The Relationship Between the Extent of Mental Model and Most Effective Timing of Feedback

Master Thesis Educational Design and Consultancy

Utrecht University, 2016

Course code: 200700224

Beitske Verheij, 3888967

Tutor: Dr. Noortje Coppens

Second assessor: Vincent Hoogerheide MSc

Abstract

Tests do not only assess what was learned, but they also enhance learning itself. This testing effect is even stronger if the learner receives corrective feedback after a retrieval test. Research concerning the optimal timing of this feedback shows various results. The number of errors could be crucial. The more errors have been made, the more beneficial immediate feedback seems to be; the less errors have been made, the more beneficial delayed feedback seems to be. Initial test score is an indication for the extent of the learner's mental model about a subject. Moreover, response certitude seems to be an indicator for the extent of mental model. The research question of this study is: to what extent does the size of one's mental model of a subject determine when corrective feedback can be given best?

In this quantitative experiment, 112 nine years old students wrote down concepts related to World War II. This subject was not taught at school. Then a test about the subject was presented. One group of students got corrective feedback immediately after answering a question, the other group received the feedback after another lesson. All students indicated how sure they were of each answer. The next day, all students did the same test again.

A multiple regression analysis was conducted, with feedback condition and extent of mental model (i.e., number of concepts written down before the first test, initial test score and mean response certitude) as independent variables and learning gain as the dependent variable. The research question concerned the interaction effect of condition and extent of mental model. This interaction effect was nonsignificant.

In this experiment, extent of mental model and optimal feedback timing were not related. Possibly, this can be explained as follows. Although extent of mental model differed a lot amongst students, all had some knowledge about the subject. Possibly, the crucial question is whether an intrinsically right answer to a test item exists or not. If students know that test items are real, they will try to retrieve the answer from long term memory. If they know that the correct answers cannot be known beforehand, taking a test is initial learning, whereby new information should be processed via the working memory. In such cases, immediate feedback might contribute to this initial processing, which could explain better results after immediate feedback in such experiments.

Key words: mental model; feedback timing; testing effect; spacing

In education, students are expected decreasingly to listen to their teachers passively and just read books (Kumpulainen, Mikkola, & Jaatinen, 2014). They are involved more and more in elaborating activities, such as retrieval tests during processing knowledge (Prince, 2004). They are also tutored more adaptively, not only by their teachers, but also through digital devices such as interactive educational computer programs, online learning material and games (Kostolányová & Šarmanová, 2014). Feedback on such learning activities is important (Hattie & Timperley, 2007), and the timing of feedback seems to influence results too (Kulik & Kulik, 1988). For both teachers and designers of digital devices it is important to know whether they should pay special attention to the timing of feedback, and whether they should make it possible to adapt this timing to individual learners. This study addresses the question which corrective feedback timing after retrieval tests concerning factual knowledge is optimal for which learners.

The testing effect

In traditional views, learning occurs during episodes of studying, and tests serve to assess what was learned (Roediger & Butler, 2010). Nevertheless, according to a meta-analytic review of Rowland (2014), many studies demonstrate that taking tests about previously presented information (*retrieval practice*) is also a learning activity, and even enhances long-term retention more than repeated restudying the information during the same amount of time (see also Roediger & Butler, 2010). This phenomenon, that has been called the *testing effect*, has been explained by the theory that through testing more elaboration occurs (Bjork, 1975). Although the testing effect has been tested mostly on college students, several studies indicate that the testing effect occurs amongst elementary school children as well (Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Roediger & Butler, 2010; Rowland, 2014). Nevertheless, the testing effect does not show in all types of information to be learned. In their study, Van Gog and Sweller (2015) indicate that the testing effect disappears if the learning material gets more complex and several elements of the information are related to each other, for example when it concerns the operation of a machine.

Retrieval effort theory

In accordance with the idea of elaboration, many theorists argue that the magnitude of the testing effect (as evidenced by the scores in the posttest) increases with the difficulty of the first retrieval practice test (Rowland, 2014). This view is based on the assumption that the learning material has been presented before the (first) retrieval test, that many items of the retrieval test have been retrieved correctly, and that no feedback has been given. Rowland (2014) calls the ideas of these theorists *retrieval effort theories* and refers to many studies confirming these theories (see for example Pyc & Rawson, 2009). Bjork and Bjork (1992) distinguish two elements in memory: *storage strength*, i.e., the degree to which a memory is durably established, or 'well learned', and *retrieval strength*, i.e., the accessibility of a certain memory at a given point in time. According to the *desirable difficulty framework* theory, retrieving memories with low retrieval strength, which is difficult, enhances the storage strength of memory (Bjork, 1994).

Spacing

Roediger and Butler (2010) mention some additional crucial aspects of retrieval practice in their literature review. Repeated retrieval seems to be more effective than only one retrieval test. Furthermore, the testing effect is weaker when retrieval tests are done shortly after one another than when there is some delay between the tests. This is consistent with the *spacing hypothesis* (Smith & Kimball, 2010), according to which different learning events should be distributed over time to be most effective, and thus should not take place immediately after one another (massed). This seems to be consistent with the desired difficulty framework (Bjork, 1994) as well: retrieval tests that are taken shortly after one another are relatively easy, because the information is fresh in mind and therefore easy to retrieve, which results in a high retrieval strength. This decreases the storage strength and thus the testing effect.

Soderstrom, Kerr, and Bjork (2015) state that the testing effect as such might be mixed up sometimes with the effect of spacing, because in the studies to which they refer, there was more spacing in the testing conditions than in the restudy conditions.

Corrective feedback

Concerning corrective feedback, Roediger and Butler (2010) indicate in their literature review some differences in testing effect research. In some experiments corrective feedback (i.e., providing the correct answer, in this research field) has been given in the retrieval practice conditions, in some experiments it has not. Several experiments focused on feedback conditions: within these studies, some participants in retrieval practice conditions were given feedback and others were not (Kang, McDermott, & Roediger, 2007; Metcalfe, Kornell, & Finn, 2009). Many studies show that when after retrieval practice corrective feedback was given, retention is better than in no-feedback conditions, and thus the testing effect is strengthened. (Roediger & Butler, 2010). This has been explained as follows: in retrieval tests without feedback, correctly retrieved information will be stored more firmly in memory (than during restudying), but forgotten information will be forgotten more definitively (Rowland, 2014). Furthermore, in recognition tests (multiple choice and true/false items) one is confronted with incorrect information, and this incorrect information may be remembered as correct if it remains uncorrected (Butler & Roediger, 2008). In a test with feedback, the learner is confronted with the whole learning content anew, not only with what he or she can remember of it. Surprisingly, without feedback the testing effect remains significant: also without feedback (so when only partially rehearsing the material), retrieval practice enhances long-term retention more than restudying, during which one is confronted with the whole learning content, as during a retrieval test with feedback (Butler & Roediger, 2008). This confirms that during testing, the information is processed more deeply than during restudying.

Timing and functions of corrective feedback

Research has addressed the question whether corrective feedback should be given immediately after each test item or should be delayed, with different conclusions (Kulik & Kulik, 1988; Smith & Kimball, 2010). These differences are not well understood.

Kulhavy and Stock (1989) describe several functions of corrective feedback. The first function is *verification*: was the initial answer correct or incorrect? If the answer was correct, it is only being reinforced by this verification. If the answer was incorrect, more elaboration is needed. The incorrect answer should be eliminated by the feedback, replaced by the correct answer, and this correct answer should be retained. This could mean that the optimal timing of feedback depends of the actual function of feedback, and therefore on the question whether the feedback comes after either a correct or an incorrect answer (which unfortunately cannot be known beforehand).

Theories about the optimal timing of feedback after wrong answers are contradictory: according to behaviouristic theories, errors should be corrected as soon as possible; on the other hand, according to Kulhavy and Anderson's *perseveration-interference theory* (1972), a longer delay of feedback is beneficial when errors have been made, because during a longer delay one forgets the wrong answers on the initial test. If this is true, however, one could argue that correct answers could be forgotten during the delay as well (because the learner does not know yet whether the given answer is correct or incorrect), and consequently that delayed feedback could eliminate the testing effect entirely. Indeed, the perseveration-interference theory was not supported by the study of Smith and Kimball (2010). From their experiments with different feedback timings, they draw the conclusion that immediate feedback is more effective after incorrect answers, but after correct answers delayed feedback enhances retention more. Kulhavy and Anderson (1972) call this the *delayed retention effect* (DRE). If receiving feedback is considered a learning event, this seems to be consistent with spacing theory (Butler, Karpicke, & Roediger, 2007; Smith & Kimball, 2010). The learner is confronted with correct learning material several times with time gaps between them. It should be noted, however, that in this view, spacing is inherent to delayed feedback after correct answers, in contrast to immediate feedback. One can question to what extent this contrast exists in reality. Within feedback timing experiments, this contrast does exist, because then delayed feedback is contrasted with immediate (massed) feedback. On the other hand, spacing can concern all kinds of learning events, not only testing and feedback. In real life, a spaced activity could be added some time after immediate feedback, as another presentation of the learning material. In short, within experiments the effect of spacing and the effect of the delay of feedback as such cannot be distinguished clearly.

Metcalfe et al. (2009) investigated effects of immediate versus delayed feedback on retention (of meanings of for the population uncommon but real words) in elementary school and amongst university students in a laboratory setting. Their actual aim was to denounce the observations of the meta-analysis of Kulik and Kulik (1988) that in most laboratory settings, delayed feedback enhances retention more than immediate feedback, but in school settings immediate feedback works better, possibly because in the classroom students pay less attention to delayed feedback (Butler et al., 2007). Metcalfe et al. (2009) obviated this problem in their procedure by obliging the school children in both conditions to copy the feedback. Hence, they concluded from their data that also in class delayed feedback is more effective than immediate feedback. Surprisingly, final test results of the college students showed no difference between immediate and delayed feedback effects.

Metcalfe et al. (2009) suggest that this difference in outcomes can be explained by the fact that the elementary school children made significantly less mistakes in their initial test than the university students in theirs. They suggest that the number of commission errors (i.e., incorrect responses to test items) in the first retrieval test determines whether feedback should be given immediately or delayed. This would be in accordance with behaviouristic theories and the findings of Smith and Kimball (2010), because the group which had made more errors benefited less from delayed feedback. Nakata (2015) conducted an experiment to test Metcalfe's (2009) hypothesis that the more commission errors have been made, the sooner feedback should be given. However, in Nakata's experiment this hypothesis was not confirmed, possibly because delayed feedback was delayed only 90 seconds, whereby the spacing effect of delayed feedback was limited, and because just before the final test a final review moment for all conditions was inserted, possibly diminishing the influence of conditions on the results.

Mental model

Based on the above, we might conclude that after incorrect answers, immediate feedback probably is most effective, and after correct answers, feedback should be delayed. However, Kulhavy and Stock (1989) give an important warning. Scores at tests do not necessarily reflect student's real knowledge. Initially correct responses at recognition tests of course can stem from correct knowledge, but could also just be guesses that might be forgotten at a second test. On the other hand, causes of initially incorrect answers can vary between an oversight and a complete non-understanding of the material. Furthermore, they argue that feedback received after a wrong answer that has been given with what they call high *response certitude*, is more likely to be elaborated on than feedback received after an answer given with low response certitude, either correct or incorrect. This elaboration could be just as effective as effortful retrievals of correct information, as intended in retrieval effort theories (Rowland, 2014).

Fazio and Marsh (2010) observe the same phenomenon: the more someone is convinced of giving the correct answer, although it is wrong, the better the correct answer will be remembered after corrective feedback. They suggest that the conviction to be right comes from an activated mental model, and after receiving corrective feedback, the right information can be attached easily to this activated mental model. This phenomenon has been called the *hypercorrection effect* (Butterfield & Metcalfe, 2001). Conversely, after right or wrong guesses, correct answers coming from feedback are much more difficult to retain than after answers given with high response certitude.

Clearly, response certitude seems to be related to the extent of somebody's mental model of the subject to be learned. If this mental model is absent, one has a very low response certitude answering the questions of a test, or, in other words, one has to guess. Extent of mental model is determined by both the background of the learner and the kind of information being learned.

Kornell (2014) distinguishes between meaningful and less meaningful information to be learned in relation to optimal timing of feedback. In his view, the extent of existing mental models about the new information might be the crucial factor. He argues that the more one already knows about a subject, the more meaningful new information is, and the more beneficial delayed feedback is, compared to immediate feedback. Kornell (2014) investigated various types of information being learned, in relation to timing of feedback. In his first experiment arbitrary word pairs, constructed especially for the test, had to be learned. In this experiment immediate feedback turned out to be more beneficial, possibly because the information to be learned could not be known beforehand. In a second experiment, the information to be learned (trivia questions) was more meaningful, and could be associated with prior knowledge. Thus, participants had more opportunity to give the correct answer in the initial test than participants in the first experiment. In this case, feedback delayed one day turned out to enhance retention just as much as immediate feedback. It should be noted that in this second experiment, participants' prior knowledge about the contexts of the trivia was not measured and possibly varied a lot, which could explain why conditions of feedback timing did not make any difference in the sample as a whole. Kornell (2014) explains the different outcomes of his two experiments as follows: answering questions about meaningful information requires more elaboration, and prior knowledge will be activated, which takes time. The reason why delayed feedback after questions about meaningful information seems to result in a better retention than delayed feedback after meaningless questions, could be that after meaningful questions the mental model remains more active as long as the question is not answered definitively. Conversely, when no mental model exists, delay of feedback is not beneficial.

Research question

The present study aimed to answer the following question: to what extent does the size of students' mental models of a subject determine when corrective feedback on retrieval practice of new information about the subject can be given best?

Within this study, the concept mental model is defined as the ideas and knowledge someone has about a subject. Linking up with the hypercorrection effect (Butterfield & Metcalfe, 2001), these ideas and this knowledge do not need to be correct to be part of the mental model, because also incorrect knowledge can give clues to retain new information. Therefore not only correct answers in the first test arise from the mental model, but also response certitude. Furthermore, (correct or incorrect) concepts written down without cues before the first test are included in the calculations of mental model. This will be explained in the Method and Analysis sections. So, actual extent of mental model was measured per participant, and not manipulated by using different learning materials, samples of different populations or different treatments of participants. This way possibly significant effects of differences in mental model cannot be confused with effects of other differences resulting from manipulations.

In this study, it was hypothesized that the more extensive the mental model is before studying the learning material, the more beneficial delayed feedback is compared to immediate feedback, and vice versa.

Method

To test the hypothesis that different sizes of mental model result in different optimal feedback timings, it was necessary that within the sample differences in mental models existed. Because children of various capacities and cultural backgrounds mostly attend the same elementary schools, the experiment was conducted at elementary schools. For this experiment, it was necessary that participants could read reasonably well, but also that they had had no or little education about the learning material (World War II) at school. Therefore the experiment was conducted in the fourth grade. World War II was chosen as subject because it was assumed from personal experience that children learn about it from their parents in varous amounts, according to their backgrounds. So, in this experiment, no learning material was presented before the first test. The information was known beforehand in various amounts by participants.

Design

This study was a quantitative experiment. Independent variables were feedback condition (dichotomous) and extent of mental model (ratio), and the interaction effect of both. The dependent variable was learning gain (ratio). Within each class, participants were assigned randomly to one of two conditions, labelled as immediate feedback (IF) and delayed feedback (DF). Extent of mental model was measured based on the number of concepts known beforehand about World War II, the number of correct answers in the first test and the mean response certitude in the first test. Learning gain was measured regarding both the number of questions that were answered incorrectly in the first test and correctly in the posttest, and the percentage of number of questions that were answered incorrectly in the first test that was answered correctly in the posttest. This will be explained in the analysis section.

Participants

Using a multiple regression and supposing a medium effect size of f^2 0.15, with 107 participants, power is 0.95 (G*Power 3.0.10). Participants were 121 students in Grade 4 from five Dutch elementary schools, but nine of them were absent during the posttest, so 112 participants (53 boys and 59 girls) remained (age: *M* = 9 years and 9.38 months, *SD* = 5.69 months). In the IF condition were 26 boys and 31 girls; in the DF condition were 27 boys and 28 girls. The schools were recruited by convenience sampling, from the researcher’s network. Permission from school principals and teachers was arranged. Parents got a letter with a description of the experiment. They were given the opportunity to exclude their child from the experiment, which nobody did.

Materials

A pilot was conducted in one class of nine students. Without cues these students wrote down in key words what they knew about World War II. On the basis of these recalls a test was constructed with short answer questions about World War II (for original items in Dutch, see Appendix B). The test consisted of ten questions concerning generally known information about World War II (‘easy questions’), and ten questions about generally unknown information, with surprising answers (‘difficult questions’). All test items could be learned as isolated facts, and therefore the testing effect could be expected to occur (Van Gog & Sweller, 2015). In the following days, the whole procedure as described in the next paragraph was tested in this pilot class, except for writing down prior knowledge. This testing did not lead to any changes in the procedure.

Procedure

The procedure as conducted in the main experiment is displayed in Figure 1 and described in detail below. First, the researcher explained to the students the objective of educational research in general, and thus the importance of doing the tasks fairly. Then the procedure was explained. It was emphasized that the teacher would not see the worksheets, and that the tasks were of no importance for the marks on school reports. To emphasize this, students did not have to write their names on the sheets, but only their date of birth and gender. Students were told that the next day another session would follow, but not that a posttest would take place, to limit communication between students about the subject in the meantime. Nevertheless, students were asked not to communicate about the questions with each other or with other people, until the session on the next day.

Next, students were given a sheet of paper and were asked to write down in key words what they knew about World War II. No cues were given. Subsequently, the sheets were handed in. Next, all students got a worksheet with twenty numbers on it (for the original worksheet in Dutch, see Appendix C).

Students in the IF condition got a set of twenty numbered cards, tied together with a string to keep them in the right order. Every card had one short-answer question about World War II on it. Easy and difficult questions were alternated. The student wrote an answer to the question, behind the corresponding number on the worksheet. It was compulsory to answer, blanks were not allowed, so that no omission errors could be made and the students had to elaborate on every question. The student indicated on a scale of 1 to 4 how sure he or she was of the answer: 1 = not sure at all, 2 = not so sure, 3 = pretty sure and 4 = absolutely sure. Then the card was turned. On the backside of the card the correct answer was printed (without the number), and the student checked his or her own answer. If the answer was wrong, the student copied the correct answer from the card behind the wrong answer on the worksheet, before going to the next card. There was no time limit to this test, to prevent reading and writing proficiency of participants to have influence. After finishing, students handed in the worksheets and cards.

After getting the worksheets, students in the DF condition got a set of twenty cards with the same questions on it in the same order as students in the IF condition, but without the correct answers on the backsides. The students filled in their answers to the questions, behind the corresponding numbers on the worksheet. Again, it was compulsory to fill in an answer, blanks were not allowed. These students also indicated on a scale of 1 to 4 how sure they were of their answer. There was no time limit to this test. After answering the questions, worksheets and question cards were handed in.

After the test, students did a drawing task until all students were finished. Next a regular 45 minutes lesson from the curriculum was given. After this lesson, students in the IF condition did the drawing task or a task given by their teacher, and students in the DF condition got back their worksheets, along with the question cards of the IF condition, to check their answers. If an answer was wrong, they copied the correct answer from the IF card behind the wrong answer. The IF cards were shuffled, to prevent students from simply copying the answers on the right places without paying attention to the questions anymore. There was no time limit to do this correction work. Afterwards, all worksheets and cards were handed in.

One day later, around the same time, the final test was completed by all students. This way the *lag to test*, i.e. the time span between the last learning activity and the test (Metcalfe et al., 2009) differed between both conditions by about 60 minutes. This was not considered a big problem, because the total lag to test was much bigger: about 24 hours. The same twenty questions as in the first test were on one sheet, in a different order than in the first test, to avoid an order effect. All students got a worksheet with numbers, filled in their answers to the questions behind the corresponding numbers and handed in the worksheet. If a student did not know an answer, blanks were allowed in this final test. There was no time limit to this test.

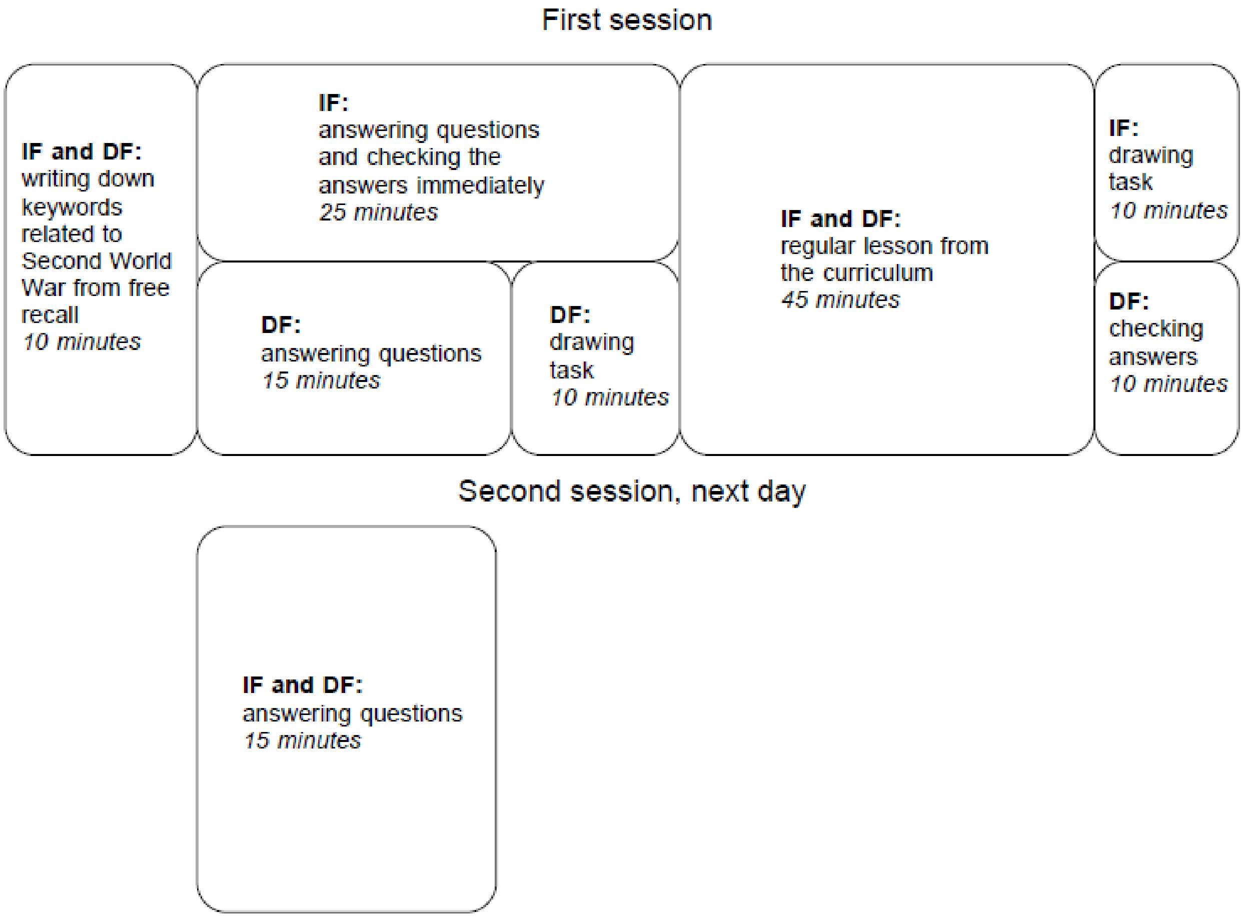


Figure 1. Procedure of the Experiment.

Analysis

Before analysing the data and addressing the central research question, the following assumptions concerning the materials used in the experiment were checked: (1) There was variance in extent of mental model. (2) There was variance in number of correct answers in the first test and in the posttest. (3) Students learned something between the first test and the posttest.

The research question concerned the interaction effect of condition (immediate versus delayed feedback) and extent of mental model on dependent variable learning gain. To answer this question, a multiple regression analysis was conducted with the independent variables condition (dichotomous) and extent of mental model (ratio), and learning gain as the dependent variable (ratio). During the experiment, gender seemed to influence the dependent variable too (in general, boys seemed to score better), and because an assumption for multiple regression is that all influential variables are included (Field, 2009), gender was included in the analysis as an independent (dichotomous) variable.

Extent of mental model was measured on the basis of (1) number of concepts known by free recall before the first test, (2) number of correct answers in the first test and (3) mean response certitude in the first test. The number of concepts known by free recall was determined as follows: a key word written down before the first test was scored as a concept on the basis of a concept list, put together by two assessors based on all results (for the original list in Dutch, see Appendix A). Also wrong concepts, for example about the dates of the war, were counted. The number of concepts was rated independently by the two assessors. Pearson Correlation between both ratings was .968, which is very high, and differences were solved by discussion. Numbers of concepts, mean response certitudes and numbers of correct responses in the first test were converted to *z*-scores. These *z*-scores were added together, and 5 was added to all scores, to make all scores positive. For each student, extent of mental model was defined as the outcome of this calculation.

It was challenging to define learning gain without bias. Initially, two ways were tried: regarding the difference between number of correct answers in the first test and number of correct answers in the posttest, and regarding the percentage of the number of questions that were answered incorrectly in the first test that was answered correctly in the posttest. Both methods of measurement had a disadvantage. According to the first method, it was easier for a participant with a low score in the first test to improve than it was for a participant with a high score in the first test. Because the maximum score was 20, a participant with a score of

13 in the first test could have a maximum learning gain of 7, while a participant with a score of 2 in the first test could have a maximum learning gain of 18. The other method favoured participants with a high score in the first test. All participants could have a maximum learning gain of 100 per cent, but a participant with a high score in the first test would need fewer additional correct answers in the posttest to reach this score than a participant with a lower initial score. Indeed, analyses showed that if the difference between initial score and posttest score defined the dependent variable learning gain, the beta was negative (β = -.16, *p* = .115). This means that the more extensive the mental model was, the less learning gain was achieved. If percentage of incorrect answers in the first test that were corrected in the posttest was considered as learning gain, the beta was positive (β = .279, *p* = .004), i.e. the more extensive the mental model was, the more learning gain was achieved. To neutralize this, scores of both variables were converted to *z*-scores and added together. To make all scores positive, 5 was added to them. For each student, learning gain was defined as the outcome of this calculation.

Assumptions associated with multiple regression were checked, and a multiple regression analysis was conducted with condition, extent of mental model and gender as independent variables, as well as the interaction effect of condition and mental model, and learning gain as the dependent variable.

Additionally, a factorial between groups analysis of variance (ANOVA) was used to compare average scores in the first test of four groups of participants: boys in the IF condition, girls in the IF condition, boys in the DF condition and girls in the DF condition. Another ANOVA was conducted with the same groups to compare average scores in the posttest. These analyses were done to measure the potential impact of some shortcomings of the experiment concerning the conditions, as explained below in the Discussion section, and for explorative reasons, because during the experiment gender seemed to have influence.

Results

All assumptions about the materials used in the experiment were met. There was much variance in mental model, and all students learned during the experiment. The mean score in the first test was 5.98 correct answers; in the posttest the mean score was 13.27 correct answers. The minimum score was 1 (in the first test), and three of the 112 participants had the maximum score of 20 (in the posttest). Figure 2 shows that both score sets were distributed normally, and thus there was no strong ceiling effect. It should be noted that in the IF condition twelve answers that were correct in the first test, given with certitude 3 or 4, and were incorrect in the posttest, were deleted. Reason for this was that participants in this condition could have seen the correct answers on the backsides of the cards before writing down their own answers in the first test. Thus, these answers were assumed to be given unjustly.

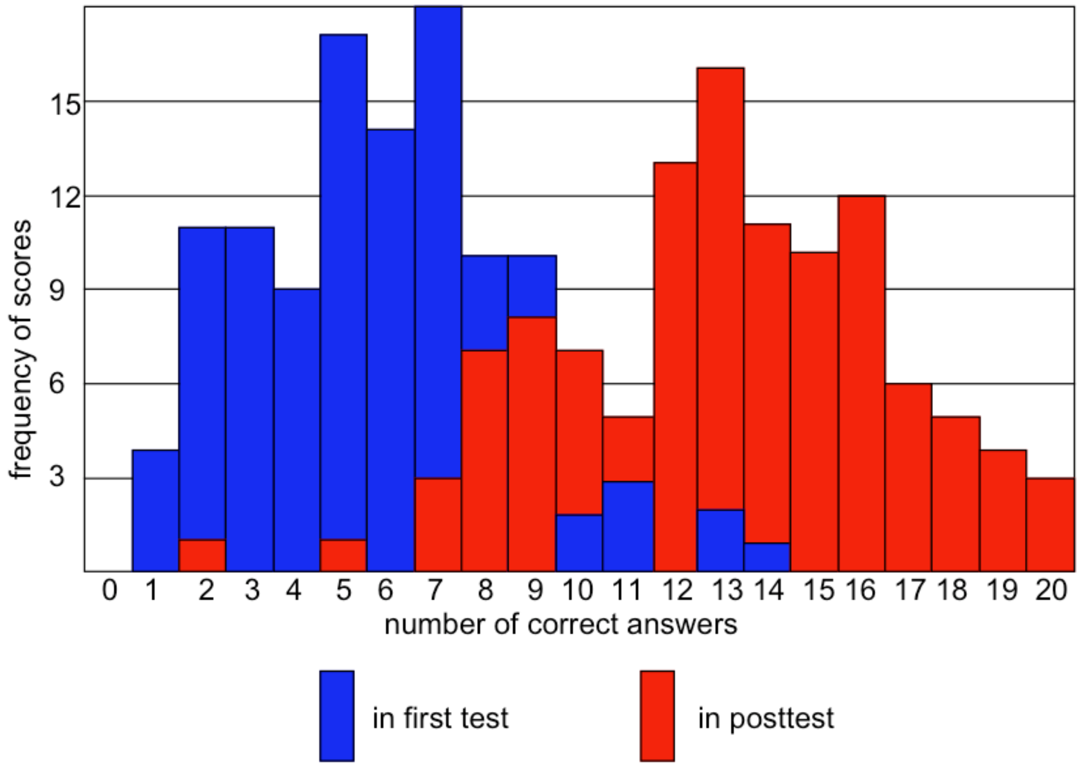


Figure 2. Score Frequencies in First Test and Posttest.

To explain the proportion of variance in learning gain caused by condition, extent of mental model and gender, as well as the interaction effect of condition and mental model, a standard multiple regression was conducted. Prior to interpreting the results, several assumptions were evaluated. Boxplots indicated that there was one univariate outlier on the variable mental model. This score was reduced to the second highest level. Assumptions of normality, linearity and homoscedasticity were met. All tolerances were > 0.1 and all VIFs were < 10, so multicollinearity did not seem to be a problem.

Results of this multiple regression analysis are shown in Table 1. Only gender had a significant influence on the outcome: boys learned more than girls during the experiment. Most striking is the fact that the interaction effect of mental model and condition, the focus of the experiment, was not significant with a *p*-value of .774.

Table 1

Multiple Regression Results of Influences on Learning Gain (*n* = 112)

	Coefficients				
Variables	*B*	*SE B*	β	*t*	*p*
(Constant)	4.46	.59		7.60	.000
Mental Model	.05	.08	.06	.63	.530
Condition	-.43	.83	-.11	-.51	.610
Gender	1.06	.38	.28	2.80	.006
Interaction Mental Model and Condition	.02	.07	.06	.29	.774
F	2.98				
p	.022				
*R*²	.10				

Figure 3 shows the individual scores on mental model and learning gain of participants in both conditions. This illustrates the absence of a significant interaction effect between condition and extent of mental model. Indeed, the visible nonsignificant interaction effect is even conflicting with the hypothesis.

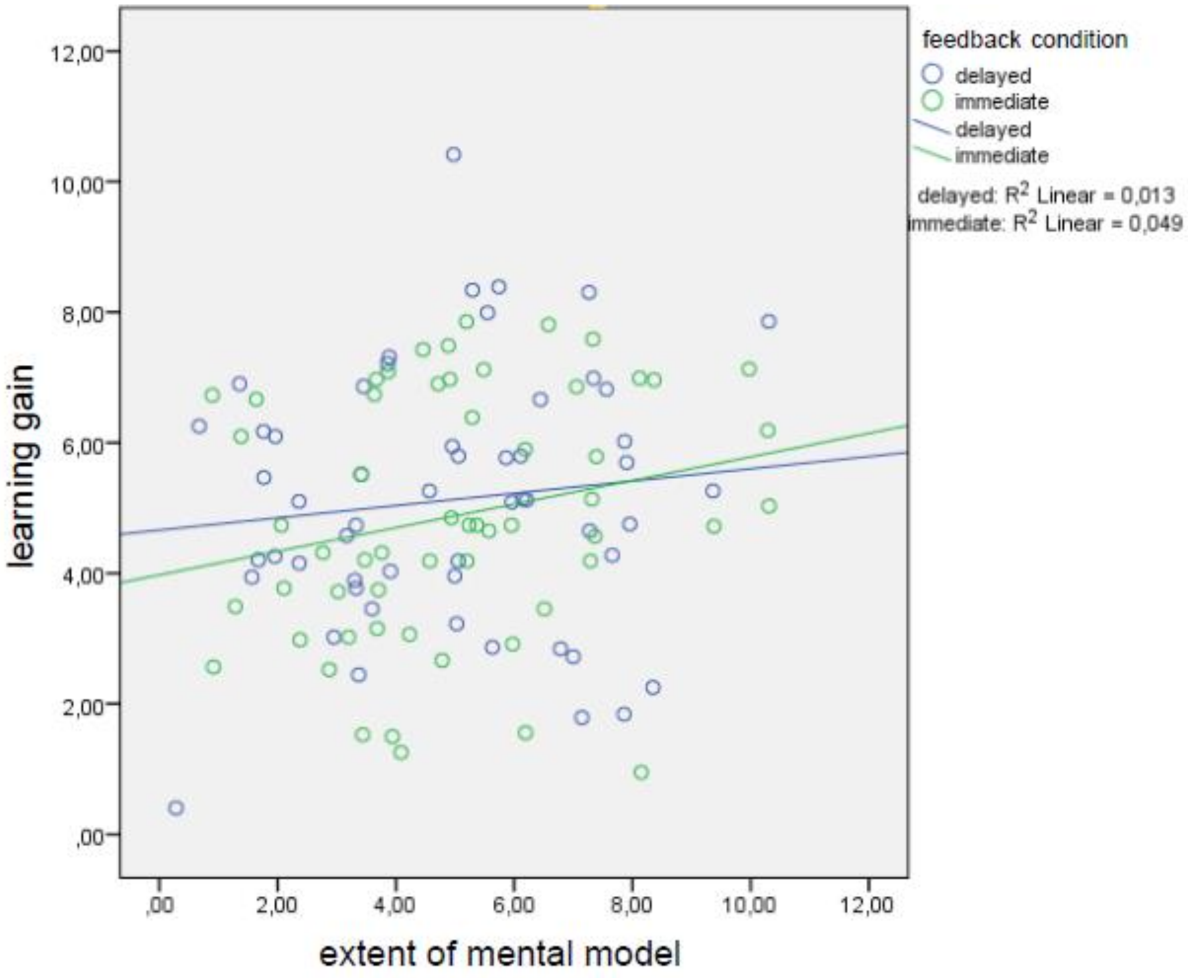


Figure 3. Interaction Effect of Condition and Extent of Mental Model on Learning Gain.

Because no experimental effect was found, participants were divided by median split into two groups of mental model: small mental model and extensive mental model. Then a factorial between groups analysis of variance (ANOVA) was used to compare average scores of learning gain of students with small mental model in the IF condition, students with extensive mental model in the IF condition, students with small mental model in the DF condition and students with extensive mental model in the DF condition. This ANOVA did not show any significant results, nor results that differed much from the multiple regression analysis outcomes. Therefore these results are not reported.

Additionally, responses were analysed, apart from the participants. The intention was to check whether this way some support for the hypothesis could be found. All incorrect answers in the first test got three dichotomous qualifications and were counted:

1. Response certitude, which was one component of the mental model, (high, response certitude 3 or 4 or low, response certitude 1 or 2)

2. Condition (immediate feedback or delayed feedback)

3. Score in the posttest (correct or incorrect), which can be considered here as learning gain, because only the incorrect answers in the first test were counted.

From the 597 incorrect answers in the first test that were given with low response certitude and followed by immediate feedback, 56 per cent was answered correctly in the posttest. From the 589 incorrect answers that were given with low response certitude and followed by delayed feedback, 54 per cent was answered correctly in the posttest. From the 260 incorrect answers that were given with high response certitude, 60 per cent was answered correctly in the posttest, in both feedback conditions. It is apparent that this analysis also does not support the hypothesis that a more extensive mental model profits more from delayed feedback, and a smaller mental model profits more from immediate feedback.

Concerning the additional ANOVAs (comparing average scores in the first test and in the posttest of boys in the IF condition, girls in the IF condition, boys in the DF condition and girls in the DF condition), assumptions of normality and homogeneity of variance were met in both analyses. The results of these ANOVAs are shown in Table 2 and Table 3.

Table 2

Mean and Standard Deviation of Scores in First Test for Condition and Gender

	Boys			Girls			Total		
Condition	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Immediate Feedback	26	7.46	2.97	31	5.34	2.66	57	6.31	2.98
Delayed Feedback	27	6.20	2.54	28	5.11	2.26	55	5.65	2.44
Total	53	6.82	2.81	59	5.23	2.46			

The main effect of gender on scores in the first test was statistically significant, *F* (1, 108) = 10.57, *p* = .002, *partial* η² = .089, with boys achieving significantly higher than girls. The main effect of condition on scores in the first test was statistically not significant, *F* (1, 108) = 2.27, *p* = .135, *partial* η² = .021, with participants in the IF condition achieving slightly higher than participants in the DF condition.

Table 3

Mean and Standard Deviation of Scores in Posttest for Condition and Gender

	Boys			Girls			Total		
Condition	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Immediate Feedback	26	15.21	3.19	31	11.73	3.23	57	13.32	3.63
Delayed Feedback	27	14.32	3.19	28	12.18	3.25	55	13.23	3.37
Total	53	14.76	3.19	59	11.94	3.22			

Likewise, the main effect of gender on scores in the posttest was statistically significant, *F* (1, 108) = 21.31, *p* = .000, *partial* η² = .165, with boys achieving significantly higher than girls. The main effect of condition on scores in the posttest was not statistically significant, *F* (1, 108) = .13, *p* = .716, *partial* η² = .001, with participants in the IF condition achieving slightly higher than participants in the DF condition. There was no interaction effect between gender and condition, nor in the first test, *F* (1, 108) = 1.08, *p* = .302, *partial* η² = .010, nor in the posttest, *F* (1, 108) = 1.23, *p* = .270, *partial* η² = .011.

Discussion

It was hypothesized that immediate corrective feedback would be more effective for participants with a smaller mental model about a subject, and that participants with a more extensive mental model would benefit more from delayed corrective feedback. This hypothesis was not supported by this study: timing of feedback did not make any difference in interaction with extent of mental model.

In former studies, variables related to the concept of mental model were manipulated by the use of different learning materials (Kornell, 2014, Metcalfe et al., 2009), different treatments (Nakata, 2014) and/or samples from different populations (Metcalfe, 2009). Thus, in these studies, other differences between learning materials, treatments or populations, not concerning the extent of mental model, could have influenced the effects of extent of mental model in relation to timing of feedback. In the present study, this problem did not exist, because samples from one population and the same learning material for all participants were used. Because the extent of the mental model was measured in every participant, and not only supposed beforehand, this variable was defined in a more precise and realistic way. Smits, Boon, Sluijsmans and Van Gog (2008) conducted a study partially concerning the relationship between optimal feedback timing and level of prior knowledge, which is only a part of the mental model. They did measure prior knowledge in participants by using a pretest, but the scores were analysed by median split, which reduces precision. In the present study, results can be considered more realistic because more aspects of mental model are involved, mental model has been measured, and has been analysed as a continuous variable. As a matter of fact, the hypercorrection effect (Butterfield & Metcalfe, 2001) seems to get some support from the response analysis, which suggests that response certitude is a relevant indicator of mental model indeed.

Possible explanation

Nevertheless, regarding the results of the present study, one can question whether the differences in mental model were defined the right way, concerning feedback timing. In his study (in which the learning material was not presented before the initial test), Kornell (2014) distinguishes between test items that have intrinsically right answers and test items that are constructed only for the experiment and do not have intrinsically right answers, such as arbitrary word pairs. If this boundary between total absence and (albeit possibly a minimal) presence of a mental model is crucial indeed, the learning material of the present study did not evoke enough differences in extent of mental model (even though these differences were huge), because all test items had intrinsically right answers and all students knew at least something about World War II, so everybody had some kind of mental model in this respect. Smits et al. (2008) suggest in their Discussion the same explanation for the lack of a significant interaction effect of feedback timing and level of prior knowledge in their experiment. Possibly, knowing that it is impossible to know the right answer to a test item evokes a different brain process while answering the question, than when one knows a correct answer exists and possibly can be found in memory, even when this does not happen.

On the basis of this presumption, one could explain the results from a cognitivist view. Completely new and unknown information comes via the senses and the sensory memory to the working memory, and can only be stored in long-term memory when processed in working memory (Woolfolk, Hughes, & Walkup, 2013). Immediate feedback after guessing contributes to processing in working memory, and probably this is more effective than delayed feedback, because in that case guessing and receiving feedback are two separate events. This is in line with for example Kornell's (2014) first experiment with arbitrary word pairs. In fact, without prestudy, this is initial learning. On the other hand, when test items can be related to a mental model, which exists in long-term memory, the timing of feedback is less important, because this mental model can be evoked any time from long term memory. Indeed, Foerde and Shohamy (2011) showed in their experiments that immediate feedback activates a different part of the brain than delayed feedback does. This could explain why in both Kornell's (2014) second experiment with trivia questions, as in the present study, timing of feedback did not have any effect on learning gain. Nevertheless, this does not explain why for example the children in the first experiment of Melcalfe et al. (2009), in which the meaning of existing words was learned, benefited more from delayed feedback than from immediate feedback, instead of equally. The spacing aspect of delayed feedback might have been the crucial factor here.

Furthermore, if immediate feedback is only more useful than delayed feedback after answering test items that have been constructed especially for an experiment, without intrinsically right answers, one can question how relevant that conclusion is for education. Generally speaking, students do not have to learn nonsensical information.

Limitations

This study has some weaknesses. One problem might be the sample size. This size was chosen expecting a medium effect size. If in fact there is only a small effect, 776 participants would have been needed (G*Power 3.0.10) to detect it. This might happen if this study is replicated with a bigger sample size. On the other hand, if the effect is only small, it is debatable whether such an outcome really should have educational implications.

Another issue is the timing of the delayed feedback. In former studies focusing on feedback timing, the lag from test to delayed feedback varied between a few seconds to several days (Smith & Kimball, 2010). The experimental effects might have been different if delayed feedback had been postponed for more than 45 minutes, thus increasing the spacing effect. On the other hand, the lag between first test and delayed feedback certainly was long enough not to occur in the same working memory activation as answering the question, because during the delay, working memory was occupied by another task. More interesting and relevant is the question to what extent the posttest would have generated different results if postponed, for example, a week. Perhaps this should be investigated if the experiment would be replicated.

Some experimental shortcomings affecting conditions in different ways need to be mentioned. During the first test, students in the IF condition could have seen some answers before answering a question, while turning more than one card at once accidentally. They could also have cribbed the answers deliberately, although they were warned that this would be noticed afterwards, and that in that case their test results would be useless and be thrown away. As noted in the Results section, correct answers in the first test, given with high response certitude, were assumed to be given unjustly if they were incorrect in the posttest, and deleted. Nevertheless, other correct answers in the first test could be given unjustly as well. Furthermore, possibly the corrective feedback after each question in the IF condition contained information or cues that made it easier to find the correct answers to the next questions. Indeed the IF group scored slightly higher in the first test than the DF group, but this difference was nonsignificant. On the other hand, lag to test was about one hour shorter for the DF group, as discussed before, which could enhance performance of the DF group in the posttest. Nevertheless, the IF group scored better in the posttest too, although only very slightly. Advantages for the IF group in the first test, combined with the advantage for the DF group in the posttest could have resulted in more learning gain for the DF group. In fact, the DF group did learn more, but also this difference with the IF group was nonsignificant. Apparently, shortcomings mentioned above did not influence the experimental outcome.

Some other factors may have influenced the outcome, although they cannot have influenced the results of the analyses that were supposed to answer the research question, because these limitations affected both conditions equally. Nevertheless, they are interesting to mention, so that in future comparable research such problems can get some attention. Almost all participants were children of native Dutch parents, and did not fully represent all students of Grade 4 in Dutch elementary schools in that sense. Writing down concepts on free recall did not fully work as expected. Twenty-four students (21 per cent) did not write anything down at all, although on average this group did give 4.4 correct answers in the first test. Obviously, they needed a cue to remember what they knew about World War II, but this cue could not be given before the test without giving away some answers to test items. Because many answers to the questions were geographical names, in the posttest participants sometimes may have guessed correct answers. Apparently there were wrong guesses too: in the posttest a geographical name that was required as an answer was often given as an answer to the wrong question. Students were asked not to communicate about the test items with each other or with other people as long as the experiment was not finished, so that learning gain would result only from the activities of the experiment. However, this could not be controlled, so some students may have violated this request. These problems should be avoided in future research, although they concerned students in both conditions equally and did not influence the experimental outcome, as said before.

The influence of gender

The only significant results concerned gender. Boys scored significantly better in the first test and in the posttest than girls, and their learning gain was significantly higher too.

Ivinson and Murphy (2003) found that boys (aged 14 and 15) more often choose war as a subject for an essay than girls do. Possibly, in general boys are more interested in the subject of war than girls are. Moreover, according to a study of Oakhill and Petrides (2007), boys benefit more from being interested in the subject of a text than girls, which means that being interested enhances the learning gain of boys more than it enhances the learning gain of girls. Oakhill and Petrides (2007) refer in their Introduction to other studies showing this phenomenon, and it was confirmed in their own experiment. Even so, this fact cannot have affected the experimental effect of the present study, because boys and girls were distributed over both conditions almost equally.

File drawer problem

Based on the present study, no interaction effect of the extent of the mental model and the timing of feedback seems to exist. Neither did a main effect of feedback timing show, although in many other studies such an effect was demonstrated (Kulik & Kulik, 1988). Because this nonexistence is so manifest in this study, one can question to what extent the publication of studies might be biased in favour of studies with significant results. Possibly, many other studies with results comparable to those of the present study exist, without being published. Opinions about this phenomenon, which has been called the *file drawer problem* (Rosenthal, 1979), are divergent. Dalton, Aguinis, Dalton, Bosco, and Pierce (2012) investigated the percentages of published and nonpublished nonexperimental studies with significant versus nonsignificant results. They deny the existence of the file drawer problem, because those percentages are almost equal. On the other hand, Franco, Malhotra and Simonovits (2014) investigated 221 social experimental studies that were conducted in the context of one scientific program, but were published only partially. They concluded that the file drawer problem does exist, not particularly because submitted studies with significant results are more likely to be published than studies without significant results, but mostly because studies without significant outcomes are much less likely to be submitted or even written at all. Maag and Losinski (2015) add to this the fact that within published articles results with less importance tend to be omitted to get the right word count. They point out the consequences for meta-analyses.

Conclusion

The present study was not able to answer the question why sometimes corrective feedback is more beneficial if given immediately, and sometimes could better be delayed. Indeed, the outcome of this study suggests that the problem itself does not exist, because timing of corrective feedback concerning real learning material did not make any difference at all. Perhaps this is good news for designers of educational devices, because they can focus on other aspects of their products instead of feedback timing.

7954 words

References

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso, *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura, *Metacognition: Knowing about Knowing* (pp. 185-205). Cambridge, MA: MITT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes, 2*, 35-67.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604-616. doi: 10.3758/MC.36.3.604

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*, 273-281. doi:10.1037/1076-898X.13.4.273

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6*, 1491-1494. doi: 1O.1037//0278-7393.27.6.1491

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology, 65*, 221-249.

Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological Science*, *21*, 801-803. doi:10.1177/0956797610371341

Field, A. (2009). *Discovering statistics using SPSS* (3 ed.). Londen: SAGE Publications Ltd.

Foerde, K., & Shohamy, D. (2011). Feedback timing modulates brain systems for learning in humans. *The Journal of Neuroscience, 31*, 3157–13167. doi:10.1523/JNEUROSCI.2701-11.2011

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-1505. doi:10.1126/science.1255484

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., 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

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 771*, 81-112. doi:10.3102/003465430298487

Ivinson, G., & Murphy, P. (2003). Boys don't write romance: The construcion of knowledge and social gender identities in English classrooms. *Pedagogy, Culture & Society, 11,* 89-111. doi:10.1080/14681360300200162

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528-558. doi:10.1080/09541440601056620

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 106-114. doi:10.1037/a0033699

Kostolányová, K., & Šarmanová, J. (2014). Use of adaptive study material in education in e-learning environment. *Electronic Journal of e-Learning, 12,* 172-182.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 68*, 505-512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279-308.

Kulik, J. A., & Kulik, C. L. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58,* 79-97. doi:10.3102/ 00346543058001079

Kumpulainen, K., Mikkola, A., & Jaatinen, A. M. (2014). The chronotopes of technology-mediated creative learning practices in an elementary school community. *Learning, Media and Technology, 39,* 53-74. doi:10.1080/17439884.2012.752383

Maag, J. W., & Losinsky, M. (2015). Thorny issues and prickly solutions: Publication bias in meta-analytic reviews in the social sciences. *Advances in Social Sciences Research Journal, 2,* 242-253. doi:10.14738/assrj.23.1044

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37,* 1077-1087. doi:10.3758/MC.37.8.1077

Nakata, N. (2014). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research, 19,* 416-434. doi:10.1177/1362168814541721

Oakhill, J. V., & Petrides, A. (2007). Sex differences in the effects of interest on boys' and girls' reading comprehension. *British Journal of Psychology, 98,* 223–235. doi:10.1348/000712606X117649

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93,* 223-231. doi:10.1002/j.2168-9830.2004.tb00809.x

Pyc, M., & 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.jml2009.01.004

Roediger, H. L., & Butler, A. C. (2010). 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

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432-1463.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 80-95.

Smits, M. H. S. B., Boon, J., Sluijsmans, D. M. A., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*, 183-193. doi:10.1080/10494820701365952

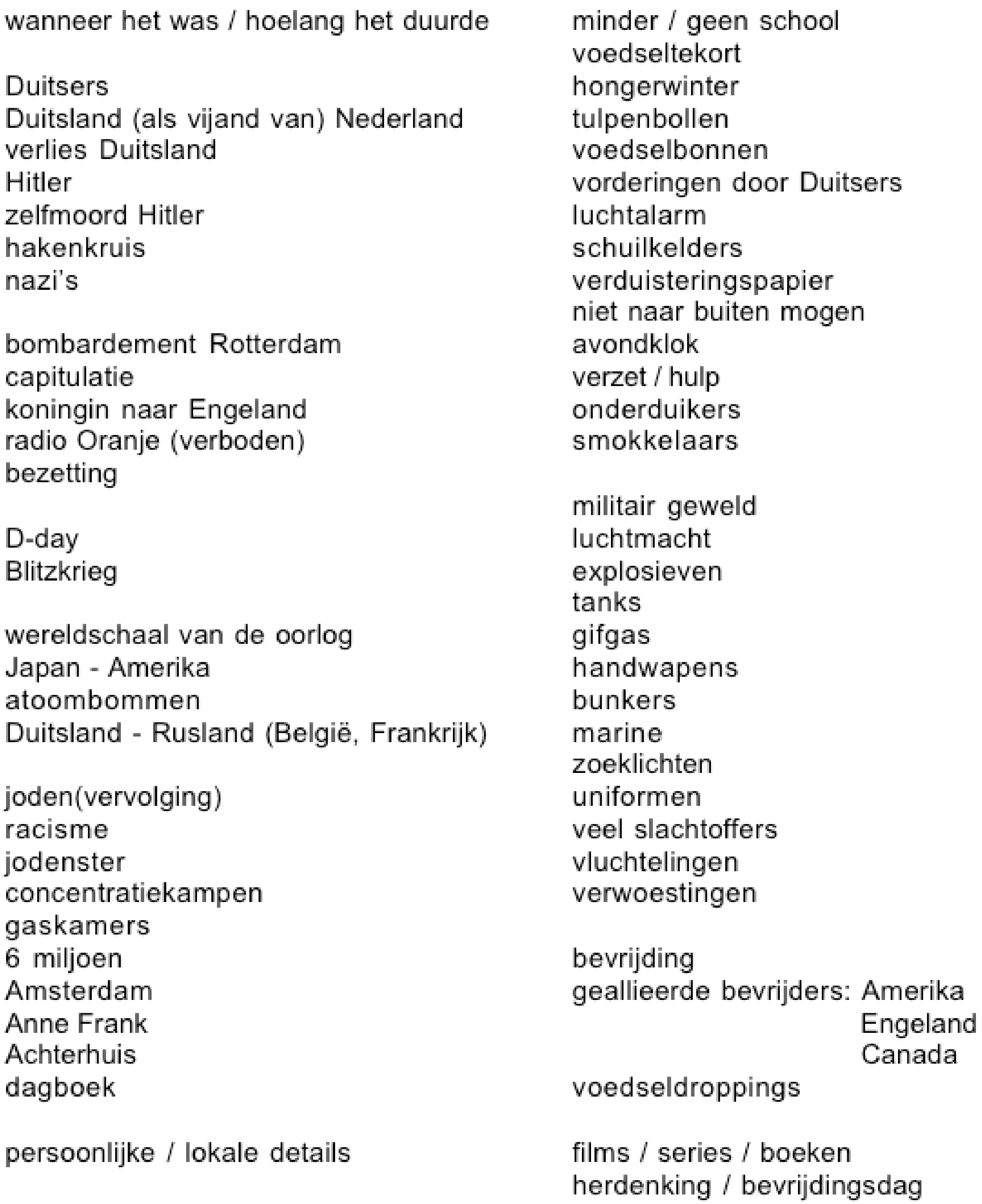
Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2015). The critical importance of retrieval - and spacing - for learning. *Psychological Science*, 1-8. doi:10.1177/0956797615617778

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychological Review, 27*, 247-264. doi:10.1007/s10648-015-9310-x

Woolfolk, A., Hughes, M., & Walkup, V. (2013). *Psychology in Education* (Second ed.). Harlow: Pearson Education Limited.

Appendix A

List of concepts



wanneer het was / hoelang het duurde

Duitsers
Duitsland (als vijand van) Nederland
verlies Duitsland
Hitler
zelfmoord Hitler
hakenkruis
nazi's

bombardement Rotterdam
capitulatie
koningin naar Engeland
radio Oranje (verboden)
bezetting

D-day
Blitzkrieg

wereldschaal van de oorlog
Japan - Amerika
atoombommen
Duitsland - Rusland (België, Frankrijk)

joden(vervolging)
racisme
jodenster
concentratiekampen
gaskamers
6 miljoen
Amsterdam
Anne Frank
Achterhuis
dagboek

persoonlijke / lokale details

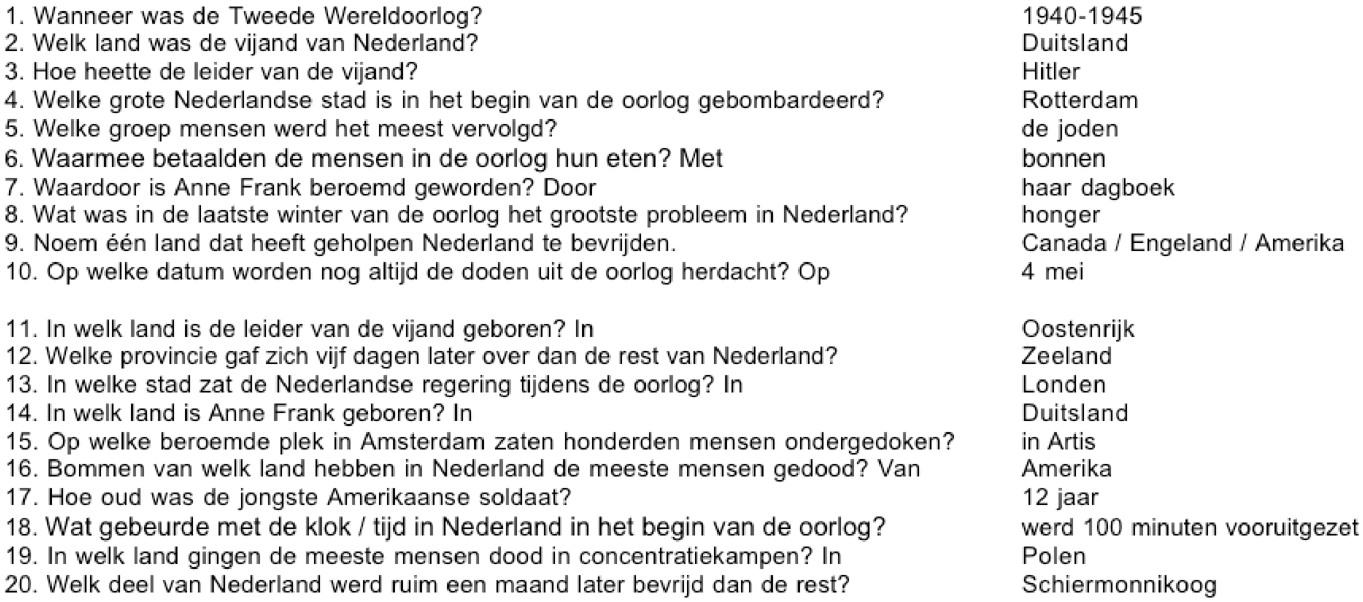
minder / geen school
voedseltekort
hongerwinter
tulpenbollen
voedselbonnen
vorderingen door Duitsers
luchtalarm
schuilkelders
verduisteringspapier
niet naar buiten mogen
avondklok
verzet / hulp
onderduikers
smokkelaars

militair geweld
luchtmacht
explosieven
tanks
gifgas
handwapens
bunkers
marine
zoeklichten
uniformen
veel slachtoffers
vluchtelingen
verwoestingen

bevrijding
geallieerde bevrijders: Amerika
Engeland
Canada
voedseldroppings

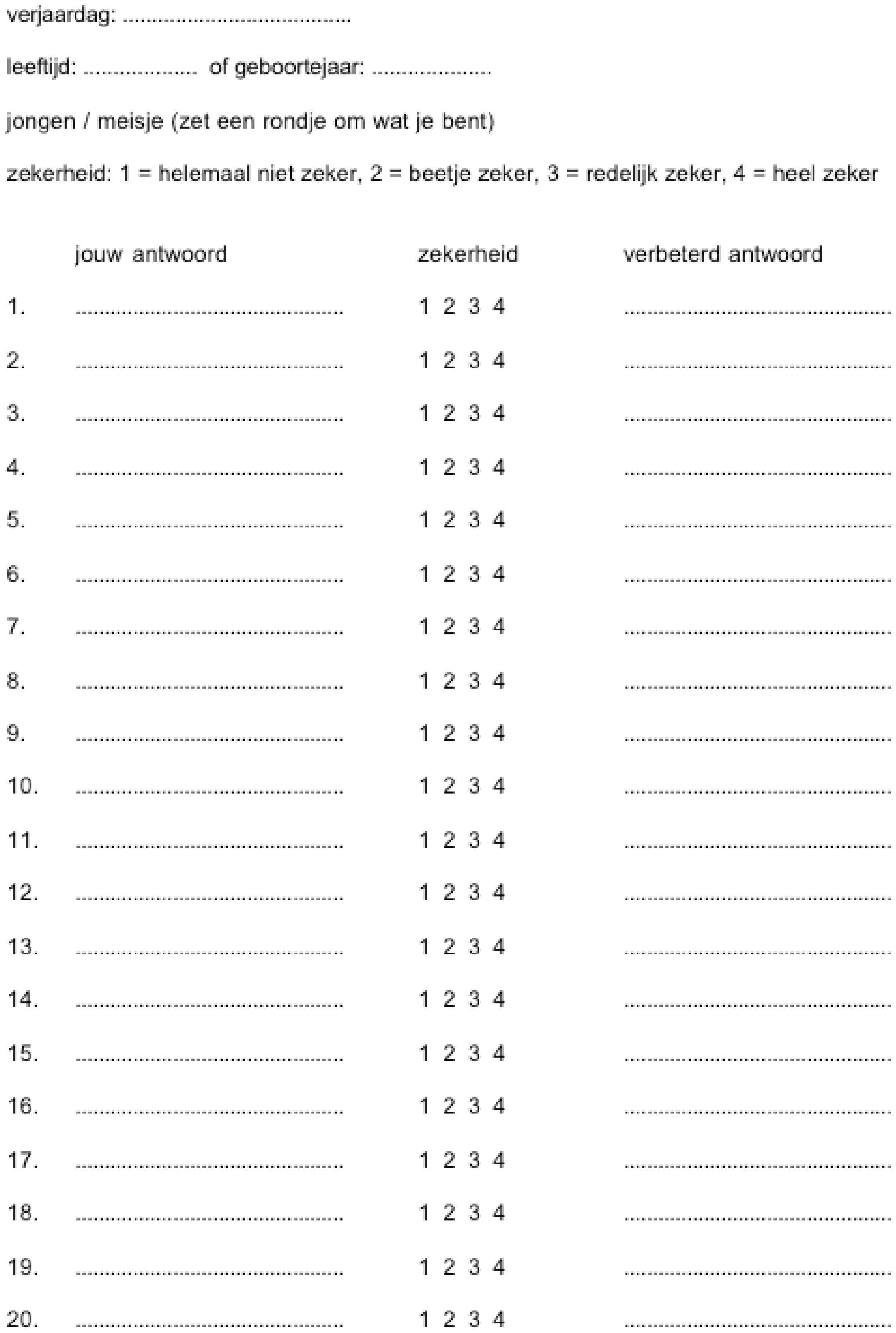
films / series / boeken
herdenking / bevrijdingsdag

Appendix B
Test items



1. Wanneer was de Tweede Wereldoorlog?	1940-1945
2. Welk land was de vijand van Nederland?	Duitsland
3. Hoe heette de leider van de vijand?	Hitler
4. Welke grote Nederlandse stad is in het begin van de oorlog gebombardeerd?	Rotterdam
5. Welke groep mensen werd het meest vervolgd?	de joden
6. Waarmee betaalden de mensen in de oorlog hun eten? Met	bonnen
7. Waardoor is Anne Frank beroemd geworden? Door	haar dagboek
8. Wat was in de laatste winter van de oorlog het grootste probleem in Nederland?	honger
9. Noem één land dat heeft geholpen Nederland te bevrijden.	Canada / Engeland / Amerika
10. Op welke datum worden nog altijd de doden uit de oorlog herdacht? Op	4 mei
11. In welk land is de leider van de vijand geboren? In	Oostenrijk
12. Welke provincie gaf zich vijf dagen later over dan de rest van Nederland?	Zeeland
13. In welke stad zat de Nederlandse regering tijdens de oorlog? In	Londen
14. In welk land is Anne Frank geboren? In	Duitsland
15. Op welke beroemde plek in Amsterdam zaten honderden mensen ondergedoken?	in Artis
16. Bommen van welk land hebben in Nederland de meeste mensen gedood? Van	Amerika
17. Hoe oud was de jongste Amerikaanse soldaat?	12 jaar
18. Wat gebeurde met de klok / tijd in Nederland in het begin van de oorlog?	werd 100 minuten vooruitgezet
19. In welk land gingen de meeste mensen dood in concentratiekampen? In	Polen
20. Welk deel van Nederland werd ruim een maand later bevrijd dan de rest?	Schiermonnikoog

Appendix C
Worksheet



verjaardag: ....

leeftijd: .... of geboortejaar: ....

jongen / meisje (zet een rondje om wat je bent)

zekerheid: 1 = helemaal niet zeker, 2 = beetje zeker, 3 = redelijk zeker, 4 = heel zeker

	jouw antwoord	zekerheid	verbeterd antwoord
1.	....	1 2 3 4	....
2.	....	1 2 3 4	....
3.	....	1 2 3 4	....
4.	....	1 2 3 4	....
5.	....	1 2 3 4	....
6.	....	1 2 3 4	....
7.	....	1 2 3 4	....
8.	....	1 2 3 4	....
9.	....	1 2 3 4	....
10.	....	1 2 3 4	....
11.	....	1 2 3 4	....
12.	....	1 2 3 4	....
13.	....	1 2 3 4	....
14.	....	1 2 3 4	....
15.	....	1 2 3 4	....
16.	....	1 2 3 4	....
17.	....	1 2 3 4	....
18.	....	1 2 3 4	....
19.	....	1 2 3 4	....
20.	....	1 2 3 4	....