Scandinavian Journal of Psychology, 2014, 55, 10–16 DOI: 10.1111/sjop.12093

Cognition and Neurosciences

Strengthening concept learning by repeated testing

CAROLA WIKLUND-HÖRNQVIST,[1] BERT JONSSON[1] and LARS NYBERG[2]

[1]*Department of Psychology, Umeå University, Sweden*
[2]*Departments of Integrative Medical Biology and Radiation Sciences, Umeå University, Sweden*

Wiklund-Hörnqvist, C., Jonsson, B. & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology* 55, 10–16.

The aim of this study was to examine whether repeated testing with feedback benefits learning compared to rereading of introductory psychology key-concepts in an educational context. The testing effect was examined immediately after practice, after 18 days, and at a five-week delay in a sample of undergraduate students (n = 83). The results revealed that repeated testing with feedback significantly enhanced learning compared to rereading at all delays, demonstrating that repeated retrieval enhances retention compared to repeated encoding in the short- and the long-term. In addition, the effect of repeated testing was beneficial for students irrespectively of working memory capacity. It is argued that teaching methods involving repeated retrieval are important to consider by the educational system.

Key words: Test-enhanced learning, memory, retrieval practice, long-term retention, feedback.

Carola Wiklund-Hörnqvist, *Department of Psychology*, *Umeå University*, *Sweden*. E-mail: Carola.WiklundHornkvist@psy.umu.se

INTRODUCTION

Traditionally, in educational settings, tests are used as tools for evaluating students' knowledge and assigning grades. Another aspect of testing that has largely been neglected by educationalists is its potential to serve as a way of facilitating learning (Bangert-Drowns, Kulik & Kulik, 1991; Butler & Roediger, 2007; Gates, 1917; Glover, 1989; McDaniel, Anderson, Derbish & Morisette, 2007; Spitzer 1939).

Findings from empirical memory studies have shown that taking repeated tests before the administration of a final retention test improve the performance compared to traditional restudy of materials, particularly on delayed recall tests (Roediger & Karpicke, 2006, 2006b; Roediger & Butler, 2011; Weinstein, McDermott & Roediger, 2010). One explanation for this beneficial effect is that repeated testing promotes active retrieval of information from memory and offers opportunities for re-encoding of information, whereas traditional restudy-techniques to a large extent rely on repeated encoding (Karpicke & Roediger, 2008). The improvement in subsequent performance after taking one or several tests is known as the *testing effect* (Kang, McDermott & Roediger, 2007; Karpicke & Roediger 2008; Martinez & Martinez, 1992). This effect has been found to be stable across different kinds of materials, such as paired-associate learning (Carrier & Pashler, 1992), facts (Carpenter, Pashler & Cepeda, 2009; McDaniel, Agarwal, Huelser, McDermott & Roediger, 2011), prose passages (Agarwal, Karpicke, Kang, Roediger & McDermott, 2008; Roediger & Karpicke, 2006b), statistics (Lyle & Crawford, 2011) and learning of skills (Kromann, Jensen & Ringsted, 2009).

The majority of previous studies on the testing effect has been conducted in laboratory settings with materials unrelated to real educational purposes (see Roediger and Karpicke, 2006, 2006b for a review). Moreover, few studies have examined the testing effect in actual learning environments with ecological materials and a more realistic retention interval from an educational perspective, that is, longer than one week (see supplementary material in Rawson & Dunlosky, 2011 for a review). Carpenter *et al*. (2009) investigated the testing effect by assessment of low-stake quizzes in an eighth-grade US history class. After a 9-month delay, the results revealed that significantly more items were recalled that initially had been tested with feedback compared to items that had been studied only or not reviewed. McDaniel, *et al*. (2007) examined the testing effect during a web-based college-course on brain and behaviour. Students either took weekly quizzes with feedback or reread some critical facts related to the weekly reading assignment. The quizzes were either short-answer (SA) questions or multiple-choice (MC) questions. Approximately five weeks after the last test, students took a final cumulative MC test. The result revealed that prior quizzing, particularly in the form of SA questions, resulted in higher scores at the final test compared to rereading (McDaniel *et al*., 2007). Thus, the results of a few studies suggest that the testing effect has the potential to facilitate learning in real educational settings. However, it has been stressed that there is a need for additional studies of test-enhanced learning using educationally relevant materials during the progression of a course (Newcombe, 2002; Rohrer & Pashler, 2010).

Knowledge of key concepts is important for students in order to gain a better conceptual understanding of the topic at hand. Several studies have shown that learning the meaning of key words is effective for reading comprehension (e.g., Beck, Perfetti & McKeown, 1982; McDaniel & Pressley, 1989). Textbook chapters used in educational settings include important key concepts that are of relevance for the specific topic. Further, lectures are aimed to help the students to orient and get an overview on the topic under study, based on those key concepts. Thus, combining a lecture with computer-assisted learning of key concepts might further improve students' knowledge level. On the basis of these prior studies, the current study aimed to examine whether repeated testing with feedback, using SA questions, promotes long-term retention relative to rereading of key concepts during the progression of an introductory university course.

A key factor in learning studies is the duration of the retention interval. Many studies related to the testing effect have applied the first retention test after two days (Thompson, Wenger & Bartling, 1978) or seven days (see Rawson & Dunlosky, 2011; Roediger and Karpicke, 2006, 2006b; for a review). A typical finding is that restudy of material is beneficial for the performance on immediate tests compared to repeated testing, whereas the opposite pattern is apparent after longer delays (Roediger and Karpicke, 2006, 2006b; Thompson *et al*., 1978; Wheeler, Ewers & Buonanno, 2003). In contrast to these findings, Carpenter, Pashler, Wixted and Vul (2008) showed in three independent experiments that testing with correct answer feedback was superior to restudy both for a five-minute and a six-week delay, and the effect increased as a function of the number of practice trials. Carpenter *et al*. (2008) argued that the immediate testing effect might be related to the inclusion of feedback, which differs from the majority of past studies (see Roediger and Karpicke, 2006, 2006b).

The results in the Carpenter *et al*. (2008) study are in line with those from an early study by Thompson *et al*. (1978), in which it was found that re-exposure (i.e. feedback) of items not recalled improved immediate performance compared to study or test without feedback. In one of their experiments they added a third condition where subjects studied a word list once, were tested by recall, and subsequently re-exposed to those words they failed to recall. The findings, on the immediate test as well as two days later, revealed that more words were retained in the recall plus re-exposure condition compared to both the study and test conditions (Thompson *et al*., 1978). Despite claims that test-restudy practice produces durable learning (Butler, Karpicke & Roediger, 2008; Cull, 2000), the results of the Carpenter *et al*. (2008) study also revealed a substantial drop in performance between the 14-day delay and 42-day delay. To further investigate the robustness of testing as a method that reduces forgetting, the present study included additional follow-up sessions 18 days and five weeks after the initial practice.

Testing promotes learning in the absence of feedback (Carpenter, 2009; Karpicke & Roediger, 2008), but, as pointed out above, providing feedback seems to further boost learning (Brosvic & Epstein, 2007; Butler *et al*., 2008; Kang *et al*., 2007; Roediger & Butler, 2011; Shute, 2008). Feedback may serve two important functions. First, feedback may prevent retrieval failures from being repeated. That is, including feedback may serve as an additional learning opportunity that reduces repeated retrieval failures. Second, including feedback in the form of correct answers prevents erroneous learning to occur. That is, if an item has been erroneously recalled and feedback is not provided to correct the error, that error response has a tendency to be well-stored and repeated later (Roediger & Marsh, 2005). Fazio, Huelser, Johnson and Marsh (2010) examined how different kinds of feedback influenced learning of non-fiction passages. Over three experiments they had participants read passages followed by either no feedback, right/wrong feedback, or correct-answer feedback. Their result revealed that providing feedback in the form of correct answers was most beneficial (Fazio *et al*., 2010). This is in line with previous research that correct-answer feedback both facilitated learning and improved retention of word-pairs one week later (Pashler, Cepeda, Wixted & Rohrer, 2005). Another important aspect is related to the timing of feedback. If the aim is to support learning, students should receive feedback immediately after their response (Brosvic, Epstein, Cook & Dihoff, 2005; Kulik & Kulik, 1988). In addition, as highlighted by Butler *et al*. (2007), a crucial aspect of feedback is that it ensures that students process the feedback regardless of their success in recall (Butler *et al*., 2007). The current study included immediate feedback in the form of correct answers regardless of recall success. In line with prior studies combining testing with feedback (Carpenter *et al*., 2008; McDaniel & Fisher, 1991; Thompson *et al*., 1978) it was hypothesized that testing with feedback should lead to better performance compared to the restudy condition both in the immediate test and at the delayed tests.

A secondary purpose of the current study was to consider how individual differences in working memory capacity (WMC) relate to learning ability (Alloway & Alloway, 2010; Yuan, Steedle, Shavelson, Alonzo & Opezzo, 2006). This issue has not been examined extensively within the testing-effect domain, and the limited prior findings show mixed results. Agarwal, Rose and Roediger (2010) found support for test-enhanced learning being particularly beneficial for individuals with lower WMC, whereas others have failed to replicate this observation (Brewer & Unsworth, 2012). These mixed results point to the importance of further considering individual differences in WMC when examining the effectiveness of test-enhanced learning, as this might clarify the educational significance of applying different learning methods in the classroom (Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013; Tse & Pu, 2012).

METHODS

Participants

Eighty-three undergraduate students registered on a cognitive psychology course participated in the study. The age ranged from 19–44 years (M = 23.8, SD = 3.94). Participation was voluntary and rewarded with 300 SEK. The experiment was administered as an addition to the curriculum. Written informed consent was obtained in accordance with the Declaration of Helsinki, and approved by the Regional Ethical Review Board, Sweden.

Materials and design

The materials used consisted of 57 key concepts from three topics in the assigned cognitive-psychology curriculum; *attention*, *memory*, and *perception* (see Appendix for examples). Each topic contained 19 key concepts, respectively. The key concepts examined were all covered in the assigned readings before the topic lecture (Galotti, 2008). Immediately after the topic lecture, the participants arrived at a computer laboratory. Participants were randomly assigned to either the repeated testing group with feedback (ST_{fb}, n = 43) or the restudy group (SS, n = 40). Each of the three learning occasions, one for each topic, included a learning phase followed by an immediate test (five-minute delay). Learning was assessed by means of a test at three different time-points; immediate, an average of 18-days later (range 15–20 days), and in a subsample at a five-week delay. For the five-week delay, 63 (ST_{fb}, n = 33, SS n = 30) of the participants were re-recruited. The relatively high number of students not participating at the five week delay was due to the fact that some students had finished their courses and left the campus. The software used for the experiment was E-prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). All items were randomly presented in the center of the screen, both during the learning phase and during the following tests.

Working memory capacity

Working memory capacity (WMC) was assessed with the Automated Operation Span (Aospan; for full task details see Unsworth, Heitz, Schrock & Engle, 2005), which requires participants to remember a series of letters while performing a concurrent task (i.e. a complex working memory task). The Aospan shows good internal consistency (0.78) and test–retest reliability (0.83; Unsworth *et al*., 2005). This task was administered immediately after the final delayed test was completed. In the Aospan, participants judge whether a math equation yields a true or false answer and then they see a letter to be remembered for later recall. After a series of equation-letter trials, with the set-sizes ranging from 3–7 items, subjects are asked to recall the letters in the correct order. Individuals are encouraged to keep math accuracy above 85% and feedback for math accuracy and letter response are provided. The dependent variable used in the analysis was the total number of correct items in the correct position (Unsworth *et al*., 2005).

Procedure

Learning phase. Immediately after the topic lecture, the participants arrived at a computer laboratory. Each participant was seated in front of a computer. To familiarize the participants with the to-be-learned material, all participants studied the key concepts at a paper for four minutes before the diverged learning phase started (cf., Experiment 1, Rawson & Dunlosky, 2011). Following these common learning phases (lecture and familiarization), the two groups were assigned to different experimental procedures namely, either rereading the facts (restudy group; SS) or taking a test with feedback (test group; ST_{fb}) six consecutive times. For the SS-group, a key-concept was presented (15 sec.), and the instruction only required subjects to study it (e.g., "*The primacy effect is the higher level of retention of information presented at the beginning of a list*"). For the ST_{fb}-group, a key-concept leaving out the keyword was presented (15 sec.) and subjects were requested to type in the correct answer at a blank screen (10 sec). This was followed by feedback in the form of the correct answer (5 sec.). For example; "*The improvement in retention of information presented at the beginning of a list?*" The participants were requested to type in the correct answer at a blank screen ("*primacy effect*") followed by feedback; "*primacy effect*" (a presentation of the correct response). For both groups, each learning phase contained six learning trials, where each learning session contained 19 randomly presented key concepts within the current topic. There was a two-minute break after the third learning session. During the break, all participants filled in a questionnaire that prompted whether they had read the assigned readings and attended the associated topic lecture. The level of difficulty and the learning procedure were held constant across occasions. Intentional learning instructions were given.

Immediate test. Learning in both groups was evaluated by a test administered after a five-minute break following the final learning session ("immediate test"). During the five-minute break, all the participants answered general questions concerning perceived difficulty and judgment of learning, which prevented the possibility of additional covert retrieval. In the immediate test, participants were presented with a fact (15 sec.) and were requested to type in the correct answer (10 sec.). After they finished the test, the participants were thanked and reminded of the follow-up session. No information was provided that the students were to take the same tests at the remaining two occasions.

Delayed tests. The average retention intervals for the delayed tests were 18 days and five weeks after the initial learning phase, respectively. On the delayed tests, participants completed all three topics within the same session using the same material, procedure, and order as in the immediate test.

Statistical analyses. Statistical analyses were conducted using SPSS 18. The effects on performance of different learning methods were assessed by two different analyses. First, a mixed-model analysis of variance (ANOVA) investigated the effects of topics, time and group. As the five-week delay involved a subsample of the full sample independent *t*-tests were conducted to investigate sample differences.[1] Partial eta-squared values indexed effect sizes for the *F*-values. Second, to explore whether individual differences in working memory capacity influenced performance in relation to the intervention, a series of simultaneous regression analyses were conducted at each delay. To be able to examine a moderating effect of working memory on performances we included an interaction term for WMC × Group (McClelland & Judd, 1993). Regression analyses were conducted with the following predictors; Group (SS, ST_{fb}) as a categorical variable, WMC as a continuous variable, and the interaction term WMC × Group. Before running the regression analysis we followed the suggestion by Aiken and West (1991) that continuous independent variables included in an interaction term should be mean centered in order to decrease collinearity. The continuous WMC as an independent variable was mean centered before entered in the regression analysis. To control for multiple tests, the alpha level 0.05 was Bonferroni corrected by the number of independent analyses (i.e., $p < 0.01$).

RESULTS

During the learning phase there was a gradual improvement in performance across the practice trials (Fig. 1). A one-way repeated (ANOVA)[2] revealed a main effect of trial, F (3.22, 109.38) = 374.22, $MSE = 0.06$, $p < 0.001$, $\eta_p^2 = 0.92$, and pairwise comparisons confirmed that there was a significant improvement between each successive practice trial (all $p's < 0.01$).

A mixed-model ANOVA with time-point *(immediate*, *18-days delay)* and topic (*memory*, *attention*, *perception*) as within-subject factors, and group (SS, ST_{fb}) as between-subject factor was used to analyze whether topics interacted with group. The analysis revealed no significant interaction effects between topic and group, F (2,162) = 1.91, $MSE = 0.04$, $p = 0.15$, $\eta_p^2 = 0.02$. Also, no significant interaction effect between topic and group was shown at the 5-week delay, F (2,122) = 0.55, $MSE = 0.01$, $p = 0.58$, $\eta_p^2 = 0.01$. Therefore, the topics (*memory*, *attention*, *perception*) were collapsed into one dependent measure that was used in the subsequent analyses. The collapsed mean proportions of performance for the two groups at the different time-points are shown in Figure 2.

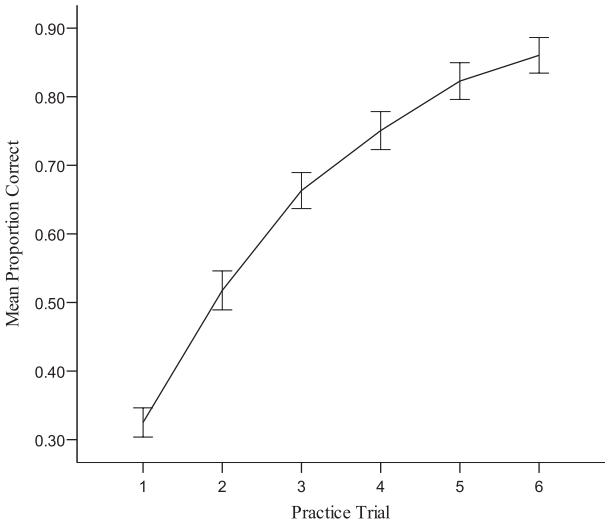

Fig. 1. Mean proportion of correct responses for the STfb group as a function of increased number of learning trials. Error bars represents ± 1 standard error of the mean.

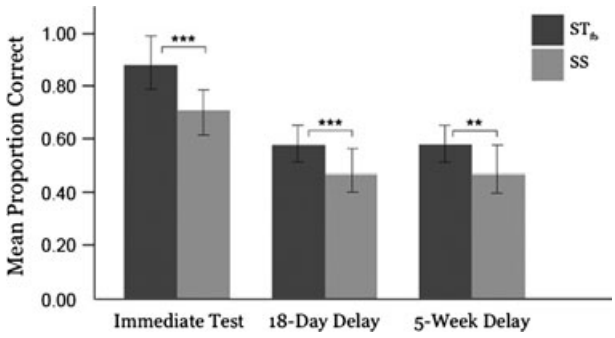

Fig. 2. The mean proportion of correct responses for the STfb and SS group for the three time-points. Error bars represents ± 1 standard error of the mean.

A mixed model ANOVA with group (SS, ST_{fb}) as between-subjects factor and time of testing *(immediate, 18-days delay)* as within-subject factor was conducted. Main effects of time of testing [F (1,81) = 451.41, MSE = 0.06, $p < 0.001$, $\eta_p^2 = 0.85$] and group [F (1,81) = 26.52, MSE = 0.02, $p < 0.001$, $\eta_p^2 = 0.25$] were qualified by an interaction between time of testing and group, F (1,81) = 8.03, $p < 0.01$, $\eta_p^2 = 0.09$. The ANOVA revealed that both groups performed at a higher level on the immediate test compared to the 18-day delay test, and that performance was in favor of the ST_{fb} group (see Fig. 2). As can be seen in Figure 2, the absolute rate of forgetting between the immediate test and the 18-day delay is greater for the ST_{fb} group (29.6%) compared to the SS-group (22.6%). In line with previous research (Roediger and Karpicke, 2006, 2006b), we calculated forgetting by a proportional measure [(initial recall – delayed recall)/initial recall]. Proportional forgetting showed that the rate of forgetting was quite similar in both groups, ST_{fb} = 33.6% and the SS group = 33.1%.

As the five-week delayed test involved a subsample (n = 63), a univariate analysis of variance was conducted to examine the effects of long-term retention between the ST_{fb} and the SS-group. There was a significant difference between the groups, F (1,61) = 10.55, MSE = 0.03, $p < 0.01$, $\eta_p^2 = 0.15$ indicating that the ST_f -group performed better than the SS-group (see Fig. 2).

The influence of working memory capacity

As the second purpose was to explore the role of working memory capacity (WMC) as a predictor of task performance, regression analyses were conducted for each time-point. Before examining the influences of working memory, a univariate analysis of variance (ANOVA) was conducted to control that WMC was comparable across the two groups [ST_{fb}-group (M = 40.81, SD = 16.61) SS-group (M = 36.50, SD = 19.18)]. There was no significant difference in WMC between the groups, F (1, 80) = 1.49, MSE = 162.61, p = 0.23. The intercorrelations among the memory performance at the different time-points and WMC can be seen in Table 1.

Table 1. *Intercorrelations among memory performance at the three different time-points and working memory capacity as a function of group.*

Measure	1	2	3	4
1. Immediate test	-	0.75**	0.73**	0.06
2. 18-day delay	0.72**	-	0.93**	0.29
3. 5-week delay	0.75**	0.93**	-	0.36*
4. WMC	0.20	0.17	0.20	-

Notes: Pearson intercorrelations (two-tailed) for the STfb group (n = 43) are presented above the diagonal, and intercorrelations for the SS-group (n = 40) are presented below the diagonal. *$p < 0.05$ **$p < 0.01$.

For the regression analyses the predictors were entered simultaneously using the enter method. Table 2 summarizes the findings of the regression analyses. For the immediate test, Group, WMC, and the interaction term Group × WMC as predictors significantly explained almost 30% of the variance in performance, adjusted R^2 = 0.269, F (3,78) = 10.94, $p < 0.0001$. For the 18-day delay, the model was significant, F (3,78) = 6.72, $p < 0.0001$ and explained 20.5% of the variance in performance, adjusted R^2 = 0.175. Finally, for the five-weeks delay the model remained significant, and accounted for significantly 21.8% of the variance in performance, adjusted R^2 = 0.178, F(3,58) = 5.40, $p < 0.002$. Table 2 gives information about the predictor variables entered into the model. As can be seen in Table 2, group significantly predicted performance at all time-points whereas WMC did not. Figure 3 shows the relationship between performance and WMC at the different time-points for each group. The advantage of the test group was seen across the range of WMC at all retention intervals.

DISCUSSION

The present study tested the hypothesis that testing with feedback will lead to better performance at longer as well as shorter retention intervals compared to restudy of key concepts. We

Table 2. *Summary of the simultaneous regression analyses predicting the performance at the three time-points with Working Memory Capacity and Group as predictors.*

Time	Predictors	B	SE (B)	β	t	P
Immediate test	Working memory capacity	–0.001	0.003	–0.073	–0.230	0.819
	group	**–0.188**	**0.035**	**–0.513**	**–5.36**	**0.0001**
	Interaction: WMC × group	0.001	0.002	0.205	0.651	0.517
18-day delay	Working memory capacity	0.004	0.003	0.431	1.28	0.202
	group	**–0.118**	**0.032**	**–0.370**	**–3.64**	**0.0001**
	Interaction: WMC × group	–0.001	0.002	–0.232	–0.693	0.490
5-week delay	Working memory capacity	0.004	0.004	0.454	1.25	0.215
	group	**–0.118**	**0.040**	**–0.346**	**–2.95**	**0.005**
	Interaction: WMC × group	–0.001	0.002	–0.202	–0.560	0.578

Note: Significant outcomes are indicated in bold.

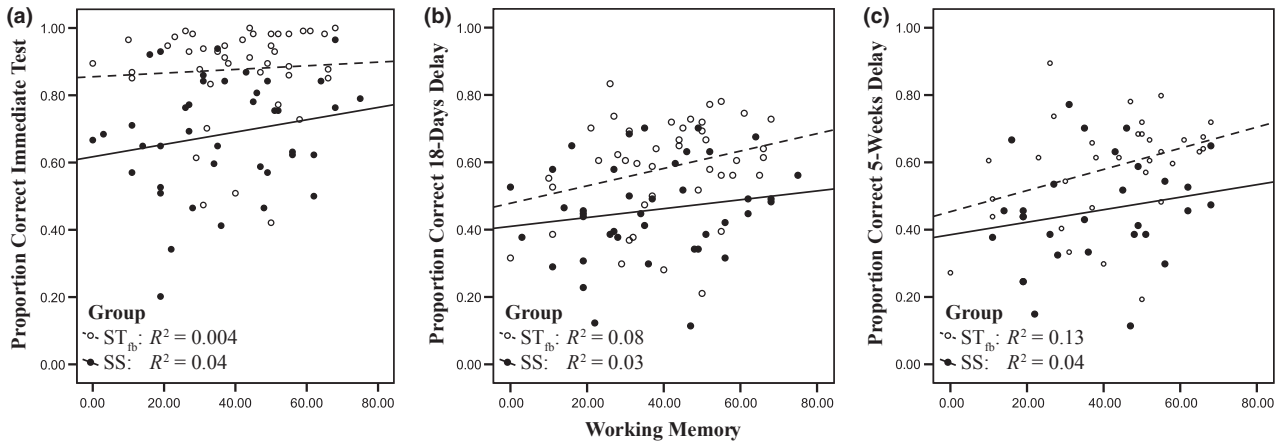

Fig. 3. Scatterplots showing the relationship between working memory capacity and mean proportion correct across time, a) immediate test, b) 18-day delay, c) 5-week delay for the different learning conditions.

found that testing with feedback was superior to the restudy condition and that the difference was sustained over all three timepoints analyzed.

As expected, the testing with feedback condition significantly outperformed the restudy condition also at the immediate test. As far as we know, this is the first study that demonstrates an immediate testing effect using course material during the progression of an on-going course. Carpenter *et al*. (2008) as well as Kornell, Bjork and Garcia (2011) have shown immediate benefits of the testing effect that are in line with the present results. They, too, used an experimental set-up in which feedback was given, although they did not use educationally relevant material integrated in an on-going course as we did. Kornell *et al*. (2011) argued that test without feedback creates a bifurcated item distribution in which items that are retrieved are high in memory strength, and items that are not retrieved are low in memory strength. When participants are provided with feedback (i.e. items are restudied), memory strength becomes high enough to surpass a threshold and the information therefore becomes recallable, hence facilitating short-term retention and preventing erroneous learning to occur (Kornell *et al*., 2011). The results from the immediate test in the current study are in line with this suggestion. The present study included feedback independently of whether the response was correct or not, which possibly resulted in items being above the threshold at the immediate test (i.e., recallable).

In addition, the tendency found that tests reduces forgetting (Roediger and Karpicke, 2006, 2006b) was not found in the current study when comparing with the restudy condition. The proportional forgetting between the immediate test and the 18-day delay test was similar in both groups. Why? One possible explanation can be related to the feedback component, which was in the form of correct answer. Speculatively, some of the items might have been recalled by the support of feedback, but those items were close to the recall threshold (Kornell *et al*., 2011) which made them recallable at the immediate test but forgotten 18 days later. Hence, a suggestion for future studies is to use functional magnetic resonance imaging techniques to investigate the neural mechanisms related to the changes in memory strength as proposed by Kornell *et al*. (2011). In general, the neurocognitive mechanisms that underlie the benefits of retrieval practice are not well understood (but see Eriksson, Kalpouzos & Nyberg, 2011).The results from the test at the 18-day delay, correspond well with the results from previous studies (Roediger & Karpicke, 2006b; experiment 2; Carpenter *et al*., 2008; experiment 2). In contrast to the results in the Carpenter *et al*. (2008; Experiments 1 & 2) study, which showed a decline in performance between later retention tests, the results of the present study revealed that the amount of information retained at the 18-day test and at the five weeks delayed test were comparable in both groups.

The interest in the testing effect has partly been driven by its educational relevance. As indicated by the performance at the first test during the learning phase, the result clearly indicates that the students' initial knowledge about key concepts is low following a lecture. However, when combined with computer-assisted learning the knowledge level improves, particularly when practicing retrieval with feedback. Historically, the idea that retrieving information from memory can increase retention is not new (see Roediger and Karpicke, 2006, 2006b for a review). The benefit of feedback-based learning for retention and shaping of responses is well established within the operant conditioning paradigm (Skinner, 1953, 1958). In the current study, the item-by-item feedback could be considered as a 'reinforcer' for learning (Skinner, 1953). However, even if testing with feedback might share some common features with the operant conditioning paradigm it should not be considered as an analogue (Kulhavy & Stock, 1989). The feedback in terms of correct answers, as in the present study, provides a repetitive or corrective feedback which is distinct from the dichotomous verification found in the "yes-no" or "right-wrong" reinforcement paradigm (Kulhavy & Stock, 1989), Recent research on feedback has also suggested that including immediate feedback is particularly important when the initial knowledge level is expected to be low and when the material is considered as complex (Shute, 2008). Thus, despite the longstanding knowledge of the benefits of test-enhanced learning and feedback within psychology, it is neither well known nor commonly implicated as a method to promote learning by students (Karpicke, Butler & Roediger, 2009). Learning is a cumulative process and a main challenge for educators is to apply methods that increase the ability for students to store and retain relevant information over long periods of time

(i.e. durable learning). A practical implication of our study is that testing with feedback produces durable learning.

A second purpose of this study was to examine how individual differences in working memory capacity (WMC) influenced the ability to benefit from the testing manipulation. The observed non-significance for WMC as a predictor provides no support for the Agarwal *et al*. (2010) suggestion that testing is especially beneficial for students with lower WMC. Rather, the findings of the present study suggest that repeated testing is beneficial compared to restudy regardless of WMC. Furthermore, across time, our results provided only weak and non-significant support for the "rich get richer" notion, which predicts that individuals with high WMC benefit the most from the testing effect (e.g., Rapport, Brines, Theisen & Axelrod, 1997). It should be stressed, though, that the sample in this study consisted of a fairly homogenous group of undergraduate university students, and may not be representative for more marked interindividual WMC variability in more heterogeneous samples. Longitudinal investigation of larger and more heterogeneous groups with regard to the WMC will be needed to further test the relation between working memory capacity and the magnitude of the testing effect.

A limitation with the present study is that it might be argued that the longer exposure duration of the items in the test with feedback condition contributed to the observed group difference. While this cannot be ruled out on basis of the present data, we note that the results from previous studies indicate that additional study opportunities and study time does not eliminate the effect (Carpenter *et al*., 2008; Karpicke & Roediger, 2008; McDaniel *et al*., 2007). In addition, the participants could be considered as relatively experienced learners, and it remains to be examined whether the present findings generalize to other populations.

In conclusion, by being based on educationally relevant materials the results of the present study extend our knowledge on the effectiveness of testing as a way to strengthen learning. The results show that test-enhanced learning is beneficial over both the short- and long term, at least when applied to the learning of key concepts integrated in an ongoing university course. The study also provides knowledge about how repeated testing and repeated study are associated with working memory capacity, and indicates that testing is beneficial irrespectively of individual differences in working memory capacity. Thus, our results generate novel information on the effectiveness of testing as a learning method and contribute to bridge the current gap between cognitive psychology and educational practice.

This research was funded by Umeå School of Education, Umeå University. We would like to thank the students for participating.

NOTES

[1] To examine whether there was a difference in performance for the remaining sample versus dropouts, we calculated the difference in performance in each group between the immediate test and the 18-day delay between the remaining sample and dropouts from the five-week delay. No significant differences were found *in* the ST_{fb} group,t (41) = 0.61, p = 0.55 (Cohens d = 0.23) or in the SS-group, t (38) = 0.36, p = 0.72 (Cohens d = 0.14). Thus, dropout should not have influenced the observed findings at the 5-week delay.

[2] Mauchlys's test indicated that the assumption of sphericity had been violated, $\chi 2$ (14) = 48.53, $p < 0.001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = 0.65).

REFERENCES

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alloway, T. P. & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*, 20–29.

Agarwal, P. J., Karpicke, J. D., Kang, S. H., Roediger, H. L. & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

Agarwal, P.K., Rose, N.S. & Roediger, H.L. (2010). Testing levels the playing field for students with lower working memory capacity. Presented at the 51st annual meeting of the psychonomic society, St. Louis, November 10, 2011.

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

Beck, I. L., Perfetti, C. A. & McKeown, M. G. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*, 506–521.

Brewer, G. A. & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*, 407–415.

Brosvic, G. M. & Epstein, M. L. (2007). Enhancing learning in the introductory course. *The Psychological Record*, *57*, 391–408.

Brosvic, G. M., Epstein, M. L., Cook, M. J. & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom. *The Psychological Record*, *55*, 401–418.

Butler, A .C., Karpicke, J. D. & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*, 273–281.

Butler, A. C., Karpicke, J. D. & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback enhances retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *34*, 918–928.

Butler, A. C. & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527.

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642.

Carpenter, S. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval, *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*, 1563–1569.

Carpenter, S. K., Pashler, H., Wixted, J. T. & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition*, *36*, 438–448.

Carpenter, S. K., Pashler, H. & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology*, *23*, 760–771.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58.

Eriksson, J., Kalpouzos, G. & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters*, *505*, 36–40.

Fazio, L. K., Huelser, B. J., Johnson, A. & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory*, *18*, 335–350.

Galotti, K. M. (2008). *Cognitive psychology in and out of the laboratory* (4th edn). Belmont, CA: Thomson Wadsworth.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*, 1–104.
Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.
Kang, S. H. K., McDermott, K. B. & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.
Karpicke, J. D. & Roediger, H. L. (2008). *The critical importance of retrieval for learning, Science*, *15*, 966–968.
Karpicke, J. D., Butler, A. C. & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when theystudy on their own? *Memory*, *17*, 471–479.
Kornell, N., Bjork, R. A. & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*, 85–97.
Kromann, C. B., Jensen, M. L. & Ringsted, C. (2009). The effects of testing on skills learning. *Medical Education*, *43*, 21–27.
Kulhavy, R. W. & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology*, *1*, 279–308.
Kulik, J. A. & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79–97.
Lyle, K. B. & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*, 94–97.
McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B. & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399–414.
McDaniel, M. A., Anderson, J. L., Derbish, M. H. & Morrisette, N. (2007). Testing the testing effect in classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.
McDaniel, M. A. & Fisher, R.P. (1991). Test and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.
McDaniel, M. A. & Pressley, M. (1989). Keyword and context instruction of new vocabulary meanings: Effects on text comprehension and memory. *Journal of Educational Psychology*, *81*, 204–213.
McClelland, G. H. & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390.
Martinez, J. G. R. & Martinez, N. C. (1992). Re-examining repeated testing and teacher effects in a remedial mathematics course. *British Journal of Educational Psychology*, *62*, 356–363.
Newcombe, N. S. (2002). Biology is to medicine as psychology is to education: True or false? *New Directions for Teaching and Learning*, *89*, 9–18.
Pashler, H., Cepeda, N.J., Wixted, J.T. & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8.
Rapport, L. J., Brines, D. B., Theisen, M. E. & Axelrod, B. N. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist*, *11*, 375–380.
Rawson, K. A. & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*, 283–302.
Roediger, H.L. & Butler, A.C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science*, *15*, 20–27.
Roediger, H. L. & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
Roediger, H.L. & Karpicke, J.D. (2006b). Test-enhanced learning. Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.
Roediger, H. L. & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology*, *31*, 1155–1159.
Rohrer, D. & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*, 406–411.
Schneider, W., Eschman, A. & Zuccolotto, A. (2002). *E–prime user's guide*. Pittsburgh, PA: Psychology Software Tools Inc.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189.
Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
Skinner, B. F. (1958). Teaching machines. *Science*, *128*, 969–977.
Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.
Thompson, C. P., Wenger, S. K. & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology*, *4*, 210–221.
Tse, C-S. & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*, 253–264.
Unsworth, N., Heitz, R. P., Schrock, J. C. & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.
Wheeler, M. A., Ewers, M. & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.
Weinstein, Y., McDermott, K. B. & Roediger, H. L. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*, 308–316.
Yuan, K., Steedle, J., Shavelson, R., Alonzo, A. & Oppezo, M. (2006). Working memory, fluid intelligence, and science learning. *Educational Research Review*, *1*, 83–98.

Received 22 February 2013, accepted 27 September 2013

APPENDIX EXAMPLES OF KEY CONCEPTS USED IN THE CURRENT STUDY

Attention

SS: Feature integration theory is a proposal that perception of familiar stimuli occurs in two stages. The first, automatic, stage involves the perception of object features. The second, attentional, stage involves the integration and unification of those features.
ST_{fb}: A proposal that perception of familiar stimuli occurs in two stages. The first, automatic, stage involves the perception of object features. The second, attentional, stage involves the integration and unification of those features?
Correct Answer: Feature integration theory

Memory

SS: Interference is a hypothesized process of forgetting in which material is thought to be buried or otherwise displaced by other information but still exists somewhere in a memory store.
ST_{fb}: A hypothesized process of forgetting in which material is thought to be buried or otherwise displaced by other information but still exists somewhere in a memory store?
Correct Answer: Interference

Perception

SS: Word superiority effect is the phenomenon that single letters are more quickly identified in the context of words than they are when presented alone or in the context of random letters.
ST_{fb}: The phenomenon that single letters are more quickly identified in the context of words than they are when presented alone or in the context of random letters?
Correct Answer: Word superiority effect