greater when delay between instances of retrieval is longer (Pyc & Rawson, 2009). This is explained by the bifurcation model as successfully retrieved items possessing sufficient strength to exceed a certain strength threshold needed for recollection, whereas items not successfully retrieved fall under this boundary, even though they may still possess variable levels of strength (Kornell et al., 2011). When a test with no feedback (which would constitute a new exposure to items) is taken, successfully retrieved items are strengthened (as per the retrieval effort hypothesis), and non-retrieved items decrease in strength, further diminishing the chance of later retrieval.
As a result, the gap, or bifurcation, between these items is created and subsequently widened with larger spacing and further tests (Kornell et al., 2012). If true, this could possibly create an illusion of tested items being forgotten more rapidly, despite memory decay rate being equal, but where the retrieved items have a larger ‘buffer zone’ before reaching the critical threshold of retrieval.

Two main considerations are made apparent from these theoretical frameworks. First, it is clear that successful retrieval of an item facilitates subsequent recollection (Bjork, 1994; Kornell et al., 2011; Pyc & Rawson, 2009). But how many retrieval events are required to do so optimally? This question was addressed by Vaughn and Rawson (2011), who found that participants who exercised to the point where four to five retrievals were successfully accomplished performed significantly better on a delayed test one week later than those who retrieved an item once. Additionally, Rawson and Dunlosky (2011) showed that, on average, six to seven repeated testing instances with feedback were needed to reach four such retrieval events. This brings up the second consideration, feedback. Test-enhanced learning methods exhibit a clear advantage over simple studying without feedback (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a). Using feedback in the form of displaying the correct target after each testing instance can improve the rate of learning during repeated test trials by as much as 29% when contrasted against a repeated test trial without feedback (Karpicke & Roediger, 2010). Using this manner of feedback is also markedly more beneficial for learning than a simple indication of a correct or incorrect answer (Fazio, Huelsner, Johnson, & Marsh, 2010). In light of the bifurcation model, showing the correct answer as feedback will provide an additional restudying instance, and hence work to prevent the division between items strengthened through successful retrieval and those already below the critical memory threshold (Kornell et al., 2011). Although unmistakably effective as a method of learning, the use of this form of feedback in studying the testing effect has been challenged. By providing an additional exposure to an item, the purity of the task is compromised and the contrast between the study and testing conditions can be viewed as weighted in favor of the testing condition, essentially lending it the benefits of both conditions (Rowland, 2014).

Although comprehensively researched in general, the efforts into exploring the testing effect paradigm had left one key area relatively uncharted until recently. How does test-enhanced learning pertain to individual differences in cognitive ability? A growing number of studies in recent years have addressed this area. A number of cognitive constructs were examined in relation to test-enhanced learning by Brewer and Unsworth (2012). This study showed a significant mediating effect of general fluid intelligence (gF) on learning, in that participants that performed well on the gF measure saw diminishing returns in the benefit of testing. No such interaction was observed for working memory. Working memory (WM) is defined by Conway and colleagues (2005) as “a multicomponent system responsible for active maintenance of information in the face of ongoing processing and/or distraction” (Conway et al., 2005, p. 170). Working memory capacity (WMC) is known as a meaningful predictor of academic achievement (Alloway & Alloway, 2010; Engle & Kane, 2004), and can be reliably assessed with a high degree of validity by a number of working memory tasks (Conway et al., 2005). Additionally, such cognitive faculties as comprehension, problem solving, and reasoning have been implicated as heavily dependent on WMC (Engle, 2002).

The results found by Brewer and Unsworth (2012) stand in contradiction to earlier results in respect to WMC. Agarwal, Rose, and Roediger (2011) found an interaction between WMC and testing, such that low-scoring individuals in the WM task, when taking a delayed test two days after the learning phase, saw greater benefit from a test-enhanced learning method than did their high-scoring counterparts. Another study sought to investigate the efficacy of test-enhanced learning in victims of traumatic brain injury with varying cognitive impairment regarding WMC and gF (Pastötter, Weber, & Bäuml, 2013). When comparing
testing results from individuals with mild, severe, or no brain injury, neither WMC nor gF had any significant effect on the efficacy of testing.

Considering the nature of the testing effect as a method of learning, the process of repeated information retrieval commonly seen in such studies must be considered a central component of the paradigm. As such, there is a remarkable dearth of studies into this process itself, in particular as it pertains to individual differences in cognitive capacity. Most efforts seem to focus on the results of the learning phase rather than the progression thereof. In order to provide more information in the contradictory climate regarding the potency of the testing effect in relation to individual differences in cognitive capacities, Nordstrand (2016) investigated the potential impact of WMC as well as gF on the learning phase of a large number of Swedish upper-secondary level students. It was shown that both measures were correlated with learning over repetitions. The gF measure, however, was so to a far more limited degree. Additionally, WMC indicated a significant interaction with learning over repetition, whereas gF did not. This study was, however, limited to only learning-phase data from the testing condition and could not give an account as to a classic testing-effect result (i.e. the testing showing more durable learning than studying). Further, that study incorporated full feedback at each trial. As no comparison to a repeat study condition was relevant, this is not necessarily a problem, but is important to note. A previous study by Vilhelmsson (2012) investigated the testing effect in regards to Pyc and Rawson’s (2009) retrieval effort perspective and a measure of WMC. Rather than using variable inter-test delay (as per Pyc & Rawson, 2009), this study manipulated effortful retrieval through easy and difficult material to be learned. Results indicated that easier materials showed a larger testing effect than difficult, somewhat discrediting Pyc and Rawson’s (2009) perspective, instead supporting alternate predictions (e.g. Carpenter, 2011). Somewhat controversially, Vilhelmsson (2012) additionally saw a significant advantage in test-enhanced learning (i.e. stronger improvement as a result of testing over studying) for high-WMC participants where previous studies have primarily ranged between no interaction of WMC (Brewer & Unsworth, 2012) and an advantage for low-WMC subjects (Agarwal et al., 2011).

In light of the conflicting accounts as to the possible relation between WMC and the efficacy of test-enhanced learning techniques (e.g. Agarwal et al., 2011; Brewer & Unsworth, 2012; Pastötter et al. 2013), as well as the link between WMC and academic success (Alloway & Alloway, 2010), it is clear that more work is needed in this area. But how does WMC pertain to retrieval effort in the learning process? If different sets of to-be-learned material is equivalent in quantity, but still exhibits variable difficulty for learning, will WMC impact the learning process? The present study will aim to first expand on Nordstrand (2016) and delve further into the learning process in regards to WMC by adding in a study condition and delayed tests for comparison, and second, to examine what the learning process preceding Vilhemsson’s (2012) findings might show in relation to WMC. The following questions will be investigated:

1. Does WMC impact the degree of learning for easy versus difficult material during the learning process?
2. Does WMC impact the degree of retention for easy versus difficult material in post-learning tests?
3. Does the learning method impact the retention of easy versus difficult material in post-learning tests?
4. Will repeated testing result in more learning than repeated studying?
Method

Participants

The participants in the study were recruited primarily through convenience sampling from the student body of Umeå University by bulletin adverts as well as of word-of-mouth. The sample consisted of 99 people ranging between 18 and 35 years of age ($M = 25.62$, $SD = 4.04$, 52% female). Out of these, 46 were recruited for the present study (age: $M = 26.66$, $SD = 3.31$ and 41% female), and an additional 53 participants (age: $M = 24.58$, $SD = 4.66$ and 60% female) were accessed from previously acquired data from 2012. The final count of participants in the testing phase was $n = 85$ as 14 were lost (1 due to misinterpreted instructions, netting negligible $n$-back scores, 4 due to corrupted data files, and 9 due to interruption or failure to return for follow-up sessions).

Criteria for inclusion in the study were defined as being of 18 to 40 years of age, not having English as native language, and being free from diagnoses of dyslexia and learning disability. All participants received written as well as verbal information that they participated of own volition, that they had the right to withdraw participation at any time without explicit reason, and that all information gathered from them was confidential, coded, and untraceable to their person and that data were handled according to PuL (1998: 204) and GDPR (2016: 679).

Design

The experimental design relied on two separate mixed within- and between-subject designs. The first, meant to explore whether there is a difference in the learning of hard vs. easy material dependent on WMC, was a $7 \times 2 \times 2$ design ($learning repetition \times easy vs. hard material \times high vs. low n$-back scores). The second, meant to investigate if WMC and learning method have any effects on the learning and retention of material dependent on difficulty, was a $3 \times 2 \times 4$ design ($test delay \times easy vs. hard \times high and low n$-back scores for both testing and studying conditions, see Table 1 for details).

![Figure 2. The present study design. The to-be-learned material, 30 easy word pairs and 30 difficult, was used for two groups: repeated studying, and alternating repeated studying and testing. Both groups were tested on the material immediately after and on two more occasions, delayed by one and two weeks respectively. On the second session, an n-back task followed the test.](image-url)
Stimuli

To-be-learned material. The material for the learning task consisted of 60 English word pairs. Of these, 30 were categorized as easy and 30 as difficult. The easy word pairs were semantically related (e.g. antler – horn) whereas the difficult word pairs were semantically unrelated and abstract (e.g. adversity – causality). The materials were taken from Nelson, McEvoy, and Schreiber (1998), and used previously by Kornell and Bjork (2009), and Vilhelmsson (2012), and exhibit strong validity and reliability (although with some variation between instances of use, see sources for evaluation). Both difficulty categories consisted of nouns and adjectives. Each word pair was shown in randomized order for 5 seconds during the study condition before moving to the next pair in the series. In the testing condition the first of the words in the pair was shown as a cue and the participant would try to remember and type the corresponding correct word. The participant had 30 seconds to type for each cue, after which the program would move to the next cue automatically, and could press the ‘enter’ key to manually move to the next cue when done. The study condition viewed each word pair a total of 14 repetitions. The testing condition viewed each word pair, was then tested on each pair, and then repeated this cycle for a total of 7 passive exposures and 7 active testing repetitions (see the discussion on successful retrieval earlier; Pyc & Rawson, 2009; see Figure 3). All cues and word pairs were presented in randomized order.

N-back. In the n-back task sequences of letters were used. Ten letters from A through J in the alphabet were presented in randomized order over 27 sequences (see Dahlin, 2009). In every such sequence each letter was shown for two seconds. A fixation cross separated each letter presentation and was shown for 500 ms. In advance of each new sequence of ten letters an instruction image was shown with information on the new task (e.g. if the coming sequence was to be a 1-back, 2-back, or 3-back task). Each of the 1-, 2-, and 3-back conditions consisted of 9 lists of 10 items. Each condition had 36 correct ‘yes’ and 54 ‘no’ answers, totaling 270 as maximum score (see the ‘scoring’ subheading).

There has been some controversy as to both the reliability and the construct validity of n-back as a measure of WMC (Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Kane, Conway, & Colflesh, 2007; Redick & Lindsey, 2013). Satisfactorily comprehensive counter-evidence has been provided, however, that the measure was deemed suitable for the present study (Schmiedek, Hildebrandt, Lövdén, Wilhelm, & Lindenberger, 2009; Schmiedek, Lövdén, & Lindenberger, 2014; Unsworth & Engle, 2007; Wilhelm, Hildebrandt, & Oberauer, 2013). This discussion will be further addressed in more detail later in this paper.

Figure 3. The cycles of the learning task. The study condition (3A) saw the same stimulus every repetition. The testing condition (3B) instead saw a test every second repetition. This continued for 14 repetitions each, 7 of which were tests for the testing condition.
Apparatus

All stimuli (word pairs in testing and study conditions, and $n$-back sequences) were presented at the center of a computer screen on a black background in white text using the font Courier New. The stimulus text was shown in 30 pt. size and instructions and information was of 18 pt. size. E-prime 2.0 was used for the presentation of all stimuli and instructions. All participants were enclosed in a controlled, solitary laboratory environment under similar sound and lighting conditions. Interaction with tasks was conducted by using letter keys of a computer keyboard. All letters were used for writing during any testing on the main learning task, and the ‘J’ and ‘K’ keys were used for the $n$-back task. All data were analyzed with IBM SPSS Statistics V. 25 (SPSS Inc., Chicago, IL).

Procedure

In total the experiment spanned three sessions, each spaced with intervals of seven days.

First session. On the first session, participants were welcomed and thanked for their participation. All participants received written and verbal information on their right to privacy and secrecy in accordance with the Personal Information Statute (PuL, 1998: 204) and the General Data Protection Regulation (GDPR, 2016: 679), as well as their freedom to withdraw participation at any moment without explicit reason. Each participant gave fully informed written as well as verbal consent. Participants were then led into a secluded room with a computer and given verbal as well as written instructions for the learning task.

Subjects in the study condition were informed that they would see a series of English word pairs and were instructed to try and memorize each, i.e. which words were presented together. Subjects were then presented with each word pair in randomized order. This repeated for seven repetitions and then followed by a three-minute resting period during which neutral visual stimuli were presented. After this, another seven repetitions followed, totaling 14 repetitions. Participants in the testing condition were given identical instructions, with the addition of instructions for the following testing condition. Every second study repetition was exchanged for a cued-recall test of each word pair where the cue word was shown followed by a blank space where the correct corresponding word would be typed (e.g. “antler - ____”). These participants were further instructed to type in the corresponding word during these test cycles. After seven alternating repetitions, there was a three-minute resting period followed by another seven alternating study-test repetitions. For passive exposures, each word pair was shown for five seconds before moving to the next item. In each test in the testing condition, the cue and blank space were present for 30 seconds before automatically moving to the next item, but the participant was informed that they could press ‘enter’ to move on manually. During the resting periods, eight images of classical works of art were shown for 10 seconds each. After 14 repetitions, both the study and the test condition received another identical three-minute resting period after which they were subject to a final test of each word pair. The studying condition would thus be tested on each word this one time only, and for the testing condition this would be the 8th test, whereas the study condition would now have seen each word 14 times, and the testing condition only 7 times in addition to any successful retrievals. Both the study and the testing condition took an average of 75 minutes to complete. After the main learning task was completed, each participant filled in a survey with individual information (e.g. age, education, etc.) as well as questions on learning strategies.
Second session. After seven days, participants returned for a second session. In this session, participants once more sat with a computer in a secluded room and were subjected to another test as in the first session. Every participant received a repetition of the same instructions to fill in the missing word that corresponded to the presented cue as for the test or tests in the first session. The same cued-recall test with items shown in a new randomized order followed.

After this seven-day-delayed test, participants performed the n-back task. Each participant was informed that they would see a series of letters in quick succession. They were instructed to try and keep the last few letters in the series in memory and continually indicate with a button press (index finger on the ‘J’ key for ‘yes’ and middle finger on the ‘K’ key for ‘no’) whether the current letter on the screen was the same as shown a certain number of letters previously (one, two, or three letters back depending on the current number in the n-back task, i.e. two letters back for 2-back). Participants received instruction by ways of a diagram as visual aid and instructed to try and indicate a yes or a no answer for each letter during the two seconds it was shown. Participants then went into the same isolated laboratory environment and were given a written summary of the instructions on the computer screen before starting. Each sequence of 10 letters was preceded by an instruction screen which indicated which finger and corresponding key was used for yes and no answers respectively, as well as a text which informed which n-back (i.e. 1, 2, or 3-back) task was relevant for the following sequence.

Third session. After 14 days all participants returned once more for a last session. In this session the participants were once more given the same instructions as the second session and then performed the same cued-recall test for each item. Cues were once more presented in randomized order.

Scoring

Scoring for all cued-recall tests considered the answer correct if no more than three letters deviated from the correct target word. Answers were considered incorrect if deviating less than three letters from the target, but also possessed a separate semantic definition (i.e. the target was ‘horn’ and the answer was ‘hornet’). Alternations on the basic word, such as verb conjugations, or number for nouns (i.e. written in plural instead of singular), were considered correct if within the three letter limit (e.g. the target was ‘horn’ and the answer was ‘horns’). Easy and difficult words were corrected separately for ease of comparison between item groups in the difficulty variable. Scoring on the n-back task considered an answer correct if the correct key was depressed during each letter presentation (i.e. if the currently showing letter was corresponding to one, two, or three letters back or not and this was correctly indicated with the correct key). As such, missing values, that is a letter stimulus with no corresponding button indication, were considered incorrect. Likewise, false alarms were considered equally incorrect. All materials were extracted from output, compounded, organized, and scored separately and independently by two different researchers. All scores were then compared and controlled.

Statistical Analysis

Two separate mixed model ANOVAs were used for data analysis. The first was performed to establish whether performance on n-back scores was associated with the degree of learning dependent on the difficulty of the material. This was a 7 (test repetition) x 2 (easy vs. difficult) x 2 (high and low n-back scores) configuration. A second ANOVA was used to
ascertain whether learning method or material difficulty were associated with long-term retention dependent on material difficulty. This was a 3 (immediate, 7 days, 14 days) x 2 (easy vs. difficult) x 4 (high and low n-back scores per condition) model. When assumptions of sphericity were violated, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. Finally, a 3 (delay) x 2 (learning condition) ANOVA will assess if the testing condition results in a higher degree of learning. Bonferroni correction was used to adjust $\alpha$-levels where appropriate. Partial $\eta^2$-values are reported to indicate effect size.

**Results**

To investigate the role of WMC in relation to testing, material difficulty, and delay, n-back scores were divided into low (lowest score through 40$^{th}$ percentile) and high (60$^{th}$ percentile through max score). This resulted in two WM groups each in the study and testing condition (see Table 1 for details).

| Table 1 |
| Descriptive statistics of n-back scores. Table shows the number of participants per group and mean proportion of correct answers (correct 'yes' or 'no' button press for each item). | Study condition | Test condition |
|---|---|---|---|
| Descriptive | Low WM | High WM | Low WM | High WM |
| $n$ | 17 | 18 | 17 | 15 |
| $M$ (SD) | .67 (.11) | .78 (.06) | .69 (.11) | .90 (.06) |

$M$: mean proportion correct, $SD$: standard deviation, $n$: number of participants

**Learning phase**

In order to ascertain if n-back scores and material difficulty were associated with a difference in learning, learning phase data were subjected to a mixed-model, 7 (repeated tests) x 2 (material difficulty) x 2 (high vs. low n-back scores) mixed model ANOVA. Mauchly’s test for sphericity indicated that the assumption of sphericity was violated for repetition ($\chi^2(20) = 161.24, p < .001$) and degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = .38$), as well as for repetition by difficulty ($\chi^2(20) = 198.28, p < .001$), where degrees of freedom were likewise corrected ($\epsilon = .28$). Corrected results revealed a main effect of learning ($F(2.3, 68.98) = 376.38, p < .001, MSE = 15.94, \eta^2_p = .926$), and of difficulty ($F(1, 30) = 131.32, p < .001, MSE = 88.1, \eta^2_p = .814$), indicating learning over time and a difference in material difficulty. Pairwise comparisons further confirmed that there was significant improvement between each successive practice trial ($p < .001$ on all accounts except repetition six and seven ($p = .16$), see Figure 4). Greenhouse-Geisser-corrected results showed an interaction of repetition and difficulty ($F(1.65, 49.42) = 25.69, p < .001, MSE = 34.2, \eta^2_p = .461$), indicating a deviation between material difficulty levels over time. Analysis did not, however, show a significant interaction of difficulty and n-back scores ($F(1, 30) = 1.05, p = .236, MSE = 294.57, \eta^2_p = .05$), indicating no relationship between n-back and material difficulty. No interaction was observed for repetition and WMC ($F(2.3, 30) = .76, p = .487, MSE = 294.57, \eta^2_p = .025$). Finally, no significant difference in performance over the learning phase was revealed...
between n-back groups, although results approached significance, with low-scoring subjects possibly performing higher in the learning phase (F(1, 30) = 3.32, p = .079, MSE = 294.57, η²_p = .01; see Figure 4). Post-hoc power analysis for between-subject comparisons indicated very low statistical power for n-back scores (.422).

**Figure 4.** Mean number of correct answers for each repetition of cued recall tests in the learning phase and top and bottom working memory groups. 4A shows learning for easy items and 4B for difficult items. Error bars represent one standard error of the mean.

**Testing phase**

In order to investigate if n-back scores and learning condition were associated with degree of learning and retention, dependent on material difficulty, testing phase data were analyzed by ways of a 3 (immediate, 7-day delay, and 14-day delay tests) x 2 (easy vs. difficult) x 4 (high and low WM for study and testing conditions as per Table 1) mixed model ANOVA. Mauchly’s test for sphericity revealed that the sphericity assumption was violated for delay (χ²(2) = 21.11, p < .001), so degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .77), revealing a significant main effect of delay (F(1.54, 92.25) = 193.06, p < .001, MSE = 17.54, η²_p = .763), indicating a clear forgetting effect over time (as illustrated by Figure 5). Delay by difficulty was likewise violating sphericity (χ²(2) = 59.34, p < .001) and corrected (ε = .61), showing an interaction (F(1.22, 73.43) = 54.06, p < .001, MSE = 10.96, η²_p = .474), indicating a more rapid decline in retrieval over time for the difficult material than for easy (detailed in Figure 5). Delay by n-back group and condition did not show significance (F(6, 60) = 1.53, p = .18, MSE = 186.49, η²_p = .07) indicating no interaction. Likewise, difficulty level by n-back and study/testing group showed no significant interaction (F(3, 60) = .55, p = .649, MSE = 186.49, η²_p = .027). Between-subjects comparisons showed no significance between the four n-back by condition groups (F(3, 59) = 2.49, p = .069, MSE = 186.49, η²_p = .112), but was approaching significance (illustrated in Figure 5). Post-hoc power analysis of the n-back by condition comparison indicated low statistical power (.588). Finally, a 3 (delay) x 2 (condition) ANOVA showed an effect of test performance and condition. Comparisons showed a significant difference in performance between test and study conditions (F(1, 83) = 9.47, p = .003, MSE = 162.61, η²_p = .102), indicating a standard testing effect.
Figure 5. Mean number of correct answers on the immediate and delayed tests and comparison groups (as per Table 1). 5A shows score for easy items and 5B for difficult items. Error bars represent one standard error of the mean.

Discussion

The first aim of the present study was to investigate whether individual differences in working memory and the difficulty of the material to be learned affects the degree of learning when using testing as a learning method. Secondly, this study aimed to explore if working memory capacity affects the efficacy of learning between a studying and a testing condition dependent on different difficulty levels of the material. It was observed that WMC had no impact on either the learning process or the performance in the testing phase. The learning of and performance on tests for the two difficulty levels were both similarly unrelated to the chosen WMC measure. Considering the weakness of the observed power in this analysis and the relative proximity to significance, however, it is not unthinkable that the low number of participants in the comparison groups obfuscate an effect. Performance for WMC groups leaned in the somewhat unexpected direction of low-WMC participants performing better during the learning phase than their high-performing counterparts. Further, there was no effect of WMC by condition group for performance in the testing phase, but similar to WMC and learning performance, statistical power was very low. Results approaching significance might similarly conceal an effect of more durable learning for the high WMC groups.

Analysis of the learning phase showed a main effect of difficulty, confirming that the material was indeed of differential difficulty. Additionally, results revealed an interaction of repetition and difficulty, showing that learning was achieved significantly faster for the easier items (see Figure 4). The material used is thus replicated as a measure of differential material difficulty and further reliability and validity is provided. The testing phase results indicated a main effect of delay, meaning that items were forgotten over time. More noteworthy, however, there was an interaction effect of difficulty and delay, signifying that items of higher difficulty were forgotten considerably more rapidly than those of lower difficulty (see Figure 5). Finally, a traditional testing effect was observed, entailing the advantages known for test-enhanced learning.

From these results, it seems that the benefits of test-enhanced learning hold stable over the difficulty of the material even when of equivalent complexity (as opposed to
previous comparisons contrasting differing levels of complexity; Karpicke & Aue, 2015), and individual differences in WMC. They do, however, spell out a possible challenge to the retrieval effort hypothesis. Pyc and Rawson (2009) say that “The basic claim of the retrieval effort hypothesis is that not all successful retrievals are created equal: given that retrieval is successful, more difficult retrievals are better for memory than less difficult retrievals” (Pyc & Rawson, 2009, p. 439). The results here appear to support the first half of this claim. The pronounced drop in performance after a week's delay for difficult items (as seen in Figure 5), however, stand in stark contrast to the second half, as the items that would require more effort to retrieve are far less likely to survive until the delayed tests. It is clear that material difficulty used here is not a valid measure of what Pyc and Rawson (2009) measure with their inter-test interval manipulation.

Another possibility is that this disparity is an effect of easier items more quickly reaching a greater number of successful retrievals, and thus facilitating more durable learning. If such an effect were to outrank the observed drop in performance from the immediate test to the first delayed test by a large enough magnitude of impact, it would effectively counter these findings. As no such analysis was included in the present study, this could be a reasonable question for further research. The bifurcation model (Kornell et al., 2011) would support this view. According to present results, the number of retrievals was observed to more rapidly reach more instances for enhancing the memory strength of these items, which would cause the non-retrieved, or less retrieved items from the difficult material to more readily fall below the critical threshold of recall. That is, the critical threshold could effectively be considered higher for more difficult material.

Yet another, rather unanticipated, interpretation from this perspective (made even more complex by later discussion of n-back tasks and WMC) is that low-WMC individuals may struggle more to remember a given item. This would cause more effortful retrieval, more substantially cementing encoding, and thus explaining the slight tendency for these individuals to perform better than their high-WMC counterparts during the learning phase. As these results did not reach significance, however, this is pure conjecture.

One possible explanation for the discrepancy shown here comes from the desirable difficulty paradigm (Bjork, 1994). If the material used here was too difficult to be desirable, learning might suffer for it. As average scores were not conspicuously low, this seems unlikely, but a number of participants did fall to zero correct retrievals on the delayed tests for the difficult material. Perhaps individual differences in some cognitive faculty other than WMC could account for this.

These results, although still useful for practical educational considerations, seemingly add yet further to the confusion on the role of WMC in the test-enhanced learning paradigm – even more so if we take into account the tendency of low-WMC individuals to almost outperform their high-capacity counterparts on the main learning task, going against all earlier results (Agarwal, Rose & Roediger, 2011; Brewer & Unsworth, 2012; Nordstrand, 2016; Pastötter, Weber, & Bäuml, 2013; Vilhelmsson, 2012). In Kane and colleagues’ (2007) view, this discrepancy, at least for the present study, is easily remedied. In their study, they concluded that n-back tasks are weakly correlated with other WM span tasks, and do thus not represent the same underlying construct. Jaeggi and colleagues’ (2010) findings, meanwhile, indicate that n-back tasks lack sufficient reliability to be useful as measures of differences between individuals. If n-backs are not valid measures of WMC, this would explain the diverging results in the present study. However, it seems that most previous conflicting studies have actually employed various complex span tasks (e.g. Unsworth, Heitz, Schrock, & Engle, 2005), such as operation span, reading span, and symmetry span (e.g. Aslan & Bäuml, 2011; Brewer & Unsworth, 2012; Unsworth, 2009) that are supported as superior measures of WMC by the critics of n-back (Jaeggi et al., 2010; Kane et al., 2007). Although
the arguments raised against n-back as a valid WMC measure have been countered with conflicting results themselves, indicating that it works equally well as complex span tasks (Schmiedek et al., 2009), and that these tasks share an underlying construct (Wilhelm et al., 2013), these conflicting accounts must be considered. Redick and Lindsey (2013) similarly found disparity between these measures, and suggested complex and simple span tasks cannot be used interchangeably with intact validity (this was then countered by results from Schmiedek et al., 2014).

Various n-back tasks are simple working memory tasks broadly used as measures of WMC (Conway et al., 2005; Kane, & Engle, 2002). Simple WMC tasks include no distracting element between mental manipulations of the material, as opposed to complex WMC tasks. It has been suggested that complex span tasks, although highly intercorrelated, might not be a good measure of pure WMC, as they involve separate unrelated tasks as distractors during the process (Schmiedek et al., 2009; Wilhelm et al., 2013). In fact, as primary WM is engaged with a distraction, it has been argued that a secondary memory system might be responsible for transfer of information to longer term storage (rather than active WM) during the distraction (Wilhelm, et al., 2013). As such, simple span tasks, such as n-back, might be cleaner measures of pure WMC altogether. If true, this would move the simple vs. complex span controversy toward synthesis, providing support for both approaches, but as measures with different appropriate applications. Moreover, as complex tasks involve separate processes (equations for automatic operation span, reading for reading span etc.), depending on which kind is used, it may bring with it unknown confounds to the study design. As this pertains to the testing-effect paradigm, as well as any design involving a learning task, if Wilhelm and colleagues’ secondary memory system is engaged, this might correlate highly with the learning task (as complex WMC tasks have done historically) because they involve the same storage processes, rather than the WM component of the task. Such an oversight would constitute a critical confound. Effects of this nature will be less pronounced when composite measures are used, as more variety in distracting elements is employed. Indeed that is precisely the case for Brewer and Unsworth (2012) who used several complex working memory tasks and found weak to no effect of WMC. Note, however, that they used no simple span tasks.

A number of issues with the present study design were made apparent during execution. A main concern is the use of English word pairs as primary material. Considering the ubiquitous penetration of English culture in western society in general, and Sweden in particular, the language proficiency levels will vary greatly. Although proficiency tends to be higher in a university population, levels will likely still be subject to high variability. Knowing the meaning of a given word in advance will likely provide an advantage over those who are not semantically familiar with the material in encoding the pairs, and if not, this is a clear confound. Some studies have used word lists with a foreign language that is highly obscure for the target population (e.g. Swahili words; Nelson & Dunlosky, 1994). This is likely a more stable approach, even though it may exclude a minority of the population that speaks e.g. Swahili. Even so, the variability in language proficiency in so doing will likely be close to binary – fluent speaker or complete unfamiliarity, and thus easily controlled. Some of the variability at play here is partly controlled by the semi-arbitrary pairing of words (e.g. causality – adversity). Another issue with the material is the pairing of the easy words. The list was initially developed in the USA (Nelson et al., 1998), and some pairs may carry greater significance in that cultural context. An example would be the word pair ‘cane – able’, which is a homophone play on the biblical reference of Abel and Cain, which in the more secular cultural context of Sweden may not carry the same connotations (as one participant volunteered about memory strategy, “I thought of a cane, which you don’t need if
you’re able to walk freely”). This seems to have played only a minor role if any, however, as there was a marked difference in material difficulty levels.

One weakness of the material would appear considerably more important. As can be seen in the learning phase data (Figure 4), the easy material was subject to a marked ceiling effect of learning in a way the difficult was not. As already made clear, the difficulty levels were still evident, however, in relation to the successful retrieval and bifurcation paradigms, this becomes a problem. Since, for many participants, the easy material reached full score before the last repetition, a larger number of successful retrievals of that material would be facilitated per item than for the difficult items. This could account for the difference in retention on the delayed tests, possibly casting doubt on other explanations. Although currently confounding, such circumstances may not be contrary to real field conditions in an educational setting as materials of various difficulty levels might share the same time restraints for learning. Thus, ecological validity might not necessarily be lacking as a result. For any further studies with this material, this issue should nonetheless be redressed and remedied, possibly by ways of increasing the number of items.

Additionally, a number of words were very similar (e.g. aptitude and attitude). Although not paired, they were both on the list and there was still a tendency to get these confused. When the scoring criteria specify a three-letter margin for spelling, the different words should strive to fall outside that margin of confusion. Further, a large number of the words in the material were four letters long, meaning that a three-letter leeway can result in a correct score for typing, for instance, one letter, or 25% or the word. In total there were 12 such words, and 10 of these were found in the easy list. This, too, may have skewed the results for the easy list. If a pre-determined letter number requirement is used, the length of the words should exceed a certain length, and strive to be of roughly equivalent length.

Lastly, a more difficult confound to address was shown in the post-learning surveys. A total of 5 participants in the study condition indicated that they used self-testing as a strategy, either mentally or by physically covering the screen with a hand for an unknown number of repetitions. What then of the possibility of more people doing the same without self-reporting? Instructing participants in advance to not engage in such self-testing is itself influencing them and creating confounding circumstances. If some sort of manual manipulation task could keep the subjects’ hands away from the screen, this might function as a distraction from the learning task. Further studies into test-enhanced learning could attempt to resolve this issue.

Although recruitment comprised a rather substantial number of participants, the nature of this study, using high- and low-performing percentile groups and two conditions, the final number in a given analysis is low. It is clear that this affected the statistical power of the analyses, and future studies could work with a narrower scope and focus on fewer variables, thus ensuring more reliable analyses.

The present study resulted in some insights into the paradigm, which might inform practical application. The benefits for test-enhanced learning hold true, further supporting its role in the development of learning methods. The results presented in this study support the understanding that testing is equally efficacious for all students, independent of individual cognitive capacity, demonstrating the suitability of such methods in the classroom. Additionally, if the bifurcation and successful retrieval explanations of the drastic reduction in correct answers for the delayed tests hold true, this could be useful in educational settings. A teacher can, for instance, identify what material is more difficult or through the course of instruction is revealed to require more attention for individual students. In response teachers can not only supply testing instances to enhance retained information, but also tailor instruction by a student’s or material’s requirements by increasing or decreasing (to free up
time for activities with higher time requirement) the number of testing instances as appropriate.

As the study condition lacks data on the learning progression, future designs similar to the present study could try to incorporate a smaller number of tests for the studying condition (e.g. every third or fourth instead of every second repetition) for comparison with a more intense testing regime, in order to gain more understanding of the learning process. More importantly, however, far deeper investigations are needed as to the role of WMC for test-enhanced learning. As has been suggested herein, the paradigm need not only more studies focusing on WMC, but also more variability thereof, with a larger focus on simple span tasks. Such studies could move to bridge the gap between the conflicting and variable results currently on display and yield important insights into the field.

In conclusion, as well-substantiated as the test-enhanced learning paradigm is as a potent tool for the facilitation of durable learning, its underlying mechanisms and their interplay with various circumstances in practical implementation must still see considerable exploration for its potential to be fully uncovered. In light of the contradictions seen in regards to individual differences in cognitive function (Agarwal et al., 2011; Brewer & Unsworth, 2012; Pastötter et al., 2013), it is clear that this field requires more attention. The growing effort to investigate this area, which the present study aims to further, is gradually showing new venues of research that may clarify the picture. Unraveling the optimal circumstances and subjects for implementation of test-enhanced learning methods in the classroom can help advance teaching practices and strengthen the educational system as a whole. Insight into the relation between WMC and testing can help teachers understand the role of students’ individual faculties and needs in a new way. Cognizance of how and when the branching of learning first happens between materials for individual students, teachers may be better able to intercept and contravene such effects, facilitating a more stable classroom. As an example, a teacher may plan more opportunities for retrieval of a material early for students in further need of assistance, offsetting a measure of the division of learning between lower and higher performing students later in the curriculum (Freeman, et al., 2014). There seem to be several interesting venues for further research into the learning process in relation to the testing-effect paradigm and individual cognitive differences.
References


