The relationship between different forms of retrieval practice and perceived mental effort, and how this relates to the performance of learners

Master thesis

University of Utrecht, the Netherlands

2018-2019

Student name: Michelle Esmee Janssen

Student no. 6229980

Thesis Supervisor: Noortje Coppens

Second reader: Gemma Corbalan

Date: 10-06-2019

Word count: 6462

Preface

In front of you, my thesis 'The relationship between different forms of retrieval practice and perceived mental effort, and how this relates to the performance of learners'. When I started doing research for this thesis, I knew nothing about retrieval practice at all. If only I had known this before, then I would certainly have studied more effectively all these years.

I am extremely thankful to my supervisor Noortje Coppens, for your listening ear, it helped me stay confident during the process, and even more important, at a time I needed it the most. Gemma Corbalan, thank you for the constructive feedback on my research plan. Thanks to my friends, for your enthusiasm, insatiable curiosity and wine. And specially, Rosan, Marianne, and Aniek, if someone looks in the dictionary for the definition of the word 'friend', they should be seeing your names. Thank you for all the support, feedback, and the helpful and critical questions as well. Thanks to all the participants, without your participation and enthusiasm during the experiment, it would not have been possible for me to write this thesis. And lastly, I want to thank my parents, for all the mental support, and of course you were right I did it!

Michelle Esmee Janssen

Juni, 2019

Abstract

Testing has been established as an effective strategy for facilitating learning. However, most learners are not aware of its benefits. Research often shows that test formats have different effects on learning and memory also known as testing effect. This can be explained by the effortful retrieval hypothesis. The effortful retrieval hypothesis states that more effortful and difficult retrieval during learning improves one's performance. It is generally assumed that different test formats yield different levels of effort. Difficult successful retrievals require more mental effort, which results in enhanced memory. Nevertheless, only one study tested the effortful retrieval hypothesis and compared perceived mental effort on different test formats using a between-subject design. Unfortunately, this study did not find any differences between the conditions in perceived mental effort during the learning phase and retention after one week. For this reason, the present study tested the effortful retrieval hypothesis and compared perceived mental effort on different test formats using a within-subject design. Results show that retrieval practice is a superior strategy of learning in comparison to restudy. Furthermore, mental effort with regard to the different learning conditions differs, free recall yields the highest mental effort, followed by cued recall, followed by recognition during learning.

Keywords: testing effect, retrieval practice, learning, retention

The relationship between different forms of retrieval practice and perceived mental effort, and how this relates to the performance of learners

Testing has been frequently used in educational settings to assess to what extent the learner has mastered the acquired knowledge (Roediger & Butler, 2010). However, research shows that testing can also improve learning (Rowland, 2014). Taking a test on studied material (i.e., retrieval practice), enhances the retention of successfully retrieved information compared to restudying the material, which is the so-called testing effect (Roediger & Karpicke, 2006). During the use of retrieval practice, learners have to recall the learned material without having it in front of them, after which they investigate how much they remember (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). Retrieval practice has been proven to be effective and to enhance learning (Bouwmeester & Verkoeijen, 2011; Carpenter, Pashler, & Cepeda, 2009; Goossens, Camp, Verkoeijen, & Tabbers, 2013; Toppino & Cohen, 2014). The study of Goossens et al. (2013) investigated whether there was a positive effect of retrieval practice of a word list in comparison to restudying the learning material on performance on the final test after one week. Indeed, they found a positive effect of recalling the words learned by retrieval practice on the final test after one week.

Nevertheless, a study of Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) showed that ineffective learning strategies such as rereading are still often used by learners. One explanation for the limited use of retrieval practice is a lack of awareness of its benefits (Karpicke, Butler, & Roediger, 2009). The study by Karpicke et al. (2009) showed that only 11% of the students used retrieval practice as a study technique. Furthermore, a study by Kornell and Bjork (2007) showed 68% of the students quizzed themselves while self-studying. An at least remarkable finding is that only 18% of the same group of students indicated that testing actually fosters learning. This clearly implies that students are not always aware of the beneficial effects of this learning strategy.

The use of Retrieval Practice and Test Formats

The use of retrieval practice can offer a number of advantages. For example, through the use of retrieval practice students gain insight into what they do and do not master (Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; McDaniel et al., 2011). As a result, learners know what to pay more attention to. Furthermore, retrieval ensures that information is strengthened

and better remembered in the long-term memory (Carpenter, 2009; Roediger, Agarwal, McDaniel, & McDermott, 2011; Roediger & Butler, 2011). However, according to Glover (1989) free recall, cued recall, and recognition have different effects on learning and memory. Free recall learning generally consists of studying a list of words, after which students have to recall as many of those words in any possible order (Karpicke & Roediger, 2010). Cued recall is the retrieval of memory with the help of cues. Cued recall therefore differs from free recall in that a cue is presented that is related to the information that has to be retrieved (Carpenter & DeLosh, 2006). An example is that in order to remember the word “leather” the word “cow” may be used as a cue.

Research done by Carpenter and DeLosh (2006) shows that fewer retrieval cues were associated with a better memory on the final test. They found that items that had been retrieved with a one-letter cue were retained better on the final test and that the final test score decreased as more cues were added (Exp. 2). To explain these effects of retrieval practice, Pyc and Rawson (2009) proposed the effortful retrieval hypothesis which will be discussed more in depth in the next section.

Effortful Retrieval Hypothesis

The effortful retrieval hypothesis states that more effortful and difficult retrieval during learning improves one's performance. Furthermore, it states that "not all successful retrievals are equal: more difficult retrievals are better for memory than less difficult retrievals" (Pyc & Rawson, 2009, p. 438). Pyc and Rawson (2009) assume that difficult successful retrievals require more mental effort, which results in enhanced memory. Two conditions have to be satisfied in order to test the effortful retrieval hypothesis. First, the difficulty of retrieval must vary. Second, retrieval during practice must be successful which means items have to be practiced until they are correctly retrieved (Pyc & Rawson, 2009).

This effortful retrieval hypothesis is based on the desirable difficulties framework (Bjork, Dunlosky, & Kornell, 2013). Desirable difficulties are beneficial, "because they trigger encoding and retrieval processes that support learning, comprehension, and remembering" (Bjork & Bjork, 2011 p. 58).

However, according to Bjork and Bjork (2011) during learning, while restudying the learning material, learners are often confronted with a recurrent problem. When the learners restudy the learned material, they tend to believe that they have learned the material, and

therefore might think that they learn effectively and have achieved a certain level of learning. While actually the learner is misled to what extent (s)he has mastered the learning material (Bjork & Bjork, 2011). This phenomenon can be explained by the fact that through restudying a sense of familiarity can arise which learners interpret as understanding the learning material. It is also seen as a method of learning where information comes to mind easily, performance improves rapidly, but fails to support transfer and long-term retention (Bjork & Bjork, 2011). Eventually, this method of learning only leads to short-term learning.

For this reason, in order to optimize transfer and long-term retention, learners should make use of methods of learning that create challenges on learning material that have to be mastered (Bjork & Bjork, 2011). Testing yourself on material is therefore known as one of the desirable difficulties. However, it is important to be aware of the fact that "if learners do not have the skills and background knowledge to respond successfully, desirable difficulties become undesirable" (Bjork & Bjork, 2011, p. 58). As then, those difficulties are not beneficial for learning. Based on the desirable difficulty framework, Pyc and Rawson (2009) assume that the difficulty of retrieval is equal to the amount of effort the learner experienced during learning. In other words, difficult but successful retrievals (i.e., free recall) require more effort in contrast to successful but easier retrievals (i.e., recognition). For this reason, Pyc and Rawson (2009) tested various test formats in their study, because they believe that those forms of test formats will elicit different forms of effort. This is known as the effortful retrieval hypothesis. The next section will provide more insight on earlier studies that have tested the effortful retrieval hypothesis regarding the mental effort during learning.

Test Formats and Mental Effort

In many studies, a found difference in learning between the different test formats is explained based on the retrieval effort. Those studies have shown that free recall testing leads to better learning outcomes than recognition testing (Carpenter & Delosh, 2006; Glover, 1989; Malmberg, Lehman, Annis, Criss, & Shiffrin, 2014). This could imply that indeed a more effortful and difficult retrieval (i.e., free recall) lead to higher invested mental effort and accordingly to more activation, which strengthens the memory and therefore lead to better learning outcomes (Endres & Renkl, 2015; Halamish & Bjork, 2011).

Nevertheless, only one study tested the effortful retrieval hypothesis and compared perceived mental effort on different test formats (free recall vs. cued recall vs. recognition) to a

baseline control condition (restudying; Coppens, De Jonge, Van Gog, & Kester, 2019). Coppens et al. (2019) expected that free recall would require the highest perceived mental effort and yield the best retention of successfully retrieved word pairs, followed by cued recall, followed by recognition, followed by restudy. Coppens et al. (2019) employed a between-subject design in which the participants were randomly assigned to one of the four conditions. Unfortunately, they did not find any differences between the conditions in perceived mental effort during the learning phase, nor in retention after one week (Coppens et al., 2019).

For this reason, the present study concerns a follow- up study to the study of Coppens et al. (2019). In order to measure the effect of different types of test formats and perceived mental effort on performance in the present study, a within-subject design will be conducted. By the use of a within-subject design, every participant will participate in the four conditions (i.e., restudying, free recall, cued recall, and recognition) and therefore participants can compare their own invested mental effort regarding the different test formats. Furthermore, the advantage of this design is that the variation in the groups remains the same because every participant participates in all measurements (Field, 2013).

Present study

Aim of the present study is to investigate whether the three different test formats (i.e., free recall, cued recall, and recognition) elicit different mental effort during learning and whether a higher perceived mental effort contributes to performance on the final test. For this reason, I compared the effect of the three test formats on perceived mental effort and performance on a cued recall test after one week.

The results have both scientific and practical relevance. Although in many studies a found difference in learning between test formats is explained on basis of retrieval effort, so far only one study has tried to measure whether there was actually a difference in effort regarding the different test formats. For this reason, the present study might contributes to the scientific body of knowledge with concern to the effectiveness of retrieval practice and mental effort on study performance of learners.

The results can contribute to educational practice as well, as the results may indicate how students' performance can be increased during learning. Nowadays self-regulated learning (SRL) receives a lot of attention in education. SRL is an active process, in which the learner takes control and responsibility for their own learning (Zimmerman, 2002). Students could integrate

retrieval practice while learning. In this way, they are able to master the subject matter independently while using retrieval practice as a learning strategy during self-studying (Roediger & Butler, 2011).

However, a study by McCabe (2011) shows that students are unaware of the effectiveness of retrieval practice as a learning strategy. Therefore, with the new insights offered by the outcomes of the present study, teachers may improve their learning programs by applying this form of testing in their lessons (e.g., flashcards, quizzes, and questions). In that way, students become familiar with retrieval practice and are able to apply this strategy during self-study. If students gain more insight on the benefits of this form of testing, they can learn in a more effective way, and because of the beneficial effects learners may receive higher grades (McDaniel et al., 2011).

Therefore, the following research question will be addressed: What is the relationship between different forms of retrieval practice and perceived mental effort, and how does this relate to the performance of learners? Based on the literature discussed above I expect that:
H1: Free recall yields the highest perceived mental effort followed by cued recall, followed by recognition, followed by restudying.
H2: Free recall yields the highest performance on the final test followed by cued recall, followed by recognition, followed by restudying.

Method

Participants and Design

The power analysis with G*Power has shown that with a medium effect size of f =. 25 and with a power of .8 a sample size of 24 is needed. Originally, 44 Dutch participants were asked in February to participate in this experiment. All the participants were made aware of their participation in the experiment and were asked to give informed consent. Due to the within-subject design, all the participants experienced four conditions: cued recall, free recall, recognition, and restudying. The order of the conditions and the allocation of the word pairs to the conditions were randomized. The reason for this was to minimize the effects of the order of combinations and the word pair-condition combinations. Furthermore, participants were tested individually on personal computers. Two participants were excluded because of errors within the experiment. For this reason, the final sample consisted of 42 participants, (31 female, 11 male) with ages ranging from 18 to 80 years old (M = 34.45, SD = .45).

Instrumentation

Performance. For the word pairs used as learning material in the experiment, the Carpenter's (2009) 'weak cue-target' list of 48-word pairs was used (See Appendix A). Half of the word pairs were used as learning material and randomly divided over the four conditions. During the learning phase, each participant learned 24 word pairs. During the recognition test, six word pairs that were not learned before were added and used as distractors. For each correctly recognized word pair, participants received 1 point and for every incorrectly recognized word pair 1 point was subtracted on the recognition test. To measure performance on the cued recall test in the learning phase, participants received the cue, and where then asked to type in the target on the computer keyboard. For each correct answer, participants received 1 point and for each incorrect or incomplete answer, they received 0 points. To measure performance on the free recall test in the learning phase, participants had to type in the word pairs on the computer keyboard. For each correct answer, participants received 1 point, for each incomplete answer participants received .05 points when one of the words was correct, and for incorrect answers, they received 0 points.

To measure performance on the final cued recall test, participants received the cue, after which the participants typed the target on the computer keyboard. For each correct answer, the

participants received 1 point, and for each incorrect answer or incomplete answer, they received 0 points.

Retrieval effort. During the learning phase, all participants indicated after each word pair how much effort it cost to recall or restudy them. To measure the amount of effort participants experienced during learning of the word pairs, the unidimensional 9-point symmetrical rating scale, developed by Paas (1992) was used. Ranging from 1; very, very low mental effort, to 9; very, very high mental effort (See Appendix B). This scale has been widely used in many educational and psychology studies because of its reliability, sensitivity, and ease of use. (Paas, Tuovinen, Tabbers, & Van Gerven 2003; Paas, Van Merriënboer, & Adam, 1994; Van Gog, Kirschner, Kester, & Paas, 2012; Van Gog & Paas, 2008).

Procedure

All participants were approached face to face or via mail (See Appendix C). The participants were asked to participate in the experiment. When the participants agreed, they filled in an informed consent form (See Appendix D). Because the experiment took place at different times and places, the researcher provided all information during the experiment to make sure that all participants received the same amount of information. The experiment took place digitally and was designed and presented in Gorilla (https://gorilla.sc). Furthermore, all the participants were tested individually in a quiet room.

The experiment consisted of two sessions as shown in Figure 1. In the study phase, participants were given an opportunity to study the word pairs one by one. The word list was divided into chunks of six word pairs, so these chunks are the different conditions (i.e., restudying, free recall, cued recall, and recognition). Each chunk was studied once, and immediately thereafter practiced 3 times in one of the four conditions. After that, the next chunk with the respective condition was presented in the same way. Every word pair was individually presented on the computer screen. For the restudying test, participants only restudied the word pairs one by one. For the recognition test, participants indicated whether they had learned the word pairs before on the computer screen. In the cued recall test, participants saw one cue word at a time and had to complete as many targets as they could. For the free recall test, the participants had to recall as many word pairs as they could by typing them on the computer keyboard. During this test, all participants indicated after each word pair how much effort it cost to recall or restudy them. After one week the participants were asked to log in again on the

computer for the final cued recall retention test. Participants had to complete the given cue with the target via the computer keyboard.

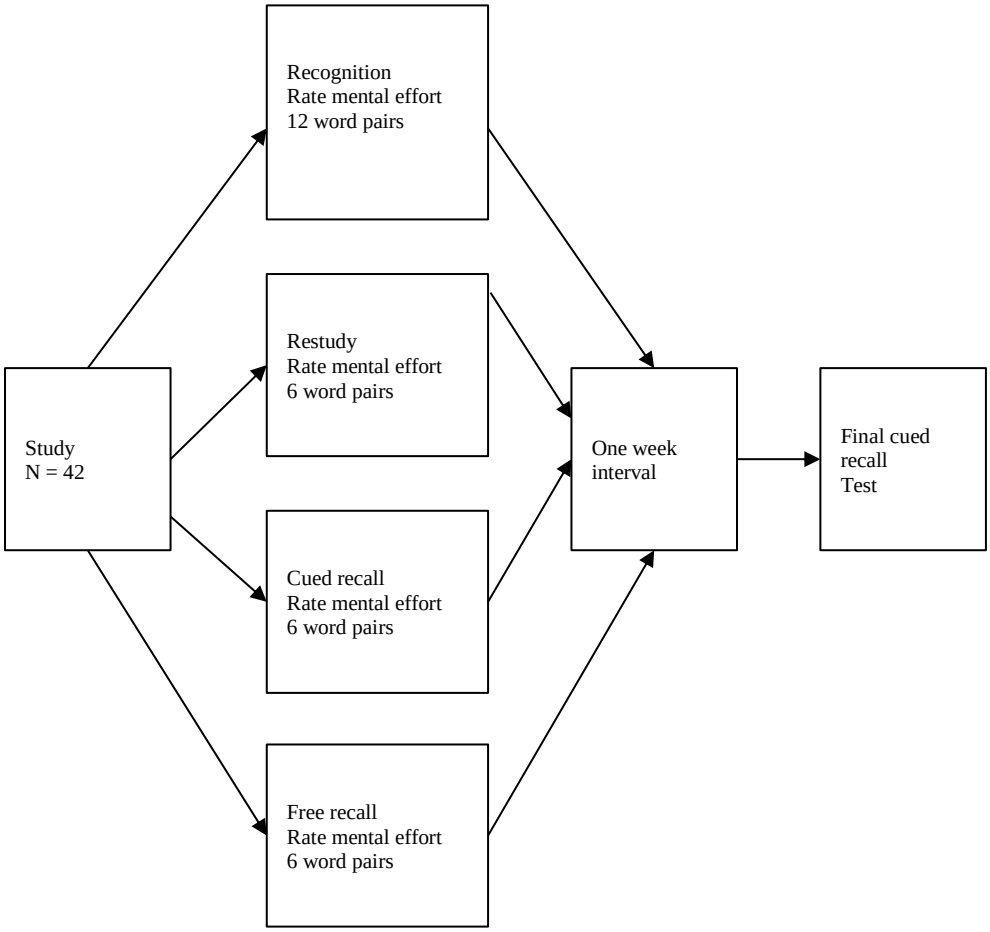


Figure 1. Schematic representation of the experiment.

Data Analysis

To be able to perform the analysis, a few variables had to be computed. First, the effort score for every condition (i.e., free recall, cued recall, recognition and restudy) originally consisted of 18 scores. For each participant, all effort scores per condition were added together and then divided by 18. Second, for the variable performance in the learning phase, all scores per condition were added together and then divided by 18. The same applies to the variable performance in the final test, although this time, the total scores per condition were added together and then divided by 24. The raw data showed that 2 participants had duplicated the final test. Therefore, it was decided to delete the second final test. Furthermore, 2 cases were removed from the data set because during the experiment an error had occurred. In order to answer the

research question, several analyses were conducted. All the analyses were conducted using IBM SPSS Statistics (version 25.0).

Second, ICCs were calculated in SPSS (Field, 2013) to determine the inter-rater reliability of the results of the learning phase and the final test. ICCs type "two way mixed" and "absolute agreement" were used (McGraw & Wong, 1996). Cichetti (1994) proposed standards for ICC results. The reliability of the ICC calculation is < .40 = low, .40 - .59 = reasonable, .60 - .74 = good, and .75 - 1 = excellent.

Third, to investigate the effectiveness of retrieval practice and mental effort on study performance the experiment was conducted with a within-subject design. In order to answer the research question, the initial plan of the researcher was to conduct a one-way repeated measures ANOVA. However, because of the violation of the assumption of normality, the non-parametric Friedman test had to be performed. Significant results were followed up by the Wilcoxon signed-rank test.

The Wilcoxon signed-rank test is a non-parametric statistical test, used to compare two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ (Field, 2013). Several assumptions should be met in order to use the Wilcoxon signed-rank test. First, the assumption of independence was checked. Each participant should participate only once in the research and should not influence the participation of others. During the experiment, every participant only participated once. For this reason, the first assumption was met. Second, the dependent variable should at least be ordinal. As in this research all the dependent variables are continuous this assumption is met as well. Third, the distribution of difference scores between the two levels of the independent variables should be roughly symmetrical. Histograms of the difference scores confirmed that the symmetry of the distribution of difference scores were indeed roughly symmetrical. Therefore, the final assumption was met as well.

Last, to interpret the effect sizes, the rules of thumb from Cohen (1988) are used. According to Cohen (1988), r = .1 is considered as small, r = .3 is considered as medium, and r = .5 is considered as a large effect size.

Results

Descriptive Statistics

First, the descriptive statistics (*Median, first and third quartiles*) of the examined variables are described and included in Table 1.

Table 1

Median and first and third quartiles of the Examined Variables

	Mental effort During learning	Performance immediately after learning	Performance in the final test
Restudy	1.17 (1.00; 2.10)		.33 (.00; .50)
Recognition	1.18 (1.00; 1.37)	1.00 (.94; 1.00)	.33 (.17; .54)
Cued recall	2.22 (1.30; 4.25)	.94 (.65; 1.00)	.58 (.29; .83)
Free recall	6.33 (3.54; 7.28)	.40 (.22; .72)	.50 (.17; .88)

Inter-rater reliability

To assess the inter-rater reliability, the given points for the answers of both the learning phase and final test were classified by both the researcher and a second, independent rater. The inter-rater reliability can be indicated as excellent for both performances on cued recall (ICC = .1.00), free recall (ICC = .98), and performance on the final test (ICC = 1.00) as well.

Statistical Analyses

Performance in the Learning Phase

It was expected that the performance in the learning phase with regard to the three different conditions (i.e., recognition, cued recall, and free recall) would be significantly different. It was assumed that during the learning phase the performance of recognition would be the greatest, followed by cued recall, followed by free recall.

A Friedman test indicated that rankings of performance in the learning phase varied significantly across the three conditions of learning (i.e., recognition, cued recall, and free recall) $X^2_f = 56.60$ (corrected for ties), $df = 2$, N – Ties = 42, $p < .001$.

When comparing difference in performance in the learning phase with regard to recognition (*Mdn* = 1.00) with performance in the learning phase with regard to cued recall (*Mdn* = .94), follow-up pairwise comparisons with the Wilcoxon signed-rank test and a Bonferroni correction adjusted α of .0250 indicated that there was a significant effect (*Z* = -3,59, *p* = < .001). This effect can be considered as medium (*r* = -.39).

When comparing difference in performance in the learning phase with regard to cued recall (*Mdn* = .94) with performance in the learning phase with regard to free recall (*Mdn* = .44), a significant effect was found (*Z* = -4.50, *p* < .001). This effect can be considered as medium (*r* = -.49). This means that during learning, testing yourself via recognition lead to better performance, followed by cued recall, followed by free recall.

Perceived Mental Effort During the Learning Phase

The first hypothesis assumed that there was a significant difference between the four conditions in learning (i.e., restudy, recognition, cued recall, and free recall) and perceived mental effort during the learning phase. A Friedman test indicated that rankings of mental effort varied significantly across the four conditions of learning (i.e., restudy, recognition, cued recall, and free recall) $X^2_f = 57.71$ (corrected for ties), *df* = 3, *N* – Ties = 42, *p* < .001.

First, when comparing differences in mental effort during free recall (*Mdn* = 6.33) with mental effort during cued recall (*Mdn* = 2.22), follow-up pairwise comparisons with the Wilcoxon signed-rank test and a Bonferroni correction adjusted α of .0167 indicated that there was a significant effect (*Z*= -4.47, *p* <. 001). Subsequent analysis indicated a large effect (*r* = -.70).

Second, when comparing differences in mental effort during cued recall (*Mdn* = 2.22) with mental effort during recognition (*Mdn* = 1.18), a significant effect was found (*Z* = -4.63, *p* <. 001). Subsequent analysis indicated a large effect (*r* = -.50).

Third, when comparing differences in mental effort during recognition (*Mdn* = 1.18) with mental effort during restudy (*Mdn* = 1.17), a non-significant effect was found (*Z* = -1.86, *p* = .063).

The found results were not entirely in line with the formulated hypothesis. Results showed that indeed free recall yields the highest perceived mental effort followed by cued recall, followed by recognition. However, no significant effect was found, when comparing mental

effort during recognition with mental effort during restudy. This means that recognition did not lead to more perceived mental effort during learning than restudy.

Performance on the Final Test

The second hypothesis stated that free recall yields the highest performance on the final test followed by cued recall, followed by recognition, followed by restudying. A Friedman test indicated that rankings of performance on the final test, varied significantly across the four conditions of learning (i.e., restudy, recognition, cued recall, and free recall) $X^2_f = 19.49$ (corrected for ties), $df = 3$, N – Ties = 42, $p < .001$.

When comparing differences in performance on the final test during free recall (*Mdn* = .50) with performance during cued recall (*Mdn* = .58), follow-up pairwise comparisons with the Wilcoxon signed-rank test and a Bonferroni correction adjusted α of .0167 indicated a non-significant effect ($Z = -.79$, $p = .65$).

Furthermore, when comparing differences in performance on the final test during cued recall (*Mdn* = .58) with performance during recognition (*Mdn* = .33), a significant effect was found ($Z = -2.91$, $p < .001$). Subsequent analysis indicated a medium effect ($r = -.32$).

Last, when comparing differences in performance on the final test during recognition (*Mdn* = .33) with performance during restudy (*Mdn* = .33), a non-significant effect was found ($Z = -1,52$, $p = .13$).

The found results were not entirely in line with the formulated hypothesis. The results showed that there was no significant effect when comparing performance during free recall with performance during cued recall. This means that free recall does not lead to better performance on the final test. Furthermore, there was no significant effect, when comparing performance during recognition with performance during restudy. This means that recognition does not lead to better performance on the final test. However, it appeared that there was a significant effect when comparing performance during cued recall with performance during recognition. This means that the use of cued recall during learning leads to better performance on the final test. This might suggest that the use of more effortful learning strategies indeed leads to better performance on the final test.

Discussion

The goal of the present study was to examine the relationship between different forms of retrieval practice and perceived mental effort, and how it relates to the performance of learners. The first expectation was in accordance with the literature of several studies, namely that free recall yields the highest mental effort, followed by cued recall, followed by recognition, followed by restudy (Bjork et al., 2013; Pyc & Rawson, 2009).

Results of the present study confirm that free recall yields the highest mental effort, followed by cued recall, followed by recognition. This research has therefore shown that the learner experienced a higher mental effort during the use of more effortful and difficult retrieval such as free and cued recall, in comparison to less useful and effortful strategies such as recognition. This is consistent with previous studies that have assumed that difficult learning strategies such as free and cued recall require more mental effort (Pyc & Rawson, 2009; Roediger & Butler, 2011).

The results show no statistically significant difference however, between recognition and restudy with regard to mental effort during learning. This can be explained by the fact that measuring retrieval effort during restudy is not possible, as no retrieval takes place. Logically, during restudy, it is only possible to measure the mental effort with regard to rereading instead of recalling the learned material.

Moreover, it can be disputed if recognition was actively different enough from restudying the learned material. In both conditions, all the participants did not have to type in the word pairs. Instead, for both conditions the word pairs were displayed on the computer screen. The only minimal difference is that distractors were added in the recognition condition. These distractors might, perhaps, have been too easy. For this reason, it could be that the participants did not experience a great difference in perceived mental effort with regard to recognition (*Mdn* = 1.18) and restudy (*Mdn* = 1.17). This might explain the non-significant difference between recognition and restudying in perceived mental effort. Nonetheless, the results still clearly indicate that indeed the difficulty of retrieval is equal to the amount of effort the learner experienced during learning (Pyc & Rawson, 2009).

Furthermore, in line with previous studies (Carpenter, 2009; Roediger & Butler, 2011; Roediger et al., 2011), the expectation was that free recall should yield the highest performance on the final test, followed by cued recall, followed by recognition followed by restudy. Results of

this study however show that both free recall and cued recall yield the highest performance on the final test followed by both recognition and restudying. There were no statistically significant differences found in performance on the final test between free recall and cued recall, or between recognition and restudying. This could be explained by the fact that the final test consisted of a cued recall test and therefore the principle of transfer-appropriate processing could have occurred (Meier & Graf, 2000). According to this principle, performance on any given task will be the highest if the characteristics of the learning procedure are similar to the characteristics of the assessment procedure (Meier & Graf, 2000 in Morris, Bransford, & Frank, 1977). In the present study, the four different conditions of testing were used during the learning phase. However, in the final test, performance of the learner was tested via a cued recall test. This thus possibly explains the non-significant results in the present study. Suggestions for further research are therefore to make a distinction between the different conditions in the final test as well, so that the conditions are tested in the way they were taught in the first place. It is worth investigating whether testing effects are maximized if the testing in the learning phase and final test are identical.

Furthermore, a possible explanation for the non-significant result between recognition and restudy could be that recognition was indeed not actively different enough from restudying the learned material during the learning phase as described earlier. It can be suggested that because both of those conditions did not require a lot of mental effort, the retention after one week was limited on the final test (Roediger & Karpicke, 2006).

Another possible explanation for the results can be found in the research of Carpenter and Butler (2011). In their study, feedback was provided after every trial in the learning phase. Giving feedback including the correct answers increases learning because the learner is able to correct incorrect answers while learning and can also maintain the correct answers. Furthermore, research from Kang, McDermott & Roediger (2007) shows that if learners do not receive feedback, the benefits of learning are limited or even absent. Not giving feedback with the right answers during learning can thus be considered a pitfall for the learner. Incorrect learning of the word pairs can lead to learners thinking that what they have learned is correct and apply this in the final test. This phenomenon has been demonstrated in several studies (Kang et al., 2007; Roediger & Butler, 2011). It might be that this phenomenon has occurred in the present study because no feedback was provided after every trial in the learning phase. The help of feedback

could therefore have ensured more successful retrieval in the final test (Butler & Roediger, 2008; Carpenter, Pashler, & Vul, 2006; Pashler, Capeda, Wixted, & Rohrer, 2005).

Lastly, the lack of a sustained benefit of free recall over cued recall and recognition over restudy in the final test could be explained by the fact that the present study used the same number of exposure trials for each condition, as opposed to for example the study of Karpicke and Roediger (2008), where participants were allowed to learn until they achieved correct recall. Pyc and Rawson (2009) indicate that in order for retrieval to be successful, word pairs should be practiced until they are correctly retrieved during practice. As this is not the case in the current study, this might also have led to reduced performance in the final test.

In sum, although somewhat different from the expectation, results are still in line previous studies and support the notion that retrieval through recall benefits subsequent test performance more than retrieval processes with recognition (Glover, 1989).

Limitations and Further Research

The results of this study must be interpreted in light of several limitations that could have affected the results in multiple ways. First, the experiment was conducted among 42 participants. Because of limited available time, it was not possible to set up an experiment with all the 42 participants present at the same day and time. For this reason, days and times differ. Although an attempt was made to have the same number of participants to participate in the experiment each day, unfortunately, this was not always feasible.

In addition, this limitation applies for the final test as well. Because of the limited available time, it was decided to let the participants take the final test exact one week later without the supervision of the experimenter. Some participants took the final test in the evening, while the learning phase was in the morning. As a result, there might be various factors that could have influenced the final performance of the participant.

For example, state-dependent learning could have occurred (Goldstein, 2011). According to the idea of state-dependent learning, one is better able to retrieve their memories if their "mood or state of awareness" is the same at the time they encode the learned material and at the time they retrieve it (Goldstein, 2011). It cannot be excluded that participants who made the first part of the experiment (i.e., learning phase) in the morning and the second part (i.e., final test) in the evening for example, experienced a difference in mood. It is possible that participants who

took the final test in the evening were much more tired after a busy working day, and for this reason, disappointing performance had occurred.

Second, because the participants made their final test without the supervision of the experimenter, there is a possibility that the so-called satisfaction has occurred during the final test (Krosnick, 1991). Perhaps, participants have not paid full attention while doing the final cued recall test. To increase the reliability, suggestions for further research would therefore be that both parts of the experiment are carried out at once, with all the participants present and under supervision of the experimenter.

Third, the external validity of the present study is low, since the results cannot be generalized. It is generally known that generalization of the results in an experiment is experienced as difficult (Aronson, Wilson, & Bever, 1998 in Aronson, Wilson, & Akert, 2017). Naturally, this limitation also applies to the present study as the experiment is conducted in an artificial situation which makes it difficult to generalize it to another situation. Future research could make use of psychological realism (Aronson et al., 1998). Future research could conduct a similar experiment during, for example, an academic university course. In that way, the experiment is then organized in such a way that participants will react the same way as they would in real-life educational settings (Aronson et al., 1998). Generalization and, in turn, external validity will then increase.

Fourth, a convenience sample was used in the present study. As a result, it is not possible to measure controlled bias and variability. Furthermore, the obtained data cannot be fully generalized beyond the sample (Acharya, Prakash, Saxena, & Nigam, 2013). Nevertheless, the present study has attempted to use a representative sample. Consideration was given to the level of education of the participants as well as the variation in age. Future research should make use of probability sampling methods, to ensure representativeness of the sample and also for the generalizability of the results to the target population (Acharya et al., 2013).

Although it is beyond the scope of the present study, it is worth to investigate how much time the learner needs while using a less effortful strategy (i.e., recognition) to achieve the same results as the more effortful strategies (i.e., cued recall, free recall). Until now, the factor time has not been studied often in relation to different learning strategies. Further studies could focus on the question whether less effective learning strategies can reach the same performance levels

if more time was invested. In this way, learners may become even more aware of the effectiveness of the use of cued and free recall.

Educational Implications

The insights offered by the outcomes of this research contribute to the educational field and underline the important role of mental effort when trying to exploit the testing effect during learning. The results of the present study show that retrieval practice indeed provides better retention than restudying alone. This suggests that testing should not only be used to measure someone's knowledge but should be also used as a learning tool (Dempster, 1992; Roediger & Butler, 2011). Teachers should help learners to develop such a learning tool with regard to retrieval practice. The development of a more difficult test such as free and cued recall might actually contribute to the performance in the final test of the learner.

Furthermore, the results of the present study suggest that retrieval practice is beneficial if it requires a lot of effortful processing (Roediger & Butler, 2011). Teachers should therefore encourage and motivate students to use this strategy during learning. Teachers must encourage students to get over the struggle during learning when using a more effortful strategy such as retrieval practice. When using a form of retrieval practice which requires more effortful processing, initially the learner will not know everything and might think that he or she is not able to learn the material. For this reason, it seems plausible that learners apply less effortful strategies such as recognition or restudying as a learning tool. Results of the present study do indicate that this form of learning will lead to better performance during learning. This might therefore have a positive effect on the learners which makes them continuously apply this strategy. However, the present study also shows that effective but effortful strategies such as free and cued recall will eventually lead to even better retention and performance.

It has been widely demonstrated by studies that taking an initial test can improve later memory and promotes long-term retention by the learner (Bouwmeester & Verkoeijen, 2011; Carpenter et al., 2009; Goossens et al., 2013; Toppino & Cohen, 2014). Although the results of the present study and earlier research show that retrieval practice is very effective, it seems to be even more effective when feedback is provided during learning (Butler et al., 2008; Butler & Roediger, 2008; Carpenter et al., 2006; Pashler et al., 2005; Roediger & Butler, 2011). Providing feedback during retrieval practice ensures that learned material is correctly entered in the memory of the learner, which will increase the performance in the final test.

Conclusion

In sum, the present study indicates that the performance will be the highest during learning with the use of recognition, followed by cued, followed by free recall. Furthermore, the present study confirms the findings of many previous studies which showed that retrieval practice is a superior strategy of learning when compared to restudy (Dunlosky et al., 2013). The present study also indicates that the mental effort with regard to the different learning conditions differs and that free recall yields the highest mental effort, followed by cued recall, followed by recognition during learning.

To conclude, while it seems tempting to adopt a learning strategy that requires less mental effort during learning, it is more beneficial to choose learning strategies that require more mental effort as this will result in better retention and, in turn, higher grades.

References

Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: Why and how of it. *Indian Journal of Medical Specialties, 4*, 330-333. doi:10.7713/ijms.2013.0032

Aronson, E Wilson, T.D. Akert, M.R., & Sommers, S.R. (2017). Sociale psychologie. Benelux: Pearson

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when ofstudy strategies: Comparing intended versus actual study behaviour. *Memory*, *25*, 784–792. doi:10.1080/09658211.2016.1221974

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology, 64*, 417-444. doi:10.1146/annurev-psych-113011-143823

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M.Hough, J. R. Pomerantz (Eds.), & FABBS Foundation, *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York, NY, US: Worth Publishers.

Bouwmeester, S., & Verkoeijen, P. P. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language, 65*, 32-41. doi:10.1016/j.jml.2011.02.005

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 918. doi: 10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *363*, 604-616. doi: 10.3758/MC.36.3.604

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & cognition, 34, 268-276.* doi*:10.3758/BF03193405*

Carpenter, S. K., Pashler, H., & Vul, E. (2006). *What types of learning are enhanced by a cued recall test?. Psychonomic bulletin & review, 13, 826-830. doi:10.3758/BF03194004*

Carpenter, S. K. (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*. doi:10.1037/a0017021

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23,* 760-771. doi:10.1002/acp.1507

Cauldron science. (2016). Gorilla [Software]. Retrieved from https://gorilla.sc/

Cichetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum

Coppens, L. C., De Jonge, M. O., Van Gog, T., & Kester, L. (2019). The effect of test modality on perceived mental effort. Manuscript in preparation.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental review, 12,* 45-75.

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. doi:10.1177/1529100612453266

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Frontiers in psychology, 6,* 1054. doi: 10.3389/fpsyg.2015.01054

Field, A.P. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll* (4th e.d.). London, UK: Sage.

Glover, J.A. (1989) The ''testing'' phenomenon: not gone but nearly forgotten. J. Educ. Psychol. *81,* 392–399. doi:10.1037/0022-0663.81.3.392

Goldstein, B. E. (2011). State-Dependent Learning. In B. E. Goldstein, Cognitive Psychology (p. 185). Belmont: Wadsworth.

Goossens, N. A., Camp, G., Verkoeijen, P. P., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*, 135-142. doi:10.1002/acp.2956

Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*, 177-182. doi:10.1016/j.jarmac.2014.05.003

IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY:IBM Corp.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. J. Exp. Psychol. Learn. Mem. Cogn. *37*, 801–812. doi:10.1037/a0023219

Kang, S. H., McDermott, K. B., & Roediger III, 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. doi: 10.1080/09541440601056620

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?*Memory*, *17*,471-479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials?. *Memory & cognition*, *38*, 116-124. doi:10.3758/MC.38.1.116

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219-224. doi:10.3758/BF03194055

Krishnan, S., Watkins, K. E., & Bishop, D. V. (2017). The effect of recall, reproduction, and restudy on word learning: a pre-registered study. *BMC psychology*, *5*1, 28.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*, 213-236. doi:10.1002/acp.2350050305

Malmberg, K. J., Lehman, M., Annis, J., Criss, A. H., & Shiffrin, R. M. (2014). Consequences of testing memory. In *Psychology of learning and motivation* (Vol. 61, pp. 285-313). Academic Press. doi:10.1016/B978-0-12-800283-4.00008-3

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition, 39*, 462-476. doi:10.3758/s13421-010-0035-2

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, 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. doi: 10.1037/a0021782

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1, 30.

Meier, B., & Graf, P. (2000). Transfer appropriate processing for prospective memory tests. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 14*, S11-S27. doi: 10.1002/acp.768

Morris CD, Bransford JD, Franks JJ. 1977. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* 16: 519±533.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84*, 429-434. doi:10.1037/0022-0663.84.4.429

Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and motor skills, 79*, 419-430.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist, 38*, 63-71. doi:10.1207/S15326985EP3801_8

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. doi:10.1037/0278-7393.31.1.3

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437–447. doi:10.1016/j.jml.2009.01.0

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, doi:10.1111/j.1745-6916.2006.00012.x

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382. doi:10.1037/a0026252

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences, 15*, 20-27. doi:10.1016/j.tics.2010.09.003

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*6, 1432. doi:10.1037/a0037559.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*, 252-257. doi:10.1027/1618-3169.56.4.252

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*, 16-26. doi:10.1080/00461520701756248

Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*, 833-839. doi:10.1002/acp.2883

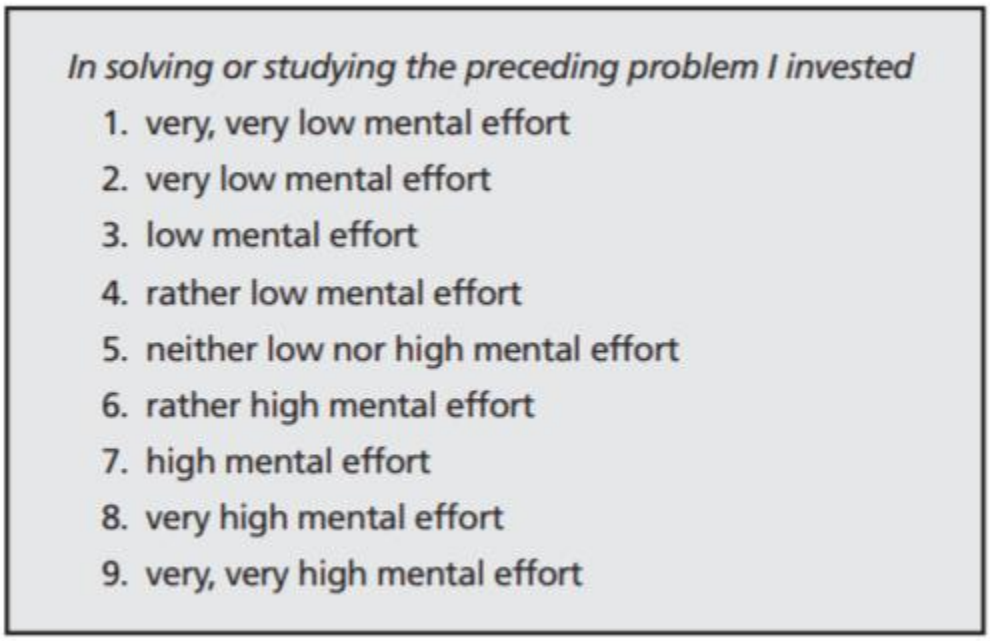
Zimmerman, B.J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, *41*, 64-70. doi: 10.1207/s15430421tip4102_2

Appendix

Appendix A

Strong cue	Weak cue	Target
List 1		
Toast (.364)	Basket (.014)	Bread
Jury (.250)	Rights (.015)	Court
Valentine (.423)	Rib (.014)	Heart
Rodeo (.477)	Camel (.014)	Horse
Stereo (.333)	Theater (.014)	Music
Neptune (.399)	Comet (.014)	Planet
Chimney (.24)	Fire (.018)	Smoke
Dentist (.459)	Lips (.014)	Teeth
List 2		
Chunk (.054)	Chisel (.01)	Block
Adolescent (.262)	Mitten (.011)	Child
Tea (.369)	Steam (.014)	Coffee
Patient (.365)	Virus (.013)	Doctor
Mow (.275)	Picnic (.014)	Grass
Fork (.37)	Kitchen (.015)	Knife
Cost (.418)	Contest (.015)	Money
Station (.083)	Airplane (.01)	Train
List 3		
Vein (.384)	Bruise (.013)	Blood
Sphere (.258)	Hole (.016)	Circle
Skirt (.295)	Maid (.011)	Dress
Blossom (.441)	Leaf (.012)	Flower
Suite (.356)	Lounge (.014)	Hotel
Adventure (.295)	Desert (.015)	Island
Switch (.459)	Morning (.014)	Light
Education (.315)	Pupil (.016)	School
List 4		
Bristle (.397)	Flick (.013)	Brush
Beverage (.493)	Ice (.016)	Drink
Mop (.244)	Elevator (.014)	Floor
Home (.333)	Barn (.016)	House
Video (.258)	Television (.014)	Movie
Folder (.322)	Scissors (.012)	Paper
Stream (.321)	Bay (.013)	River
Main (.315)	Directions (.013)	Street
List 5		
Shell (.25)	Raft (.011)	Beach
Couch (.288)	Hammock (.013)	Chair
Alarm (.388)	Dial (.019)	Clock
Prom (.221)	Spin (.013)	Dance
Wilderness (.264)	Meadow (.014)	Forest
Shatter (.412)	Bead (.016)	Glass
Saliva (.262)	Speak (.017)	Mouth
Rock (.269)	Building (.017)	Stone
List 6		
Minister (.349)	Soul (.014)	Church
Lunch (.269)	Manners (.014)	Dinner
Picket (.384)	Barrier (.014)	Fence
Plum (.299)	Seed (.011)	Fruit
Note (.299)	Print (.014)	Letter
Call (.378)	Ear (.014)	Phone
Cuff (.247)	Jacket (.013)	Shirt
Hose (.473)	Mist (.013)	Water

Appendix B



The Paas (1992) subjective rating scale.

Appendix C

Beste,

Mijn naam is Michelle, student van de Universiteit Utrecht. Ik doe onderzoek naar leren middels diverse test formats en de effectiviteit daarvan. Met dit onderzoek hoop ik een bijdrage te kunnen leveren en meer inzicht te kunnen verkrijgen in de wijze waarop lerenden studeren, en door verschillende test-formats te vergelijken, er achter te kunnen komen op welke manier men het beste onthoudt. Ik ben op zoek naar participanten die mee willen doen aan mijn onderzoek.

Wat houdt het onderzoek in?

Het experiment bestaat uit 2 fasen. In de eerste fase zullen de participanten verschillende woordparen leren a.d.v. verschillende test formats. In de tweede fase zal d.m.v. een “test” worden onderzocht, welke manier het meest bijdraagt aan het onthouden van de geleerde woord paren. Beide fasen zullen ongeveer een kwartier van je tijd in beslag nemen.

Doelgroep

Ik ben op zoek naar participanten 18+ die het leuk vinden om mee te werken aan mijn experiment.

Opbrengst

Indien gewenst kan ik na afloop per participant een rapportage van de algemene uitkomsten toesturen. Daarnaast kan ik ook een persoonlijke rapportage sturen. Hierin is zichtbaar hoe de scores van jou zich verhouden tot de scores van de andere deelnemende participanten. Let op in beide gevallen zijn geen gegevens van andere individuele deelnemende participanten te herleiden, jouw gegevens blijven dus altijd privé.

Privacy en vertrouwelijkheid

Alle gegevens worden vertrouwelijk behandeld en anoniem verwerkt. De gegevens worden alleen voor onderzoeksdoeleinden gebruikt en niet verstrekt aan derden.

Heb je interesse in deelname aan dit onderzoek, neemt u dan contact op met Michelle Esmee Janssen
m.e.janssen2@students.uu.nl

Met vriendelijke groet,
Michelle

Appendix D

Betreft: Leren a.d.v. verschillende test-formats experiment

In te vullen door de deelnemer

Toestemmingsverklaring voor gebruik gegevens ten behoeve van het onderzoek

Hierbij geef ik toestemming aan de voor het onderzoek verantwoordelijke onderzoeker van de Universiteit Utrecht om de gegevens die zijn verkregen tijdens het experiment te gebruiken voor onderzoek.

Mijn gegevens worden door de onderzoekers vertrouwelijk verwerkt.

Ik verklaar hierbij volledig te zijn ingelicht over de procedure van het onderzoek. Ik ben in de gelegenheid gesteld om vragen over het onderzoek te stellen en mijn (eventuele) vragen zijn naar tevredenheid beantwoord.

Ik heb genoeg tijd gehad om te beslissen of ik mee zou doen.

Ik weet dat meedoen helemaal vrijwillig is en weet dat ik op ieder moment kan beslissen om toch niet mee te doen. Daarvoor hoef ik geen reden op te geven.

Naam:……

Plaats: ....., Datum:……

Handtekening deelnemer:

In te vullen door de onderzoeksleider

Ik heb een mondelinge en schriftelijke toelichting gegeven op het onderzoek. Ik zal resterende vragen over het onderzoek naar vermogen beantwoorden. De deelnemer zal van een eventuele voortijdige beëindiging van deelname aan dit onderzoek geen nadelige gevolgen ondervinden.

Naam:……

Plaats: ....., Datum:……

Handtekening onderzoeksleider:

Appendix E

APPLICATION FORM FOR THE ASSESSMENT OF A RESEARCH PROTOCOL BY THE FACULTY ETHICS REVIEW BOARD (FERB) OF THE FACULTY OF SOCIAL AND BEHAVIOURAL SCIENCES

General guidelines for the use of this form

This form can be used for a single research project or a series of related studies (hereinafter referred to as: "research programme"). Researchers are encouraged to apply for the assessment of a research programme if their proposal covers multiple studies with related content, identical procedures (methods and instruments) and contains informed consent forms and participant information, with a similar population. For studies by students, the FERB recommends submitting, in advance, a research programme under which protocol multiple student projects can be conducted so that their execution will not be delayed by the review procedure. The application of such a research programme must include a proper description by the researcher(s) of the programme as a whole in terms of the maximum burden on the participants (e.g. maximum duration, strain/efforts, types of stimuli, strength and frequency, etc.). If it is impossible to describe all the studies within the research programme, it should, in any case, include a description of the most invasive study known so far.

Solely the first responsible senior researcher(s) (from post-doctoral level onwards) may submit a protocol.

Any approval by the FERB is valid for 5 years or until the information to be provided in the application form below is modified to such an extent that the study becomes more invasive. For a research programme, the term of validity is 2 years and any extension is subject to approval. The researcher(s) and staff below commit themselves to treating the participants in accordance with the principles of the Declaration of Helsinki and the Dutch Code of Conduct for Scientific Practices as determined by the VSNU Association of Universities in the Netherlands (which can both be downloaded from the FERB site on the Intranet[1]) and guarantee that the participants (whether decisionally competent or incompetent and/or in a dependent relationship vis-a-vis the

[1] See: https://intranet.uu.nl/facultaire-ethische-toetsingscommissie-fetc

researcher or not) may at all times terminate their participation without any further consequences.

The researcher(s) commit themselves to maximising the quality of the study, the statistical analysis and the reports, and to respect the specific regulations and legislation pertaining to the specific methods.

The procedure will run more smoothly if the FERB receives all the relevant documents, such as questionnaires and other measurement instruments as well as literature and other sources on studies using similar methods which were found to be ethically acceptable and that testify to the fact that this procedure has no harmful consequences. Examples of studies where the latter will always be an issue are studies into bullying behaviour, sexuality, and parent-child relationships.

The FERB asks the researcher(s) to be as specific as possible when they answer the relevant questions while limiting their answers to 500 words maximum per question. It is helpful to the FERB if the answers are brief and to the point.

Our FAQ document that can be accessed through the Intranet provides background information with regards to any questions.

The researcher(s) declare to have described the study truthfully and with a particular focus on its ethical aspects.

Signed for approval[2]:

Date:

[2] The senior researcher (holding at least a doctoral degree) should sign here.

A. GENERAL INFORMATION/PERSONAL DETAILS

1.

a. Name(s), position(s) and department(s) of the responsible researcher(s):

Michelle Esmee Janssen, Master student Educational Sciences faculty Behavioral sciences

Name(s), position(s) and department(s) of the executive researcher(s):

2. Title of the study or research programme - Does it concern a single study or a research programme? Does it concern a study for the final thesis in a bachelor's or master's degree course?:

Master's Programme in Educational Sciences, Final thesis master's degree

3. Type of study (with a brief rationale):

- experimental

4. Grant provider:

no

5. Intended start and end date for the study:

01-11-2018 / 10-06-19

6. Research area/discipline:

Cognition and ICT

7. For some (larger) projects it is advisable to appoint an independent contact or expert whom participants can contact in case of questions and/or complaints. Has an independent expert been appointed for this study?[3]:

[3] This contact may, in principle, also be a researcher (within the same department, or not) who is able to respond to the question or complaint in detail. Independent is to say: not involved in the study themselves. The FERB upholds that an independent contact is not obligatory, but will be necessary when the study is more invasive.

no

8. Does the study concern a multi-centre project, e.g. in collaboration with other universities, a GGZ mental health care institution, a university medical centre? Where exactly will the study be conducted? By which institute(s) are the executive researcher(s) employed?:
no

9. Is the study related to a prior research project that has been assessed by a recognised Medical Ethics Review Board (MERB) or FERB?
no
If so, which? Please state the file number:
no

B. SUMMARY OF THE BACKGROUND AND METHODS

Background

1. What is the study's theoretical and practical relevance? (500 words max.):

The results have both scientific and practical relevance. Although in many studies a found difference in learning between test formats is explained on basis of retrieval effort, so far only one study has tried to measure whether there was actually a difference in effort regarding the different test formats. For this reason, the present study might contributes to the scientific body of knowledge with concern to the effectiveness of retrieval practice and mental effort on study performance of learners.

The results can contribute to educational practice as well, as the results may indicate how students' performance can be increased during learning. Nowadays self-regulated learning (SRL) receives a lot of attention in education. SRL is an active process, in which the learner takes control and responsibility for their own learning (Zimmerman, 2002). Students could integrate retrieval practice while learning. In this way, they are able to master the subject matter independently while using retrieval practice as a learning strategy during self-studying (Roediger & Butler, 2011).

However, a study by McCabe (2011) shows that students are unaware of the effectiveness of retrieval practice as a learning strategy. Therefore, with the new insights offered by the outcomes of the present study, teachers may improve their learning programs by applying this form of testing in their lessons (e.g. flashcards, quizzes, and questions). In that way, students become familiar with retrieval practice and are able to apply this strategy during self-study. If students gain more insight on the benefits of this form of testing, they can learn in a more effective way, and because of the beneficial effects learners may receive higher grades (McDaniel et al., 2011).

2. What is the study's objective/central question?:
What is the relationship between different forms of retrieval practice and perceived mental effort, and how does this relate to performance of learners?
3. What are the hypothesis/hypotheses and expectation(s)?:

H1: Free recall yields the highest perceived mental effort followed by cued recall, followed by recognition, followed by restudying.
H2: Free recall yields the highest performance on the final test followed by cued recall, followed by recognition, followed by restudying.

Design/procedure/invasiveness
4. What is the study's design and procedure? (500 words max.):
Participants and Design

The power analysis with G*Power has shown that with a medium effect size of f =. 25 and with a power of .8 a sample size of 24 is needed. Originally, 44 Dutch participants were asked in February to participate in this experiment. All the participants were made aware of their participation in the experiment and were asked to give informed consent. Due to the within-subject design, all the participants experienced four conditions: cued recall, free recall, recognition, and restudying. The order of the conditions and the allocation of the word pairs to the conditions were randomized. The reason for this was to minimize the effects of the order of combinations and the word pair-condition combinations. Furthermore, participants were tested individually on personal computers. Two participants were excluded because of errors within the

experiment. For this reason, the final sample consisted of 42 participants, (31 female, 11 male) with ages ranging from 18 to 80 years old (M = 34.45, SD = .45).

Procedure

All participants were approached face to face or via mail (See Appendix C). The participants were asked to participate in the experiment. When the participants agreed, they filled in an informed consent form (See Appendix D). Because the experiment took place at different times and places, the researcher provided all information during the experiment to make sure that all participants received the same amount of information. The experiment took place digitally and was designed and presented in Gorilla (https://gorilla.sc). Furthermore, all the participants were tested individually in a quiet room.

The experiment consisted of two sessions as shown in Figure 1. In the study phase, participants were given an opportunity to study the word pairs one by one. The word list was divided into chunks of six word pairs, so these chunks are the different conditions (i.e., restudying, free recall, cued recall, and recognition). Each chunk was studied once, and immediately thereafter practiced 3 times in one of the four conditions. After that, the next chunk with the respective condition was presented in the same way. Every word pair was individually presented on the computer screen. For the restudying test, participants only restudied the word pairs one by one. For the recognition test, participants indicated whether they had learned the word pairs before on the computer screen. In the cued recall test, participants saw one cue word at a time and had to complete as many targets as they could. For the free recall test, the participants had to recall as many word pairs as they could by typing them on the computer keyboard. During this test, all participants indicated after each word pair how much effort it cost to recall or restudy them. After one week the participants were asked to log in again on the computer for the final cued recall retention test. Participants had to complete the given cue with the target via the computer keyboard.

5.

Which measurement instruments, stimuli and/or manipulations will be used?[4]:

[4] Examples: invasive questionnaires; interviews; physical/psychological examination, inducing stress, pressure to overstep important standards and values; inducing false memories; exposure to aversive materials like a unpleasant film, video clip, photos or electrical stimulus; long-term of very frequent questioning; ambulatory measurements, participation in an intervention, evoking unpleasant psychological or physical symptoms in an

Instrumentation

Performance. For the word pairs used as learning material in the experiment, the Carpenter's (2009) 'weak cue-target' list of 48-word pairs was used (See Appendix A). Half of the word pairs were used as learning material and randomly divided over the four conditions. During the learning phase, each participant learned 24 word pairs. During the recognition test, six word pairs that were not learned before were added and used as distractors. For each correctly recognized word pair, participants received 1 point and for every incorrectly recognized word pair 1 point was subtracted on the recognition test. To measure performance on the cued recall test in the learning phase, participants received the cue, and where then asked to type in the target on the computer keyboard. For each correct answer, participants received 1 point and for each incorrect or incomplete answer, they received 0 points. To measure performance on the free recall test in the learning phase, participants had to type in the word pairs on the computer keyboard. For each correct answer, participants received 1 point, for each incomplete answer participants received .05 points when one of the words was correct, and for incorrect answers, they received 0 points.

To measure performance on the final cued recall test, participants received the cue, after which the participants typed the target on the computer keyboard. For each correct answer, the participants received 1 point, and for each incorrect answer or incomplete answer, they received 0 points.

Retrieval effort. During the learning phase, all participants indicated after each word pair how much effort it cost to recall or restudy them. To measure the amount of effort participants experienced during learning of the word pairs, the unidimensional 9-point symmetrical rating scale, developed by Paas (1992) was used. Ranging from 1; very, very low mental effort, to 9; very, very high mental effort (See Appendix B). This scale has been widely used in many educational and psychology studies because of its reliability, sensitivity, and ease of use. (Paas, Van Merriënboer, & Adam, 1994; Paas, Tuovinen, Tabbers, & Van Gerven 2003; Van Gog, Kirschner, Kester, & Paas, 2012; Van Gog & Paas, 2008).

---

experiment, denial, diet, blood sampling, fMRI, TMS, ECG, administering stimuli, showing pictures, etc. In case of the use of a device (apparatus) or administration of a substance, please enclose the CE marking brochure for the relevant apparatus or substance, if possible.

What does the study's burden on the participants comprise in terms of time, frequency and strain/efforts?:
for the learning phase one quarter and for the final cued recall test one quarter.

Will the participants be subjected to interventions or a certain manner of conduct that cannot be considered as part of a normal lifestyle?:

no

Will unobtrusive methods be used (e.g. data collection of uninformed subjects by means of observations or video recordings)?:
no

Will the study involve any deception? If so, will there be an adequate debriefing and will the deception hold any potential risks?:
no
6. Will the participants be tested beforehand as to their health condition or according to certain disorders? Are there any inclusion and/or exclusion criteria or specific conditions to be met in order for a participant to take part in this study?:
no
7. Risks for the participants -
Which risks does the study hold for its participants?:
To what extent are the risks and objections limited? Are the risks run by the participants similar to those in daily life?:
I do not think that there are risks for my participants, the only "risk" i can think of , is that they may experience a sense of fear doing the "test". However, everyone will take the test in their own house/room, i hope it makes that less scary. Furthermore, most of the participants i am close with, so i do not think that they will be scared or experience a sense of fear.

8. How does the burden on the participants compare to the study's potential scientific contribution (theory formation, practical usability)?:

For the participants it will only 2 times take one quarter of their time, so i do not think that will burn them a lot. However, it might be that the results will show a significant effect. Which can contribute to science. I am sure that the participants who are willing to participate do not mind that this experiment will take a quarter of their time.

9. Will a method be used that may, by coincidence, lead to a finding of which the participant should be informed?[5] If so, what actions will be taken in the case of a coincidental finding?:
no

Analysis/power

10. How will the researchers analyse the data? Which statistical analyses will be used?:

Data analysis

To be able to perform the analysis, a few variables had to be computed. First, the effort score for every condition (i.e., free recall, cued recall, recognition and restudy) originally consisted of 18 scores. For each participant, all effort scores per condition were added together and then divided by 18. Second, for the variable performance in the learning phase, all scores per condition were added together and then divided by 18. The same applies to the variable performance in the final test, although this time, the total scores per condition were added together and then divided by 24. The raw data showed that 2 participants had duplicated the final test. Therefore, it was decided to delete the second final test. Furthermore, 2 cases were removed from the data set because during the experiment an error had occurred. In order to answer the research question, several analyses were conducted. All the analyses were conducted using IBM SPSS Statistics (version 25.0).

Second, ICCs were calculated in SPSS (Field, 2013) to determine the inter-rater reliability of the results of the learning phase and the final test. ICCs type "two way mixed" and "absolute agreement" were used (McGraw & Wong, 1996). Cichetti (1994) proposed standards for ICC results. The reliability of the ICC calculation is < .40 = low, .40 - .59 = reasonable, .60 - .74 = good, and .75 - 1 = excellent.

[5] For instance: dementia, dyslexia, giftedness, depression, extremely low heartbeat in an ECG, etc. If coincidental findings may be found, this should be included in the informed consent, including a description of the actions that will be taken in such an event.

Third, to investigate the effectiveness of retrieval practice and mental effort on study performance the experiment was conducted with a within-subject design. In order to answer the research question, the initial plan of the researcher was to conduct a one-way repeated measures ANOVA. However, because of the violation of the assumption of normality, the non-parametric Friedman test had to be performed. Significant results were followed up by the Wilcoxon signed-rank test.

The Wilcoxon signed-rank test is a non-parametric statistical test. It is used to compare two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ (Field, 2013). Several assumptions should be met in order to use the Wilcoxon signed-rank test. First, the assumption of independence was checked. Each participant should participate only once in the research and should not influence the participation of others. During the experiment, every participant only participated once. For this reason, the first assumption was met. Second, the dependent variable should at least be ordinal. As in this research all the dependent variables are continuous this assumption is met as well. Third, the distribution of difference scores between the two levels of the independent variables should be roughly symmetrical. Histograms of the difference scores confirmed that the symmetry of the distribution of difference scores were indeed roughly symmetrical. Therefore, the final assumption was met as well.

Last, to interpret the effect sizes, the rules of thumb from Cohen (1988) are used. According to Cohen (1988), $r = .1$ is considered as small, $r = .3$ is considered as medium, and $r = .5$ is considered as a large effect size.

11. What is the number of participants? Provide a power analysis and/or motivation for the number of participants. The current convention is a power of 0.80. If the study deviates from this power, the FERB would like you to justify why this is necessary:

> 42 for power analysis see above!

C. PARTICIPANTS, RECRUITMENT AND INFORMED CONSENT PROCEDURE

1. The nature of the research population (please tick):

1. General population without complaints/symptoms YES

2. General population with complaints/symptoms NO

3. Patients or population with a diagnosis (please state the diagnosis) NO

2. Age category of the participants (please tick):

18 years +

3. Does the study require a specific target group? If so, justify why the study cannot be conducted without the participation of this group (e.g. minors):

participants are all above the age of 18.

4. Recruitment of participants -

How will the participants be recruited?:

Via mail and/or face to face

How much time will the prospective participants have to decide as to whether they will indeed participate in the study?: one week

5. Does the study involve informed consent or mutual consent? Clarify the design of the consent procedure (who gives permission, when and how). Does the study involve active consent or passive consent? If no informed consent will be sought, please clarify the reason:

I will receive informed consent form, from all the participants if they agree to participate in my experiment.

6. Are the participants fully free to participate and terminate their participation whenever they want and without stating their grounds for doing so?:

Yes

7. Will the participants be in a dependent relationship with the researcher?:
Well ¨ dependent¨ . Most of the participants that participate in my experiment i know. But they are not dependent of me, at least i hope so.

8. Compensation
Will the participants be compensated for their efforts? If so, what is included in this recompense (financial reimbursement, travelling expenses, otherwise). What is the amount?
No
Will this compensation depend on certain conditions, such as the completion of the study?
No

D. PRIVACY AND INFORMATION

1.
Will the study adhere to the requirements for anonymity and privacy, as referred to in the Faculty Protocol for Data Storage[6]?:
anonymous processing and confidential storage of data (i.e. storage of raw data separate from identifiable data): YES
the participants' rights to inspect their own data: YES
access to the data for all the researchers involved in the project: YES

If not, please clarify.

Has a Data Management Plan been designed?
No
2.

[6] This can be found on the Intranet: https://intranet.uu.nl/wetenschappelijke-integriteit-facultair-protocol-dataopslag

Will the participant be offered the opportunity to receive the results (whether or not at the group level)?: Yes

Will the results of the study be fed back to persons other than the participants (e.g. teachers, parents)?: No

If so, will this feedback be provided at the group or at the individual level?

No

3.

Will the data be stored on the faculty's data server?: Yes

Will the data that can be traced back to the individual be stored separately on the other faculty server available for this specific purpose?:

No

If not, please clarify where will the data be stored instead?:

Because there is no other server! It will be stored on the faculty's data server.

E. ADDITIONAL INFORMATION

Optional.

F. FORMS TO BE ENCLOSED (CHECKLIST)

Text (advert) for the recruitment of participants
Information letter for participant
Informed consent form for participants
Written or oral feedback information (debriefing text)
(Descriptions of) questionnaires
(Descriptions of) measurement instruments/stimuli/manipulations
Literature/references

Signature(s):[7] Date and place: 23-01-19 Leiden

Name, position: Michelle Esmee Janssen, Student at Utrecht University

[7] The senior researcher (holding at least a doctoral degree) should sign here.