Test-enhanced learning, working memory and fluid intelligence

Dennis Nordstrand
Dennis Nordstrand

Under det senaste decenniet har testbaserat lärande väl etablerats som ett effektivt sätt att främja hållbar inlärning. Många sorters material och omständigheter har utforskats i relation till denna metod. Endast nyligen har dock individuella skillnader i relation till testbaserat lärande fått uppmärksamhet som ett forskningsområde. Ett område hittills förhållandevis outforskat är relationen mellan individuella skillnader i kognitiv kapacitet och inlärningsprocessen med återhämtning som inlärningsmetod. Denna studie hade för avsikt att utforska denna relation genom att mäta generell flytande intelligens och arbetsminneskapacitet för ett urval av gymnasieelever (n = 189, M = 16.89 år gamla) som använde testbaserat lärande som inlärningsmetod. Resultaten indikerar att arbetsminne och flytande intelligens båda är relaterade till inlärningsprocessen, men att det förstnämnda är så till en signifikant högre grad än det sistnämnda.

During the last decade, test-enhanced learning has been thoroughly cemented as an efficient way to promote durable learning. Many materials and conditions have been explored in relation to this method. Only recently, however, have individual differences in relation to test-enhanced learning received attention as an area of study. An area as of yet relatively unexplored is the relationship between differences in cognitive ability and the process of retrieval as a method of learning. The present study set out to explore this relationship by measuring general fluid intelligence and working memory capacity in a sample of upper secondary level students (n = 189, M = 16.89 years of age) who used a test-enhanced learning method. The results indicate that working memory and fluid intelligence are both related to this learning process, however the former to a significantly higher degree than the latter.

One of the prime applications of the science of learning and memory in terms of utility is within the field of education (McDaniel, Roediger, & McDermott, 2007). Teachers and students alike have a limited amount of time at their disposal to build a solid, durable knowledge base in a given subject. Although constructing durable knowledge is, in theory, a trivial task, practical applications of proven methods are seldom realistic. For instance, rereading of materials has long been shown to increase the learning rate of a material as a function of the number of repetitions (Rothkopf, 1968). This method, however, is time consuming and an indefinite number of repetitions of every issue in every subject is hardly workable. The challenge is, thus, to promote durable learning as efficiently as possible (Rawson & Dunlosky, 2011). Depending on discipline and school of thought, the terms learning and memory can be ascribed many and varied meanings and implications. A concise but serviceable definition in the realm of cognitive psychology was posited by Gluck, Mercado and Myers, and reads as follows: “learning is the process by which changes in behaviour arise from experience through interaction with the world; memory is the record of past experiences acquired through learning” (Gluck, Mercado and Myers, 2008, p. 39). When these
terms are used in this report, this is the context in which they are to be viewed. Since memory and learning have been subject to research for over a century a wealth of knowledge is available on the topic. It seems, however, that a very limited amount of this information is turned to practical use in actual schools. Indeed, Dunlosky and colleagues (Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013) evaluated ten of the most common learning techniques used by teachers and students in schools and their findings support this view. In their evaluation they took into account four main factors: learning conditions (e.g. settings), student characteristics (e.g. ability level), materials (e.g. from simple to more complex materials), and criterion tasks (outcome measurements of learning). Of the techniques evaluated, all but two were deemed as being of low to moderate utility in their ability to promote efficient and durable learning, with several aspects of these appraised as insufficient. Among many other things, this conclusion demonstrates the importance of, and need for, a solid empirical basis in teaching methods. One of the two techniques deemed as having high utility was, unsurprisingly, practice testing. Practice testing is defined in the study as “self-testing or taking practice tests over to-be-learned material” (Dunlosky et al., 2013, p. 45).

Although the educational system has traditionally used testing as a means of assessing already established knowledge and proficiency of a given subject, it has been repeatedly demonstrated that the process of testing is, in actuality, itself an effective method of facilitating learning (Dunlosky et al., 2013; Karpicke & Aue, 2015; Karpicke & Roediger, 2008; McDaniel et al., 2007; Roediger & Karpicke, 2006a, 2006b). Indeed, ample evidence exists that the process of retrieval of previous stimuli strengthens the ability to recall a given piece of information again at a later point in time (e.g. Dunlosky et al., 2013; Karlsson-Wirebring et al., 2015). This phenomenon is commonly referred to as the testing effect. The testing effect was initially observed as early as a century ago, but as a field of research it experienced a revival with Karpicke and Roediger’s attention a decade back (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a, 2006b).

Typically, the testing effect is examined by one of two methods (see Figure 1). The first entails exposing subjects to a certain stimulus, or to-be-learned material and then dividing the subjects into two groups, one of which will repeatedly restudy the material and one of which will repeatedly be tested on it (a between-group design). The second is to, after initial exposure, divide the material into two conditions, one to be restudied and one retested (a within-group design). Both variants are usually followed by an immediate test on the material directly after the repeated study or test condition (i.e. the learning phase) and then learning is assessed again by a delayed test days or weeks thereafter (see Figure 1).
Figure 1. A typical testing-effect study. After exposure to the to-be-learned material participants or material is divided into two groups or conditions, one of repeated studying, one of repeated testing. Following this learning phase two or more tests are administered, one immediately after learning and one or more delayed.

At the immediate test it is usually found that the repeated study condition confers a higher rate of retention than the repeated test condition. At the delayed test, however, the subjects who were placed in the repeated test group tend to show a significantly higher retention rate than those within the repeated-study group, illustrating the typical testing effect.

The testing effect has been demonstrated to be remarkably stable across a wide array of various conditions and alternate materials to be learned. Many types of to-be-learned material have been investigated; word pairs (Wiklund-Hörnqvist, Jonsson & Nyberg, 2013), semantic facts (Carpenter, Pashler & Cepeda, 2009), functions (Kang, McDaniel & Pashler et al., 2011), and text content (Karpicke & Blunt, 2011; Karpicke & Roediger, 2010), being only a handful of examples. Other variables explored include repeated vs. single tests (Larsen & Butler, 2013), age of subjects (Lipowski & Pyc, 2014), spacing between tests (Larsen & Butler, 2013), and free or cued recall testing (Wiklund-Hörnqvist et al., 2013; Zaromb & Roediger, 2010). Additionally, this method has been contrasted to other, alternate methods, such as group discussions (Stenlund, Jönsson & Jonsson, 2016) and concept maps (Rawson & Dunlosky, 2011; Karpicke & Blunt, 2011). Yet the testing effect stands in the face of many and varied challenges, in regards to simple as well as complex materials (Karpicke & Aue, 2015), and its potential benefits in education are undeniable.

Although the testing effect has been extensively and thoroughly investigated in regards to material to be learned and test circumstances, it has been claimed that the theoretical explanation models have fallen behind. In more recent years, however, a number of more specific models have been developed as the abundance of studies examining the effect has increased further (Lehmann, Smith & Karpicke, 2014). The Desirable difficulty perspective, initially proposed by Bjork (1994), is essentially the notion that items with higher retrieval difficulty during encoding will be more easily recalled once successfully retrieved than will ones of lower difficulty. A similar model, the retrieval effort hypothesis posits that retrieval during practice is more effective in promoting durable learning if it is both successful and requires more effort (Pyc & Rawson, 2009). Investigations into this model have manipulated the interval between stimuli in order to examine whether more clustered or spaced testing leads to more stable learning. Findings indicate that the more spaced successful retrieval instances are the more stable the memory of the item is (Pyc & Rawson, 2009).
Bjork & Bjork’s (1992) model of memory strength asserts that representation of memory is defined by two different kinds of strength: storage strength and retrieval strength. Storage is the main factor in long-term retention. When an item is stored it is made available for later retrieval and the higher the degree of storage strength, the higher the likelihood of retrieval. Retrieval strength, however, concerns transitory access to information in the short term. This model posits that encoding as well as retrieval events have the capacity of increasing both storage and retrieval strength, but that retrieval-oriented activities have a higher potential for facilitating such improvement. Because of the continuum oriented nature of this proposed configuration of memory, a stored item may possess a certain degree of memory strength and yet not successfully be recalled. A consequence of this configuration was proposed by Kornell, Bjork and Garcia (2011). As previously stated, the gains from retrieval seem to increase with longer delay (Pyc & Rawson, 2009). According to the bifurcation model this may be due to an effect of the recalled items being above a certain recollection threshold in strength, but non-recalled items being situated below this critical boundary (Kornell et al., 2011). Tests without feedback will, thus, create a divide between successfully retrieved items (see the retrieval effort hypothesis) and those not recalled, with items recalled growing stronger and those not recalled lagging behind with repeated tests and longer delays. This causes a gap, or bifurcation, in the distributions of these items (Kornell et al., 2011). Hence, this may create an illusion of the tested items being forgotten at a more rapid pace even though memory decay may happen at an equal rate with the tested items having a larger ‘buffer zone’ to the threshold limit value of retrieval.

From the above offered theoretical frameworks two main considerations become apparent. The first issue is that of successful retrieval. Apparently items successfully retrieved seem to be more readily recalled when later tested (Bjork, 1994; Kornell et al., 2011; Pyc & Rawson, 2009), but is one successful instance of recalling enough? If not, how many repetitions are needed in order to achieve the full potential of successful retrieval? Vaughan and Rawson (2011) discovered that students who practiced until four to five successful retrievals were accomplished performed significantly better on a one week delayed test than did those who achieved only one successful retrieval. Rawson and Dunlosky (2011) further demonstrated that in order to achieve four successful retrievals, six to seven repeated tests with feedback were required. This raises the second issue, which is feedback. Although repeated testing holds a significant advantage over repeated studying of a material even when feedback is not provided (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a), testing with feedback seems to be even more beneficial (Karpicke & Roediger, 2010). Karpicke and Roediger (2010) found that a repeated testing condition where feedback was provided increased performance, during learning and on a one week delayed test, by approximately 29% compared to testing without feedback. When feedback takes the form of showing the correct answer rather than a right/wrong indication it not only lets a subject know when a correct answer is given but also provides an additional instance of studying. When Fazio, Huelser, Johnson and Marsh (2010) investigated these two alternative forms
of feedback the former was indeed found to be the most advantageous. As seen in the bifurcation model, providing an opportunity for restudying of all items in the form of feedback after correct or incorrect answers is also theorized to prevent the gap between successfully retrieved items and non-retrieved items from growing wider and may facilitate encoding of previously lagged items (Kornell et al., 2011).

As thoroughly explored as the testing effect is in many respects, one area, however, seems as of yet to largely avoid scrutiny. The relationship between individual differences regarding various cognitive constructs and the efficacy of testing on learning a given material yet remains a weak point in the mapping of the testing effect phenomenon. Two such constructs that have previously been investigated in relation to the testing effect are general fluid intelligence (gF) and working memory capacity (WMC) (Agarwal, Rose & Roediger, 2011; Brewer & Unsworth, 2012; Dunlosky et al., 2013; Pastötter, Weber & Bäuml, 2013). The gF construct reflects an individual's ability for inductive and analytic reasoning, and a capacity for problem solving without depending on knowledge gathered from prior experience. It is a representation of an ability to discern and deduce relationships and a capacity for adaptation to new situations (Cattell, 1963). Working memory can be explained 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) and its capacity can be measured with a high degree of reliability and validity by various working memory span tasks (Conway et al., 2005). Working memory capacity has been shown to be not only its own cognitive skill, separate from intelligence, but also an important predictor for measures of academic success (Alloway & Alloway, 2010). Furthermore, WMC has been shown to be a central component of a wide array of cognitive behaviors, including comprehension, reasoning, and problem solving (Engle, 2002).

Brewer and Unsworth (2012) investigated a number of constructs and their impact on the testing effect. It was found that general fluid intelligence (gF) had a significant mediating influence as high-performing participants on gF tasks saw diminishing returns from testing. Pastötter et al. (2013) examined retrieval practice (another term for practice testing) in people with mild, severe, or no traumatic brain injury. It was found that despite varying degrees of general cognitive impairment, in regards to general fluid intelligence as well as working memory, there were no differences in the potency of the testing effect. Due to these variations in results it is clear that more research is needed in order to ascertain to what degree, and for whom, the testing effect may best be utilized to facilitate efficient learning. In the Brewer and Unsworth (2012) study, some results appear to be in conflict with earlier findings regarding the working memory capacity construct (WMC) as no difference between high- and low-performing participants was found (Brewer & Unsworth, 2012). Conflictingly, Agarwal et al. (2011), found a significant link, where low scorers on WMC tasks saw improved benefit from retrieval practice compared to those with higher scores at a test two days later (Agarwal et al., 2011).
Considering the implications of the testing effect, the actual process of encoding (see Figure 1) should be seen as an operative part of the mechanism of test-enhanced learning. Surprisingly this element of the phenomenon is severely underrepresented in the testing effect literature. Furthermore taking into account the link between WMC and academic achievement (Alloway & Alloway, 2010), and the conflicting results pertaining to its potential link to the potency of the testing effect (Agarwal et al., 2011; Brewer & Unsworth, 2012; Pastötter et al. 2013), the importance of further examination of WMC and its impact on the learning process becomes clear. In summary, it seems that the mechanisms of the actual learning process during the intervention phase of testing effect studies still widely constitutes unknown territory. Moreover, cognitive ability as it pertains to this process of learning promises to unearth useful insights into the workings of the phenomenon. The prevailing absence of such studies is surprising and should be remedied (Dunlosky et al., 2013). Given the previous work done on the relationship between WMC, gF, and the testing effect, the purpose of the present study was to investigate whether working memory and/or general fluid intelligence is related to the learning process during test-enhanced learning. More specifically put: is there a significant relationship between the degree of learning during test-enhanced learning, and scores on cognitive tests for the assessment of WMC and/or gF respectively?

Method

Participants
The sample consisted of 189 first year upper-secondary level students from natural science, technological, and child-care programs of a Swedish upper-secondary level school. The age range was 16-20 ($M = 16.89, SD = .65$) and 38.8% were females. All participants were given verbal as well as written information before the data collection, and participation was voluntary. Written informed consent was collected in accordance with the Helsinki declaration and the study was approved by the Regional Ethics Committee of Umeå University (2015/382-310).

Material & design
The current study used a within-group design examining the testing effect. In order to explore the impact of individual differences in cognitive ability on the potency of the testing effect, measurements of gF and WMC were also collected.

To-be-learned material
The material to be learned consisted of 60 Swahili-Swedish word pairs used in previous studies (Karlsson-Wirebring et al., 2015; Nelson & Dunlosky, 1994). This material was presented in a learning phase consisting of studying and cued recall testing conditions (see Figure 2). Participants restudied 30 of the word-pairs, and were repeatedly tested on the other half. This cycle was repeated totaling 6 instances each of study and test conditions. Feedback in the form of the correct word-pair was provided after each test item independently of correctness of the response. Each participant was given a unique, randomized list of test and study
items respectively from the total pool of the 60 word pairs in order to avoid item effects. All word pairs were presented in the middle of the computer screen. Study items were presented for 8 seconds, and test items provided 8 seconds to type the corresponding Swedish word followed by immediate feedback in the form of the intact word pair shown for 1 second. As no immediate or delayed test was used in this study, the test condition is the only one relevant for upcoming analyses (see Figure 2).

**Figure 2.** Study design in its entirety and the phase focused on in this study (highlighted). The repeated test condition is the main point of interest.

**General fluid intelligence**

To measure gF, Raven’s advanced progressive matrices (APM) were used (Bors & Stokes, 1998). APM is a visuo-spatial reasoning task which measures aspects of gF. Each item in APM consists of a 3x3 setup of cells in a matrix of patterns and various geometric elements dictated by non-disclosed rules with the last location left empty. Provided were 8 alternatives containing similar geometric elements from which participants could choose. Subjects were instructed to determine which alternative was appropriate to fill in the empty space. The position of the correct answer was randomized to be equally distributed between the alternatives by the end of the task. Participants were given 25 minutes to complete the entire task and were continuously shown the remaining amount of time. The version of APM used was Bors and Stokes’s (1998) short-form version. This version contains half the amount of items (24) of the original and exhibits strong correlation with the original long form ($r = .88$; Bors & Stokes, 1998). APM was chosen as a measure for gF due to having been thoroughly tested and having exhibited robust reliability (Alderton & Larson, 1990; Chiesi, Ciancaleoni, Galli, Morsanyi, & Primi, 2012).

**Working memory**

The automatic operation span task (Aospan) was used for measuring WMC. The Aospan tool was chosen due to its high internal consistency (.78) and test-retest reliability (.83; Unsworth, Heitz, Schrock, & Engle, 2005). Aospan entails the assessment of a math equation and the correctness of its provided answer. Following each such assessment the subject is exposed to an alphabetic character and is instructed to remember these letters and their order of presentation. To ensure that participants do not trade off a lower score on the equations for a higher score on the letters a math-equation accuracy of 85% was encouraged. In this study this was done by showing the subjects a diagram in the top right corner of the screen, which indicated when scores were above or below the 85% limit. This cycle was repeated in set sizes of 3 to 7 items, with sizes varying randomly, prohibiting participants’ prediction of set sizes. After each set, participants were tasked to recall the series of letters in the correct order as these were presented.
Feedback was provided for math-equation accuracy as well as letter responses. As this task involves remembering information parallel to simultaneous processing of information it constitutes a complex working memory task (Unsworth et al., 2005). The final score on the task was the total number of letters correctly recalled in their correct positions.

Scoring procedure and apparatus
Answers entered during cued-recall testing were automatically scored by the Damerau-Levenshtein Distance algorithm (Damerau 1964, Levenshtein 1966) and later manually rescored where given answer bore lexical similarities to the correct answer but carried an unrelated semantic definition (e.g. “tunna – tunn”, Eng. “barrel – thin”), and thus automatically scored as correct. Conversely, instances where a clearly correct answer was given but obfuscated by various artifacts (e.g. “pärla – pärlaaaa”) were rescored as correct. In total 56 such scores (out of a total of 20,235 answers for .18%) were corrected, and out of these 51 were rescored as incorrect and 5 were rescored as correct (meaning 91% of corrections were negative). All tasks used were accessed on an online platform by personal laptops provided by the school.

Procedure
Data used in the current study were collected as part of a larger study on the memory decay of learned items in relation to the testing effect (see Figure 2). The data relevant for this study were collected over two sessions, spaced one week apart. In the first session each participant performed APM and AoSpan in order to measure WMC and aspects of gF. In the subsequent session one week later, participants underwent the learning phase and were subjected to the to-be-learned material (as indicated in Figure 2). All data collection was carried out within the participants’ scheduled school hours. During data collection participants were situated in prepared classrooms, divided into respective normal class. The exception to this was one instance where 3 classes were situated together in the school canteen (also prepared for the purpose), due to issues of convenience. In all cases the participants were placed in an orderly fashion and spaced in such a way as to prohibit line of sight to neighboring screens. These provisions in conjunction with the random, personalized lists were made in order to minimize risk of cheating. The order and method of test administration, apart from randomized elements, were identical for all participants. In the second session, all participants were first instructed to learn the upcoming word pairs and were then exposed to the 60 Swahili-Swedish word pairs by seeing each intact word pair once in succession. Immediately following this initial exposure, half of the items were tested with feedback and half were restudied. Test items were administered by means of cued-recall testing, presenting the Swahili word followed by a blank space (e.g. Mashua - ______). Each participant was instructed to type the corresponding correct Swedish word and then received immediate feedback in the form of the correct, intact word pair being displayed for 1 second (e.g. Mashua - Båt). Test and study conditions were both performed during the same session in a randomly interspersed order.
**Statistical analysis**

In order to establish if learning significantly improved during the testing condition, a repeated measures ANOVA was conducted on scores across repetitions. To investigate if APM and Aospan are related to the degree of learning, a Pearson correlation analysis was conducted. A repeated measures ANOVA was then performed with APM and Aospan as covariates to determine any main and interaction effects of cognitive measures and repetition. If Mauchly's test indicated violation of sphericity, degrees of freedom was corrected using Greenhouse-Geisser estimates of sphericity.

**Results**

In order to establish that learning had occurred, a repeated measures ANOVA was performed. The ANOVA showed that Mauchly's test indicated that the assumption of sphericity had been violated, $x^2(14) = 528.54$, $p < .001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = .39$). The corrected results revealed that the effect of repetitions was statistically significant; $F(1.93, 330.19) = 797.33$, $MSE = .02$, $p < .001$, $\eta^2_p = .82$. Pairwise comparisons further confirmed that there was significant improvement between each successive practice trial ($p < .001$ on all accounts, see Figure 3).

![Figure 3. Performance on the cued-recall task across repetitions. Error bars represent one standard error of the mean.](image-url)
The scores on APM ranged from 2 to 24 ($M = 13.5, SD = 4.67$), with a skewness of -.02 ($SE = .19$) and kurtosis of -.51 ($SE = .38$). The scores on Aospan ranged from 0 to 73 ($M = 53.51, SD = 13.55$), with a skewness of -1.26 ($SE = .19$) and kurtosis of 1.73 ($SE = .38$). To investigate the relationship between each repetition and the cognitive measures, Pearson correlations were calculated (see Table 1.) Significant positive correlations were found between Aospan and repetition as well as APM and repetition. Aospan showed significant correlations from repetition 2 and onward, with increasing significance and correlational strength between repetitions. APM showed a similar pattern of correlation with repetition, although to a lesser degree, not exhibiting significance until repetition 3. Furthermore Aospan and APM were significantly intercorrelated (Table 1).

Table 1. Pearson correlations for measures of cognitive ability and repetitions.

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Repetition 1</td>
<td>_</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Repetition 2</td>
<td>.88**</td>
<td>_</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Repetition 3</td>
<td>.80**</td>
<td>.91**</td>
<td>_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Repetition 4</td>
<td>.75**</td>
<td>.85**</td>
<td>.93**</td>
<td>_</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Repetition 5</td>
<td>.69**</td>
<td>.81**</td>
<td>.89**</td>
<td>.96**</td>
<td>_</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Repetition 6</td>
<td>.66**</td>
<td>.78**</td>
<td>.87**</td>
<td>.93**</td>
<td>.96**</td>
<td>_</td>
<td></td>
</tr>
<tr>
<td>7. APM</td>
<td>.07</td>
<td>.10</td>
<td>.16*</td>
<td>.19*</td>
<td>.20*</td>
<td>.20*</td>
<td>_</td>
</tr>
<tr>
<td>8. Aospan</td>
<td>.16*</td>
<td>.20*</td>
<td>.23**</td>
<td>.28**</td>
<td>.28**</td>
<td>.31**</td>
<td>.35**</td>
</tr>
</tbody>
</table>

Notes: Aospan = automatic operation span; APM = Raven’s advanced matrices (short form); *p < .05; **p < .01

To examine whether individual differences in cognitive ability influenced the degree of learning, a repeated measures ANOVA was performed with APM and Aospan as covariates. Mauchly’s test showed violation of sphericity; $x^2(14) = 441.5, p < .001$, Greenhouse-Geisser estimates of sphericity ($\epsilon = .40$). There was a main effect for repetition, $F(2.01, 301.85) = 60.34, MSE = .02, p < .001, \eta^2_p = .29$, and a significant interaction effect of Aospan, $F(2.01, 303.66) = 3.61, MSE = .02, p = .03, \eta^2_p = .02$. However, no interaction was observed for APM, $F(2.01, 303.66) = 2.21, MSE = .02, p = .11, \eta^2_p = .01$. For illustrative purposes the participants were grouped, based on Aospan scores, into quartiles of top and bottom 25% and the middle 50% (see Figure 4).
Discussion

The aim of the current study was to investigate whether there is a relationship between individual differences in cognitive abilities and the learning process when using a practice-testing method. First, it was established that test-enhanced learning with feedback improves learning across repetitions in a large sample of students in upper secondary school. It was then shown that there was a significant interaction effect of working memory capacity and repetition, indicating that individuals with higher WMC tended to learn more (see Fig 4). Contrary to what has been indicated by previous studies (Brewer & Unsworth, 2012; Pastötter et al. 2013) general fluid intelligence did not exhibit an interaction effect. Considering the respective differing nature of referenced studies, however, one using composites of multiple gF-measures (Brewer & Unsworth, 2012) and one testing subjects with traumatic brain injury (Pastötter et al. 2013), there is much leeway for diversity of results without these standing in opposition to one another. Working memory was revealed as an important factor in the learning process, as those with higher WMC showed a higher degree of learning.

Historically in testing effect studies, repeated studying has been related to a higher score on the immediate test, whereas repeated testing has been related to a higher score on a delayed test (Roediger & Karpicke 2006a, 2006b). Subjects with higher WMC have performed at a higher level in general. The effectiveness of testing as compared to restudy, however, has been more beneficial for those low in WMC, i.e. showing a stronger testing effect for these subjects (Agarwal et al., 2011). The
results exhibited in the current study add evidence to that finding as even those of lower WMC show a higher degree of learning than what would be expected from a restudying condition given results from previous studies (Agarwal et al., 2011; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a, 2006b). From the significant interaction effect of WMC and repetition, and correlational data (see Table 1), it is seen that the degree of significance grows increasingly stronger by each subsequent repetition. That is, although high and low WMC subjects start out with equal trial scores, those of higher WMC accelerate more rapidly than their lower scoring counterparts. By the sixth and final repetition they have achieved a substantially higher degree of learning, meaning they will likely have a larger knowledge base from which to draw during later retrieval. Though the present study did not follow the “learning to criterion” method used by some earlier studies (e.g. Vaughn & Rawson, 2011) an interesting comparison to that paradigm begs attention. According to the concept of successful retrieval discussed earlier, retrieval is at its most effective after four to five successful instances (Vaughn & Rawson, 2011), and in order to facilitate this number of retrieval successes an average of six to seven repetitions are needed (Rawson & Dunlosky, 2011). Considering the present design of six repetitions and the differences observed between high and low scorers on the WMC task, it is not unthinkable that those of higher WMC are capable of reaching the level of four successful retrievals at an earlier point than their counterparts. This would add evidence to an explanation for why those of higher WMC also perform at a higher level at delayed tests than those of low WMC (Agarwal et al., 2011). Expanding the reasoning of the preceding argument on advantages that higher WMC confers opens up another interesting path of speculation. Connecting the divided WMC groups to the bifurcation model explained earlier in this paper (Kornell et al., 2011) may help inspire new lines of inquiry in future research. Following that the WMC groups are, themselves, bifurcated, there might be a possibility that the WMC construct is related to the split that is proposed to occur due to non-recalled items falling behind those successfully retrieved. If the results reported here constitute an effect of memory strength in general it could follow that higher working memory ability keeps the memory strength of a larger number of items robust enough not to fall under the critical boundary for retrieval (Kornell et al., 2011). As further explained, feedback may help reduce this split between items of high and low memory strength respectively (Kornell et al., 2011). It is thus conceivable that the utilization and form of feedback in this study might have obfuscated any such effect by increasing the memory strength of otherwise weaker items. A similar design to the current one, investigating the learning process in test-enhanced learning and cognitive ability, but with different feedback conditions, could be an interesting venue for further research. As feedback is an instrumental part in maximizing the efficiency of testing (Karpicke & Roediger, 2010), changing the feedback conditions would certainly impact the degree of learning shown here. However, since testing has been shown to be more beneficial for durable learning than studying even when feedback is not used (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a), such a design should not be unthinkable.
Although the results regarding the gF construct presented here may at first seem contradictory or counterintuitive in regards to those discussed in previous studies (Brewer & Unsworth, 2012; Pastötter et al., 2013), it is important to note that no immediate or delayed retention tests were used. The results presented here are solely an investigation across repetitions during the process of test-enhanced learning, which preempts the delayed tests previously investigated (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a, 2006b). Building on the sizeable basis of test effect studies that have reliably demonstrated that the studying condition is linked to higher performance on an immediate test, and that the testing condition is linked to higher performance on a delayed test, we can speculate as to the role of gF (e.g. Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a, 2006b). Using this basis for an assumption of a similar progression here, it is not inconceivable that gF may be a more important factor when retrieving from long-term storage than it is in the short term. Although it is clearly conjecture at this stage, this notion warrants further research.

In conducting this study, some limitations were observed which should be deliberated when designing future studies of a similar nature. Firstly, there was the concern of pre-existing language ability. Though no participant was reported to have any noteworthy knowledge of Swahili at initiation, Swedish as a first language was not stated as a prerequisite. On the other hand, taking into account the robust sample size and effects found, such a provision is likely to have had only an insignificant impact, if any, on the final results. Secondly the measurements chosen for the appraisal of cognitive constructs may also be argued to comprise too narrow a range. The robustness of APM as a measurement of gF notwithstanding (Alderton & Larson, 1990; Chiesi et al., 2012), it is a task specifically focused on visuo-spatial reasoning and it is a distinct possibility that other aspects of gF would yield different results. It is thoroughly possible that a composite of several measurements such as that used by Brewer and Unsworth (2012) could produce a more valid analysis. Likewise, though Aospan was here revealed as significantly related to the learning process, this measure’s breadth is still more limited than such a composite, and other measures of WMC could yield more understanding. Further efforts in this area could expand on the design of the present study and improve on its scope. Apart from widening or altering the selection of cognitive measures used, future studies could take into account immediate and delayed retention tests. Such a design would be better tailored to make comparisons of test-effect strength between WMC groups for example, and would thus not need to rely on previous results to the same extent as this analysis. Furthermore, as this study focused exclusively on the learning process of the testing condition, the cognitive measures and their relation to the study condition yet remains unexplored. Adding retention tests could offer new, important discoveries into the field.

From the many insights that can be gleaned from studies within the testing effect paradigm it has long been apparent that practice testing is a potent tool for efficient promotion of durable learning. However, with the investigation of the relationship between individual differences and test-enhanced learning, to which
the current study adds new evidence, the picture is broadening. Being aware of the best ways and for whom to implement test-enhanced learning in the classroom may help improve the efficiency of teaching. The differences we have here observed in WMC and gF could also help improve teachers’ awareness that students have different pre-existing conditions and capabilities to take into account when interacting with their class. Being cognizant of when and how the schism between high and low performers first appears, a teacher might act early and counteract this development, thus leveling the playing field for students of differing abilities. For instance planning extra opportunities for retrieval practice in the earlier stages of the curriculum may facilitate a sufficient amount of successful retrievals for effective learning to occur. Hence the gap between higher and lower WMC students may be reduced and overall conditions for the class at large improved (Freeman et al., 2014).

In summary, the results presented within this thesis indicate that working memory and general fluid intelligence are both involved in the learning process of test-enhanced learning, although the former to a significantly higher degree than the latter. The implications of these results open up several new promising venues for further research. Armed with the knowledge available within the testing effect paradigm in general, and the specifics currently being uncovered, teachers may have the opportunity to better equip themselves for the classroom.
References

Agarwal, P. K., Rose, N. S., & Roediger, H. L. (2011). Testing levels the playing field for students with lower working memory capacity. Presented at the 51st annual meeting of the psychonomic society, St. Louis, 10th November, 2011.


