The methodology for research about ease of learning judgements
- does sequential and simultaneous judgements create different results?

Caroline Gremillet

Handledare: Fredrik Jönsson
C-UPPSATS I PSYKOLOGI, 15 hp, 2011

Stockholms Universitet
Psykologiska Institutionen

THE METHODOLOGY FOR RESEARCH ABOUT EASE OF LEARNING JUDGEMENTS - DOES SEQUENTIAL AND SIMULTANEOUS JUDGEMENTS CREATE DIFFERENT RESULTS?

Caroline Gremillet

Metacognition involves different evaluations of your own thinking- and learning process. Ease of learning (EOL) means judging how difficult for example a word is to learn. When researching about EOLs, different methods have been used in the past. In an experiment that tested glossaries, the methodology for measuring EOLs was investigated in this study. In the experiment, a comparison was made between making a sequential or simultaneous judgement of the difficulty of the words. A simultaneous judgement means judging one item while seeing the other items on the list, and a sequential judgement means judging while only seeing the word pair you are to judge. The result of this experiment was in line with previous research in that EOL judgments significantly, but only moderately so, predict the actual learnability of the items. However, the difference between a sequential and simultaneous judgement and their correlation with recall was not significant. There was a tendency towards better EOLs being made with simultaneous judgements, having a 30 % higher correlation with recall than sequential judgements.

Learning is a part of life and especially a part of a student's life. For some students it is not the lack of motivation that is keeping them from getting good grades. It is their lack of knowledge about effective study techniques and when to apply them (Roher & Pashler, 2010). To know this, a student needs to evaluate how difficult the study material is, and choose an appropriate method to study it. A lot of research has been devoted to learning in life, starting out with Pavlovs theory about conditioned stimulus and response, and Skinners theory about reinforcement (Passer et al, 2009). Unfortunately there is not so much research conducted on the metacognitive processes activated when learning study material. Most research within this field has been executed the past decade and it has now become a popular subject.

Memory

When the topic of memory and learning was approached in the 1960s the three-stage model (Atkinson & Shiffrin, 1968) appeared. It involved a sensory input to the sensory memory, where attention and encoding passed to short-term (working) memory and rehearsal occurred. A back-and-forth process between long-term and short-term memory involving encoding and retrieval takes place when the brain rehearses information. The depth of processing is crucial for information to pass over to long-term memory. On a structural and phonological level you *look* at, for example, the word you are trying to learn. On a semantic encoding level you *pay attention* to what the word really means, by for example trying to fit it into a sentence. Craik and Lockhart (1972) found support for the latter being an effective method when studying

and suggests that students organize hierarchical outlines or detailed flashcards in order to translate this level of processing into real world study techniques. When getting involved in elaborative rehearsal, looking closer at the meaning of the material and elaborating on it, is when you are truly on your way to mastering it.

Different aspects will be affected by how difficult a student perceives study material to be. One aspect is how that judgement will determine which method the student chooses to use to learn the material, as well as which parts of the material that will be studied. There are some study techniques that are far better than others to use (Rohrer & Pashler, 2010). For them to be effective for the student, the student needs to know how much time he needs to spend on studying the material. If the student makes an incorrect judgement about how hard the material is, the wrong study technique might be applied and the student might underestimate how much time that is needed for the material to be learned, resulting in bad grades. The process of estimating how difficult a subject is to learn and how much you have learnt is called a metacognitive process.

Metacognition

Serra and Metcalfe (2009) define metacognition as thinking about thinking, to have knowledge about your own thinking, and to be able to survey and control the metacognitive process. Nelson and Narens (1990) describe how metacognition works. Before studying for example words, one makes ease of learning (EOL) judgements about them, how difficult the words will be to learn. This will effect the selection of processing method of the information to-be-learned, the strategy for the kind of mnemonic process to apply. At the stage of acquisition, a person makes judgements of learning (JOL), which is a judgement about how much you have learned. From this, a person further allocates the study time at hand and decides what study techniques to apply. Then, at the retention stage, the person tries to maintain the knowledge previously studied, and makes feeling of knowing judgements (FOK), if the person can recognize but not recall the words. At the retrieval stage, the person uses a self-selected and self-directed strategy for retrieving information, and answers for example questions at a test. The entire metacognitive process can be divided into two parts. One part of the process is monitoring, a stage where a person makes EOL-, JOL-, and FOK-judgements. This stage is also called the prospective part (Nelson, 1996), where a person makes predictions about future performance. The other part is the controlling stage, where the person determines what to study and for how long, and which techniques to apply. After the person finishes her studies and completes a test, she can make a judgement of confidence for retrieved information (Nelson, 1996). Confidence of retrieved information means that the person states how sure she is about that the answers given on the test are correct. This stage is called retrospective where a person makes post dictions about performance (Nelson, 1996).

During the monitoring part of the metacognitive process, several important judgements are being made. Sheffer and Bar (2004) talks about JOL as a persons judgement about its own learning, which is important for a student to be able to tell which areas need to be studied further. It is of great importance that the student does not misinterpret the recognition of the material as knowledge “saved” in the long-term memory, making a JOL that is not corresponding with the person’s true knowledge of the material. JOLs are made to evaluate the current state of mastery during the monitoring stage, and to see if the current state of mastery is lesser than the degree of mastery desired (Nelson, 1996). If the current state of mastery is less than degree of

mastery desired the person will allocate subsequent study time during the controlling part of the metacognitive process. JOL is followed by a judgement of knowing (JOK), a judgement made after a study session regarding how well a person thinks he or she knows the material (Leonesio & Nelson, 1990) and FOK judgements about material that the person is not able to recall but perhaps recognize. JOK and FOK do not have a strong correlation, indicating that they are not based on the same cues thus tapping different aspects of the memory. Furthermore, delayed JOLs are more accurate than immediate judgements (Schwartz, 1994), implying that the same judgement at different times will be based on different sources.

One possible explanation for failure on an exam for students' who thought they were doing well, is that students sometimes misjudge which techniques that are effective (Roher & Pashler, 2010). This happens because some techniques that have been proved effective in the long run will in fact make one remember less while studying, thus making the student think it is an ineffective technique. This means that the student would make an incorrect judgement of learning during the study session when using certain techniques, resulting in incorrect conclusions on how to proceed.

Several aspects have been targeted when it comes to learning, among them are the so called test-effect, spacing-effect and mixed lists. Research (Carrier & Pashler, 1992; Cull, 2000) has shown that something called the test-effect, to test oneself on material just studied, is a lot more effective than for example re-reading a chapter. Passer et al (2009) clarifies this by an example: you have seen coins thousands of times in your life, but that does not mean that you can correctly paint or describe a coin if asked to pay attention to its details. Hence, simply repeating information is often a poor study strategy. Kang, McDermott and Roediger (2007) demonstrated through their experiments that students remember more when answering questions on tests with their own words, rather than answering multiple-choice questions, working on a deeper processing level in the controlling part of the metacognitive process.

Guérard, Hughes and Tremblay (2008) researched about the so-called von Restorff, or isolation, effect. The von Restorff effect means that people remember information better when they are distinct from its surrounding. Distinctive words or associations could be something bizarre, funny, unexpected or surprising. An example of this could be when words like flower, stone, cancer, sea, star, sex, shirt, blue, vomit and phone are presented to someone. The words cancer, sex and vomit will be better remembered than the other words, and this is called the von Restorff effect. In Guttentag and Carrolls (1998) study, low frequency words (words not occurring often in daily speech) were studied and tested together with high frequency words, and it turns out that people are not explicitly aware of the fact that low frequency words are easier to remember. That is, their EOL judgments were opposite to their actual performance after learning the material. As a result, people might misjudge the level of difficulty when for example learning glossaries, causing an incorrect monitoring process, leading to inappropriate use of study techniques and perhaps to little time spent on studying certain items.

Son and Metcalfe (2000) describe how students use metacognitive judgments to control their allocation of time while studying to determine what to learn. A discrepancy-reduction model describes the process of judging how and what to study. The model says that a person will first select what material to study, and then judge how difficult the material is. From there, the person will decide what level of

difficulty to study the most. During the time the person studies, he will constantly judge whether or not he has learned the material yet and continues to study until he has learned it, and then moves on to other areas. This process is called monitoring of learning. People exhibited a clear preference for studying the more difficult material over the interesting material if there is an expectancy for a test. Most study time will however be allocated to an intermediate state of learning, things that are not fully learned but not completely unlearned. This means that strategies and allocation might be changed when there are factors as time pressure, interest and motivation involved. This being said, performance might not be affected significantly just because a student changes her strategy when studying. Even though motivation could influence the strategy being chosen for studying, the mere intention is not the factor that will cause a different performance, it is the use of different cognitive strategies. In the experiments in Son and Metcalfe's (2000) study, the participants performed better on topics judged as interesting and easy, which means that people are good in telling what subjects they will have difficulties learning.

Ease of learning

The focus in this study is the monitoring aspect of the metacognitive process, and in particular, the type of judgment termed ease of learning (EOL). The methodology for research about EOL judgments is investigated. EOL means that a person judge how difficult a topic is (Son & Metcalfe, 2000) and can for example adjust their reading to spend more time on difficult to-be-learned material. Students attain significantly low to moderately high correlations when their estimations on how difficult a word will be to learn; their EOLs, gets correlated with their actual recall (Koriat, 1997; Keleman, Frost & Weaver, 2000; Leonesio & Nelson, 1990; Nelson, 1996; Schwartz, 1994; Son & Metcalfe, 2000; Underwood, 1966; Zechmeister & Bennett, 1991).

Methodological aspects when measuring EOL

Compared to other popular topics in memory research, there is not that much research carried out regarding EOL judgments. A flaw regarding current EOL research is that the studies are not executed in the same fashion making it is hard to draw conclusions about the validity of the predictive ability of EOL judgements. Obviously, the predictive validity of EOL judgments is not only a function of the participants' ability to correctly judge the ease or difficulty of learning a material, but also by methodological constraints imposed by the experimental setup. EOL judgements depends on factors such as how you perceive your memory in general, how many times you will be allowed to study certain material, which methods you will be allowed to use, if and what kind of test is expected and earlier specific experience (Koriat, 1997).

A problem with previous studies concerning EOL is that they are not conducted in the same fashion. Underwood (1966) even used different ways to judge the difficulty of a word in the same study. One way was to draw a line to show how difficult a word was, as perceived by the participant. The longer the line, the more difficult the word was according to the participant. Other participants were in the same study told to assign the value 100 as the level of difficulty to one word in the word list that the participant thought would be of medium difficulty degree to learn. All other words were to be given a proportional number, for example 200 if twice as hard to learn, and less than 100 if easier than the word with the value of 100. In Son and Metcalfe's study (2000), they used a scale from one to eight, ranging from most difficult to

easiest item to learn. Except for the difference in which rating scale is used and how the judgment is made, studies about EOL have different numbers of study trials, and different conditions for them. In Underwood's (1966) study, the participants were not even told about the first upcoming test, which makes the conditions for the tests different compared to studies (Leonesio & Nelson, 1990; Zechmeister & Bennett, 1991) where the participant was told about an upcoming test. As concluded earlier, strategies might change when aware of a forthcoming test, and can therefore influence the result.

The amount of time the participants were allowed to study the words, and, if part of the experiment, the amount of time spent on distractions was different in each study. All this makes it harder to compare studies with different or similar results. Also, previous studies do not take the possible different effects of simultaneous and sequential judgments about difficulty of words into consideration. Earlier studies have not mentioned that there might be a difference between those two conditions: judging difficulty of a word simultaneously, while looking at all the other words they will study too, and sequentially, just looking at the word they are to judge. The judgement type will be further investigated in this study, to see if there in fact is a difference between making a simultaneous judgement and a sequential judgement.

Studies regarding EOL have been using different dependent and independent variables. In Kelemen, Frost and Weavers (2000) study, they used the number of trials to learn each item as a dependent variable, arguing that it is a more sensitive measure and thus making it a better criterion measurement than for example correlations with judgements about rated difficulty of items. In this study, both ways of making EOL judgements will be tested to see if it makes a difference whether you ask the participant to judge how difficult the item is, and when the participant will have learned the item (regarding their correlation with recall). Zechmeister and Bennett (1991) used the rate of presentation of to-be-judged items as their independent variable, varying the amount of time the participants had to judge the items. In this study, the independent variables will be about the difference between making a sequential and simultaneous judgement.

The aim of the study

The overall purpose is to establish predictive validity by looking closer at possible differences in correlations, so all studies conducted in the same field can be compared on the same basis. The purpose is to investigate the correlation between ease of learning judgements and actual recall to see if the correlation is affected by making EOLs in different ways, sequentially or simultaneously. Furthermore, the ease by which a participant will learn an item will be studied in two ways, by asking the participant to rate the level of difficulty of the item by assigning it a value, and by saying at what point they will know the item. The purpose of this is to see if the different ways of phrasing the question will influence the judgement being made, and thus creating a different correlation with recall. The study will also be measuring the overall correlation between EOL and recall at the different tests given in the experiment, and between EOL and trials to acquisition.

The methodology of the current study is designed to create a uniform way of measuring EOLs to establish predictive validity about whether peoples' EOL are corresponding with their actual recall to a larger extent or not. Students will be asked

to rate the level of difficulty with which they will learn a specific word pair. They will also be asked at which test they think they will know that item. Then, the participants will practice on the glossaries and be tested on all of them. When tested, you can tell whether or not they could judge correctly how difficult a word would be to learn, i.e. the accuracy of their EOLs, by looking at the correlation between EOLs and recall. The questions posed before the study session about how difficult a word is and when the participant will know the word basically ask the same thing: about the ease with which they will learn the items. However, a hypothesis is that not all participants will realize this, and they might therefore respond differently on those two questions, as suggested by Keleman Frost and Weavers (2000). In future references, this will be called phrasing of question. It is important to see if the participants perceive the questions as identical, to be able to see if they understood what judgement was asked of them and if that affects the correlation with recall.

The participants will make judgements about the level of difficulty of the items in two different ways. In one session they will make that judgement sequentially, meaning they will only be looking at the word pair they are judging while judging the level of difficulty. In the other session they will look at the items simultaneously, meaning they will be able to see the entire list of word pairs while making a judgement about an individual item. The condition for which type of judgement they make will in this study be called Judgement condition. This is a key element in this study, for earlier studies make no difference between judgement conditions and their possible effect on the predictive validity of the EOL judgments, that is, the correlation between the EOL judgements and later recall.

Simultaneous EOLs are more common in real life, where a person would for example estimate the entire list of word pairs spontaneously before studying it, and not just by looking sequentially, separately, at every single word to judge the level of difficulty. Therefore, one hypothesis is that the simultaneous judging type will be superior to the sequential judging type because people have more experience in making those judgements. Furthermore, if this proves true, future studies should consider two things. First, they can no longer randomly choose one judgement type and then generalize the material as if it did not matter which type they had used, and which type other studies used. Second, simultaneous judgements should be used in experiments where the purpose is to make generalizations to real life settings where those judgement types in fact are used more often than sequential ones. As mentioned earlier, different cues are used when making different judgements, and the same judgement at different times. This might be the case when having, or not having, other items to compare the level of difficulty with, thus affecting how well that judgement correspond with actual recall.

Method

Participants

In the study, the participants were 29 students (23 females) participated with a mean age of 25.07 years ($SD = 5.92$, *range*: 19 - 45). Students were recruited at Stockholm University by a note on a billboard. The note informed them that students participating in the experiment would receive either course credit or a ticket to the movie theatres. All the students currently enrolled at the Department of psychology at

Stockholm University received an email about the experiment with the same information. Other students were found by a convenience sample and attended Stockholm University, Stockholm School of Economics and Södertörn Högskola. Students from the convenience sample received the same information given on the billboard and in the email.

Before the experiment took place all the participants were informed that they could withdraw from the experiment whenever they felt like it without having to explain why, and signed a written informed consent. They were also informed that all instructions were given on a computer, and that they were allowed to ask questions before and after the experiment. Since the experiment was conducted in Swedish, and e-prime can only read English letters when typing in an answer, the participants were instructed not to use the Swedish letters “å, ä, ö” but to replace those letters with “a, o”.

The participants were divided between four different versions of the test that was distributed to the participants. The four different versions were combinations of the two word lists and the two judgement types. The test consisted of two different lists of words and about 50 per cent of the participants received a version where one list of words came first, and the other 50 per cent got the same list in their second session. The purpose of this was to prevent systematic effects of the lists used. Another difference between the versions distributed was that one group did sequential judgments about the glossaries first, and the other group did simultaneous judgments first. Participant number one got the first version, participant number two got the second version and so on. In sum, list order and task order was fully counterbalanced.

Apparatus

The computer software e-prime 2.0 (Psychological software tools, Oittsburgh, PA) was used during the entire experiment, and it was conducted on a HP computer (PC).

Two word lists with 20 word pairs each were used in the experiment. An example of a word pair is “tabibu” (Swahili, meaning chiropractic) and “doktor” (Swedish, meaning doctor) if the combination was Swahili-Swedish, and “Adhd” and “problem” if the combination was Swedish - Swedish.

Procedure

EOL rating phase

First each participant was asked to judge the word pairs. The first question was “how difficult do you think this item will be to learn?”. Answers were given on a scale from one to six, where one represented difficult and six represented easy. All these options were displayed on the screen at the same time, and the participant typed in a number to rate the difficulty of the item. The word that they were to learn was always the right word in the item displayed on the screen. On each list, ten of the word pairs were Swedish-Swedish combinations, and ten were Swahili-Swedish combinations. The second question was “by which test will you correctly recall the item?”. The options to make this prediction was: on the first, second or third memory test, or not even by the time of the third test.

There were two different tests distributed to every participant and about half of the participants started with a session where they could make these judgments simultaneously, meaning being able to look at the entire list of word pairs to the left

on the screen while judging the difficulty of the items. The other half started their session with the sequential judgement type, where they only saw the item they were currently judging while making the judgement about the level of difficulty. When they completed one session they were given the other version with a new set of words pairs and another judgement type. In total, there were four versions combined by the two words lists and two judging types.

Study phase and test phase

After the EOL phase, the participants studied the word pairs for five seconds each, in total 80 seconds. The word pairs were presented in a random order in every study session, and the computer automatically switched to another word pair after five seconds had passed. After the study session, the question “how many word pairs do you think you will remember?” emerged on the screen, and the participant could choose a number from zero to twenty to type in, ranging from knowing none of the words to knowing all of them.

This was followed by equations and the participant was encouraged to not answer them until sure about whether they were correct or not. One meant yes it is correct, zero meant no it is not. However, the participant was told that after 15 seconds the time would be up, and another equation would emerge if the previous one had not been answered. If the equation was answered before the 15 seconds were up, a new equation emerged and so it went on for 45 seconds. An example of an equation is: $65 \times 11 = 705$. After that, the computer automatically moved on to the next step in the experiment. This section was inserted into the experiment solely to prevent short-term memory effects from influencing answers given on the test if the test had been given immediately after the study session. Therefore, this is not a part of the results and will in no way be analyzed.

When finished with the equations, the first test was given. The participant was given 15 seconds to type in the word that during the study session had been paired with the word presented on the screen. The participant was shown one of the words in the word pair and were to type in the other one before 15 seconds had gone by. If the participant had not answered before 15 seconds had passed the word disappeared and was replaced by another word from the study session and assigned the participant zero points for the previous word.

When the test was finished, the participants solved equations for 45 seconds. Then, they were to study the word pairs for five seconds each again, and type in how many words they thought they would remember on the following test, as well as completing equations before taking the second test. After completed test number two, which had the same procedure as test number one, equations were showed for 45 seconds, followed by another study session. Then, they typed in how many words they thought they would remember on the last test. Equations were then solved, and then the participants took test number three.

After test number three was completed, the screen displayed a watch and a text informing the participant it was time for a small break before continuing. After the break, the whole procedure was repeated with a new list of words, consisting of 20 word pairs as well. The difference was now the judgement type. If the participant had made judgements about the words while being able to see the entire word list before, judging simultaneously, they were given the sequential test where those judgements

were made without the possibility to see all the other words while judging the difficulty of the words. Those who had made judgements sequentially in the first session were given the test with simultaneous judgements in the second session.

The experiment lasted for approximately 40 minutes per person. Some participants participated simultaneously, but were separated with a wall shielding them from each other.

Data processing

The two scales for question one and two in the experiment were scaled in the opposite direction. The first questions degree-of-difficulty rating of the item went from hard to easier EOL, the second question went from easy to harder judgement about the difficulty of the word, resulting in having opposite meanings for the numbers one to six. For this reason the first scale was inverted before analysis to match the responses given on the second scale, before correlating EOLs and recall/trials to acquisition.

Correlational analyses were made in SPSS 16.0 between EOL and recall. The Goodman Kruskal gamma (Nelson, 1984) correlation was used to calculate the correlation for test one between EOL and recall. There were three tests for recall per session in the experiment, and two sessions in total. Because recall was close to 100 % on memory test number two and three, making them a constant variable, they were excluded from the analysis. The correlation for test one was calculated between simultaneous judgement and recall for question one and two, as well as sequential judgement and recall for question one and two.

Results

In all of the analyses an alpha level of .05 was used as the criterion for statistical significance. The data was entered into a 2 x 2 repeated measures ANOVA with judgment condition (sequential/simultaneous) and phrasing (Question1/Question2) as independent variables and the gamma correlation between (i) the EOLs and the first cued recall test and (ii) trials to acquisition as the dependent variables. The overall gamma correlation between EOL judgements and recall, without consideration to judgement type and phrasing of question, was $G(20) = .45$, $p = .05$, using data from 28 of the participants.

The correlation between EOLs and recall performance at the first test

In the first part of the analysis, EOLs for all the conditions below were correlated with actual recall on the first test. There was no main effect of Question, meaning that neither the first nor the second way of eliciting EOLs had any impact on the predictive validity of the EOL judgments ($F(1,27) = .15$, $\eta_{\text{partial}}^2 = .01$, $p = .71$). The mean gamma correlation between EOLs and recall at the first test for Question one, asking the participants to rate the level of difficulty of learning the items, was .44 ($SD = .04$), whereas for Question two it was .46 ($SD = .04$). There was no main effect of Judgment Type, $F(1, 27) = 1.78$, $\eta_{\text{partial}}^2 = .06$, $p = .19$. Simultaneous presentation ($M = .51$, $SD = .05$) was associated with a nominally higher correlation, than the sequential presentation ($M = .39$, $SD = .07$). The interaction ($F(1,27) = .03$, $\eta_{\text{partial}}^2 = .00$, $p = .86$) was not significant.

The correlation between EOLs and trials to acquisition

In the second part of the analysis, EOLs for all the conditions below were correlated with trials to acquisition. There was no main effect of Question, meaning that neither the first nor the second way of eliciting EOLs had any impact on the predictive validity of the EOL judgements ($F(1,27) = .19$, $\eta_{\text{partial}}^2 = .01$, $p = .66$). The mean gamma correlation between EOLs and trials to acquisition for Question one was .50 ($SD = .05$) and .53 ($SD = .05$) for Question two. There was no main effect of Judgement type, $F(1, 27) = 2.83$, $\eta_{\text{partial}}^2 = .10$, $p = .10$, although there was a tendency. Simultaneous judgements ($M = .60$, $SD = .06$) were associated with a nominally higher correlation than the sequential judgements ($M = .44$, $SD = .07$). The interaction ($F(1,27) = .00$, $\eta_{\text{partial}}^2 = .00$, $p = .96$) was not significant.

Discussion

The purpose of this paper was to study EOL judgements and see whether there was a difference between sequential and simultaneous judgements and their correlation to recall, and whether the phrasing of the EOL question mattered. This was tested in order to establish a methodology for EOL experiments. The idea behind looking at the difference between judgement types, and ways of phrasing the EOL judgement, was to attain predictive validity. In that way, a study can be compared with other studies within the same field when executed in the same fashion, i.e. with the same judgement type, if there was a difference between the two types investigated in this study.

Metamemory accuracy

As stated earlier in this paper, previous research indicates that people attain low to moderately high correlations between their EOL judgements and actual recall (Koriat, 1997; Keleman et al., 2000; Leonesio & Nelson, 1990; Nelson, 1996; Schwartz, 1994; Son & Metcalfe, 2000; Underwood, 1966; Zechmeister & Bennett, 1991). The present study was in line with research in showing that, when not considering judgement type nor phrasing of question, the mean correlation between EOL and recall was moderate.

Judgement type

Furthermore, what kind of judgement type is used in the experiment appears to have importance, even though the result was not significant in this study. Comparing the correlation for sequential judgements for question one and two with simultaneous judgements, one can see that the judgement type does matter. There was a clear difference between simultaneous and sequential judgement types: the simultaneous judgement generated on average a 30 percent higher correlation with recall for both questions. One can assume that simultaneous judgements correspond better with recall because simultaneous judgements are made more often in real life than sequential ones, which is why it is important to differentiate the judgement types when researching about EOL. The effect size for this sample difference, calculated with Cohen's d , was .7, indicating that the true effect was rather large.

The most likely reasons for not attaining a significant result in the current study is in part a power problem due to too few participants, and the ceiling effect due to too easy words. Future research should use a larger sample as well as increasing the level of difficulty of the items. The correct answer rate was close to 100 percent already on the first test in this study, and should be around 50 percent. This means that in future

research more difficult words need to be chosen for the word pairs, to be able to analyse test number two and three as well. When choosing the level of difficulty, several things need to be considered. The point at which the participant makes a judgement about the difficulty of the word is itself a study session and will therefore make it easier for the participants to learn the items. The students learn more about the material being studied if they test themselves on the material while studying, the so-called test-retest effect, which should also be considered when choosing the level of difficulty. Also, Kang, McDermott and Roediger (2009) concluded in their experiment that students learn more from filling out the answer themselves rather than just recognising words (multiple choice questions), which makes this test a study session in itself. Another topic discussed was the von Restorff effect, where words that stand out are more easily remembered. Swahili words stand out a lot compared to the Swedish vocabulary, so the level of difficulty can be increased when considering which Swahili words to include in the experiment.

The methodology for research conducted in the same field needs to be constructed in the same manner to be able to draw conclusions about correlation and causation. Differences have already been pointed out earlier in this paper between some of the studies testing the EOL-judgment. One difference regarding the design of the experiments is the placement of the recall-test. In Leonesios (1990) study, the test is held four weeks after the study session, in Brittons (1991) study the test is one day after the study session, while in Guttentag (1998), Underwood (1966), Son and Metcalfe (2000) studies the test is being held immediately after the study session. The placement alone can affect the test result and therefore distort EOL correlations with recall. In the present study, the test was held immediately after the study session leaving room only for the distractions, the equations, lasting 45 seconds to prevent the short-term memory from influencing recall. In future studies, it is suggested that the same design is used when EOL judgements are not correlated specifically to recall in the distant future.

The true effect size should also take part in determining how easy the effect will be to discover in an experiment. The difference between simultaneous and sequential judgements correlated with recall might not be that much larger than the 30 percent-difference found in this study, just large enough to be significant. If the true effect size is small, making it harder to discover, all the factors mentioned above could influence the correlation which therefore makes it necessary to be cautious about the details of the design of a study.

The purpose of this paper was not to decide that only one of the judgement types should be used in *all* EOL experiments. It was to show that different judgement types entail different results, which have to be taken into consideration when comparing the result from a study to either another study executed with another judgement type, or generalizing to a real life situation where another judgement type was used.

Phrasing of question

There was another purpose with this experiment, to see if the participants perceived question one and two as two different questions, and therefore responding differently, one perhaps being a better tool to use when measuring EOLs. The means for question one and two were almost identical, indicating that the participants did not perceive the questions as two different ones.

In this respect, it makes no difference which of the two ways of phrasing EOL you correlate with recall, it does not affect the correlation at all. In all, EOL judgements will not be affected by how you ask the participants to make their EOL judgements.

Conclusions

Even though the difference between judgement types was not significant in the experiment carried out in this study, there was a tendency towards better EOLs being made with the simultaneous judgement type, prompting for a strict use of either simultaneous or sequential judgements when conducting research about EOLs. By executing research on EOLs in a strict manner, carefully designing each part of the experiment, it might influence the perception researchers have on the predictive ability of EOLs.

References

- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *advances in the psychology of learning and motivation: research on theory* (Vol. 2). New York: academic Press.
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and cognition*, 20, 632-642.
- Craik, F. I. M, & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Experimental Psychology: general*, 104, 268 – 294.
- Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14, 215-235.
- Guérard, K., Hughes, R. W., & Tremblay, S. (2008). An isolation effect in serial memory for spatial information. *The Quarterly Journal of Experimental Psychology, Psychology press*, 61, 752-762.
- Guttentag, R., & Carroll, D. (1998). Memorability judgements for high- and low-frequency words. *Memory & Cognition*. 26, 951-958.
- Kang, S. H, McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology*, 19, 528-558.
- Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability, *Memory & Cognition*. 28, 92-107.
- Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgements of learning. *Journal of Experimental Psychology: General*, 126, 349 – 370.
- Metcalf, J. (2009). Metacognitive judgments and control of study. *A Journal of the Association for Psychological Science*, 18, 159-163.
- Nelson, T. O. 1. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95, 109–133.
- Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, 51, 102-116.
- Nelson, T. O. 1., & Narens, L. (1994). Why investigate metacognition? In Metcalfe, Janet, Shimamura, Arthur P. *Metacognition: Knowing about knowing*. (pp. 1-25). Cambridge, MA, US: The MIT Press.

Passer, M., Smith, R., Holt, N., Bremner, A., Sutherland, E., & Vlieg, V. (2009). Psychology: the science of mind and behaviour. Berkshire: McGraw Hill.

Richter, L., & Kruglanski, A. W. (1998). Seizing on the Latest: Motivationally Driven Recency Effects in Impression Formation , *Journal of Experimental Social Psychology* 34, 313–329.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, 39, 406-412.

Son, L., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 204-221.

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.). *Handbook of Metacognition and Education*, 278-298. New York, NY: Routledge.

Zechmeister, E. B., & Bennett, D. J. (1991). How easy is it to judge ease of learning? *Bulletin of the Psychonomic Society*, 29, 36-38.