The Effect of Points in Retrieval Practice on Performance and Intrinsic Motivation

Bente van Thuijl (5701112)

Utrecht University

First assessor: Gesa van den Broek

Second assessor: Pieter Wouters

Date: 4th of July 2020

Abstract

Retrieval practice – the recall or *retrieval* of information from memory – is a powerful learning strategy, but students often choose less effective learning strategies like rereading over retrieval practice. Adding points to retrieval practice, as part of so-called *gamification*, may lead to better performance and increased motivation because points might function as a form of performance feedback and can make a task more interesting. The present study investigated the effect of points in retrieval practice on performance, retention, and intrinsic motivation. Additionally, it was tested if the effect of points on intrinsic motivation was different for people with a low and high performance during practice. A within-subject design was used in which 90 participants (aged 16 to 70) practiced French words in an online retrieval practice program both with and without points. An intrinsic motivation questionnaire was used to measure motivation for the different retrieval conditions. Retention was measured with a posttest. Results showed that points did not lead to better performance during practice nor better posttest retention. However, higher intrinsic motivation was reported in the points condition independent of low or high performance of participants. The findings add new insights in the implementation of points as a single game-element.

Keywords: Retrieval practice, points, gamification, intrinsic motivation, performance, retention

The Effect of Points in Retrieval Practice on Performance and Intrinsic Motivation

Retrieval practice - the recall or *retrieval* of information from memory - is a powerful learning strategy to promote long-term retention (Karpicke & Roediger, 2008; Storm, Bjork, & Storm, 2010). Such retrieval of information from memory produces better retention over time than restudying for the same amount of time, an effect known as the *testing effect* (Roediger & Butler, 2011) However, students do not frequently use retrieval practice on their own (Grimaldi & Karpicke, 2014). They often choose to learn by rereading instead of retrieval practice (Karpicke, Butler, & Roediger, 2009; Tullis, Finley, & Benjamin, 2013), which results in suboptimal learning. Moreover, when students use retrieval practice, they find it difficult to make effective study choices and overestimate their answer correctness (Dunlosky & Rawson, 2012). This could result in too few or short practice sessions, while multiple retrieval practice sessions lead to ideal retention (Karpicke & Roediger, 2007). The question therefore arises as to how students can be motivated to engage more in retrieval practice. The present study tests whether introducing gamification in retrieval practice can lead to higher motivation and better learning outcomes.

Gamification is a process that uses elements of game design that motivate the player and implements these elements in a non-game context (Chapman & Rich, 2018; Deterding, Dixon, Khaled, & Nacke, 2011). Gamification is used to increase students' motivation and performance. According to a recent literature review (Hamari, Koivisto, & Sarsa, 2014), most gamification studies in education reported positive effects on learning and enhanced motivation. A systematic mapping study (Dicheva, Dichev, Agre, & Angelova, 2015) concluded, however, that while there are many publications on educational gamification, there is a lack of empirical research on the effect of the incorporation of specific game elements in learning environments. In particular, the motivating effects of specific single game elements in educational contexts need more substantial empirical research (Dicheva et

al., 2015; Sailer, Hense, Mandl, & Klevers, 2014). Such research could help instructors decide what game elements to use in specific contexts (Nacke & Deterding, 2017).

Points in education

The present study focuses on the game element 'points', because this is an often used game element in education (Dicheva et al., 2015). Points can be defined as numerical units indicating progress (Seaborn & Fels, 2015). Points as a gamification feature, can provide the student with rewards during the learning process. Rewards are incentive systems, in education intended to motivate or reinforce learning (Deci, Koestner, & Ryan, 2001). Reward systems are often used in games, and it is seen as an important factor to keep players motivated for a game (Wang & Sun, 2011). Including rewards in educational settings might have two positive consequences.

First, rewards can provide performance feedback (Mckernan et al., 2015). For instance, score-keeping gives the opportunity to track progress towards a desired goal (Garris, Ahlers, & Driskell, 2002; Werbach & Hunter, 2012). In addition, Kelle, Klemke and Specht (2013) suggested that points allow players to quantify their skill level and therefore help the player to understand what they can improve.Whether points are perceived as performance feedback is unclear (Attali & Arieli-Attali, 2015). According to Butler (1987) the locus of interest of students influences the effect of performance feedback. Points, as a form of feedback, may be perceived as information about task performance and perhaps direct attention to the task, or as a reward for good performance and direct attention to the self (Attali & Arieli-Attali, 2015). Normative grading in schools is an example of feedback that can direct attention on the self. Instead of focussing on the mastery of a task, students might focus on outcome and social comparison. When attention is directed to the self this might lead to reversed effects of feedback because it interferes with task performance (Butler, 1987).

Secondly, literature suggests that rewards may enhance the learning outcome by motivating students to pursue a challenging task that they otherwise would find less interesting (Cameron, Pierce, Banko, & Gear, 2005; Mckernan et al., 2015). The reward may motivate the player to continue the task and be more precise. However, there is some controversy about usage of rewards in educational settings as a way to increase performance and motivation. On the one hand, it is found that rewards can promote intrinsic motivation (Cameron et al., 2005). On the other hand, it is argued that rewards are only effective in getting people to do a task, but once the reward is not available anymore, the intrinsic motivation is undermined (Deci et al., 2001). Currently, it is recognized that the effectiveness of rewards is dependent on the type of rewards and the context they are used in (Cameron et al., 2005; Ryan & Deci, 2020). The cognitive evaluation theory suggests that how rewards influence intrinsic motivation depends on whether the reward is experienced as supporting the need for competence and autonomy (Deci & Ryan, 2002).

Previous studies on points as single gamification feature

Whether points implement as a single gamification element in educational settings lead to higher intrinsic motivation is unclear. Studies, concerning the implementation of points as a single gamification element have found inconsistent results regarding the effect on intrinsic motivation (see Table 1). On the one hand, points did not affect intrinsic motivation (Mekler, Brühlmann, Tuch, & Opwiss, 2017) but in earlier study a significant effect of points on intrinsic motivation was detected (Mekler, Tuch, Brühlmann, & Opwiss, 2013).

Table 1

Summary of Studies where Points are used as Single Gamification Element

Study	Overview of experiment	Relevant conditions	Results
Arieli-Attali, 2015	Examined the effect of different types of points shown during adaptive testing on test-taker behaviour.	• Control condition. • Number Right score, total number of correctly answered items. • Percent Correct score, percentage of	Points had an effect on all aspects of the test-takers behaviour. Reward Points seemed to work best for the engagement and performance of the test-taker. It is suggested that assigning different weights to item difficulty and a monotonically

		correctly answered items. • Reward Points, weighted score for each correct response number of points corresponding to the difficulty level of the item. • Ability Estimate score, score obtained on a predetermined arbitrary scale.	increasing score are properties of points that have a positive effect for the test-taker.
Attali & Arieli-Attali, 2015	Examined the effect of points as a gamification element in mathematics assessment on performance during the task for adults (study 1) and teenage students (study 2). In study 2 the likeability and perceived effort for the two conditions were also measured.	• Control condition. • 10 + 5 points condition, 10 points were awarded for each correct answer and 0-5 points for the speed of the response. • 1-10 points condition, 1-10 points for accuracy depending on the speed of response (only in study 1).	In study 1, a small effect (d = .28) on speed of response was found for the points conditions but no effect on accuracy of the performance. In study 2, the same effect was found for the two aspects of performance. Teenage students had a higher likeability rating for the points condition and their perceived effort was higher for the points condition.
Kelle, Klemke, & Specht, 2013	Examined the effect of points on knowledge gain and user experience in basic life support training.	• Control condition • Points condition • Points condition with time limit	The combination of points condition with time limit had a positive effect on knowledge gain and user experience. For the points condition without time limit no effects were found.
Mekler et al. 2013	Examined the effect of points on intrinsic motivation and practice performance in an online image annotation task.	• Control condition • Points condition, 100 points for each tag.	Points had a significant effect on the quantity but not on the quality of the performance on the task compared to the control condition. Points led to significantly higher intrinsic motivation compared to the control condition.
Mekler et al. 2017	Examined how points influenced intrinsic motivation, perceived competence and practice performance in an online annotation task.	• Control condition • Points condition, 100 points for each tag.	Points did not affect intrinsic motivation and perceived competence compared to the control condition. Points had a significant effect on the quantity but not on the quality of the performance on the task compared to the control condition.

Several studies (Attali & Arieli-Attali, 2015; Mekler et al., 2013; Mekler et al., 2017) reported that points led to an increase in the quantity of task performance, yet not to better quality or accuracy of task performance. Kelle, Klemke and Specht (2013), did find a positive effect on knowledge gain in the points condition when it was combined with a timer.

Points as a form of feedback during learning are a novel feature (Arieli-Attali, 2015), therefore not much research is done in different types of points (Attali & Arieli-Attali, 2015).

In the dissertation of Arieli-Attali (2015) the effect of different types of points on test-taker behaviour were examined in self-adapted testing. Weighted scores seemed to work the best to increase the engagement and performance of the test-taker. It was suggested that assigning different weights to test items depending on item difficulty and a monotonically increasing score, where incorrect answers do not decrease the total score, are properties of points that have a positive effect for the test-taker. However, results of a recent study (Nebel, Schneider, Beege, & Rey, 2017) suggests that penalties might enhance learning. Students played a educational video game with or without penalties. Students with penalties learned and recalled information more efficiently than students who practiced without penalties (Nebel et al., 2017).

There is a lack of empirical research on the effect of the incorporation of specific game elements in learning environments (Dicheva et al., 2015). A small number of studies investigated the effect of points in a non-game contexts. Points have led to higher task quantity but not to higher task quality (Mekler et al., 2017; Mekler et al., 2013; Attali & Arieli-Attali, 2015), however it is not yet tested whether higher task quantity can also lead to improved retention. Nebel et al. (2017) found that penalties in a educational game have let to better retention. Still it is not proved whether points, that can also decrease, can be effective for retention, especially when only one game-element is used.

The present study

The present study is different from previous studies where points were used as single gamification elements, because none of these studies focused on the effect of points on retention. This is particularly interesting for retrieval practice since it promotes long-term retention (Roediger & Butler, 2011). Efficient retention is important in retrieval practice therefore, penalties, in the form of a small point decrease, were added in the points system.

This paper contributes to the understanding of how retrieval practice can be effectively implemented to motivate and increase the performance of students, using points.

Rewards, such as points, can function as performance feedback (Garris, Ahlers, & Driskell, 2002; Werbach & Hunter, 2012) and may motivate people to pursue a challenging task (Cameron et al., 2005) and therefore might increase the intrinsic motivation and performance during retrieval practice. The first research question was: Do points influence performance during retrieval practice? It is not yet tested whether effects during practice will also translate into higher learning outcomes on a delayed posttest. The second research question of this study was thus: Do points lead to a better retention of the words *after* retrieval practice? The third research question was: Do points influence intrinsic motivation for retrieval practice? A form of points that has been present in education for a long time is grading. There is little evidence that grading in education enhances motivation or learning performance, however there is evidence of negative effects on intrinsic motivation and learning outcomes (Ryan & Deci, 2020), especially for lower ability students (Klapp, 2015). Therefore, it might be expected that points have a different effect on people with low performance. Therefore the fourth research question was: Is the effect of points on intrinsic motivation different for people with low and high performance in retrieval practice? To test these questions, people practiced retrieval in a condition with and without points, followed by a delayed retention test two to three days after practice.

The hypotheses are as follows: First, points in retrieval practice will lead to better performance during retrieval practice. Previous studies observed higher task quantity when points were used as a single gamification element (Mekler et al., 2017; Mekler et al., 2013; Attali & Arieli-Attali, 2015). Second, points in retrieval practice will lead to better retention on a delayed posttest. This has not been tested yet, but higher knowledge gains in an educational training with points (Kelle, Klemke, & Specht, 2011) and better retention after

playing an educational game with penalties (Nebel et al., 2017) demonstrate that points as an added gam-element potentially might increases retention. Third, points in retrieval practice will lead to higher intrinsic motivation. Mekler et al. (2013) found significantly higher intrinsic motivation when points were added as a single gamification feature in an image annotation task. The fourth hypothesis is, motivational effects of points on intrinsic motivation are expected to be larger for high performing students than for low performing students. This is expected since grades have negative effects on the intrinsic motivation and learning outcome of low performing students (Klapp, 2015). Negative effects of grades have been explained with the emphasis that is placed on evaluating what a student does with grades. This supposedly leads to performance goals and the desire to be better than others (Pulfrey, Butera, & Buchs, 2011). These negative consequences, demonstrate the importance to examine the effect points as an added game-element in educational settings. Especially, because gamification is more and more used in education (Dicheva et al., 2015), but the effect of single game elements (e.g. points), as said before, is still unclear.

Method

Participants

A total of 90 participants (M_{age} = 28.08, SD_{age} = 11.77, age range: 16-70 years, 72.2% female) were recruited by convenience sampling via multiple social media platforms such as Instagram, Facebook and WhatsApp. The participants had studied French between 0 and 6 years (M = 4.14, SD = 1.71). Of the 90 participants, see flowchart (Appendix H), six did not participate in the second session, resulting in six missing values for the variable Posttest Retention.

A gift voucher worth 25 euros was raffled per 50 participants who completed the second session. The necessary sample size was calculated a priori by using the G*Power 3.1.9.4 (Faul, Erdfelder, Lang, & Buchner, 2007). For a one-way repeated measures

multivariate analysis of variance (MANOVA) with a medium effect size f of 0.25 (Cohen, 1988) a minimum of 56 participants was required for a power level of 0.80. For the paired samples *t*-test with a medium effect size of 0.50 (Cohen, 1988), a minimum of 34 participants was required to get a power of 0.80. The sample size is thus sufficient.

Stimuli

The stimuli in this study were 78 French words with Dutch translations. The words were selected from the vocabulary lists of a French teaching method for Grade 10 pre-university level. The French words were split up into two lists of 39 words.

Design and experimental control

The randomized experiment had Scoring as within-subject factor (No Points versus Points), and three dependent variables, specifically Practice Performance (average of the correct responses and the average reaction time in the first session), Posttest Retention (percentage of correctly translated practiced words on the posttest in the second session) and Intrinsic Motivation (three measures: Interest/Enjoyment, Perceived Competence and Value/Usefulness). The Scoring condition was manipulated through randomization: participants practiced each condition for twelve minutes (48 participants first practiced with the Points conditions then with the No Points condition, 42 participants vice versa). The order of the two word list was also randomized across participants. The participants were unaware of the condition assignments.

Retrieval practice. In the current study, a digital retrieval practice program 'SlimStampen' was used (Van Rijn, 2010). The program continuously updates the estimate of the forgetting rate for each item based on the students' response time and accuracy (see Sense, Behrens, Meijer, & Rijn, 2016). The word order in which the program presents/repeats the words for retrieval practice is adapted based on the student's response time, accuracy and the number of previous retrievals. Consequently, the amount of words each participant

practiced differed based on their performance during practice. When a French word appeared for the first time, the Dutch translation was given and the participants were asked to retype this translation (i.e., study trial). After the initial study trial, the retrieval trials started for this specific word. In the retrieval trials, a French word was given without translation and participants needed to retrieve and type in the Dutch translation from memory. Participants had to press 'Enter' to receive feedback: When students had entered the correct translation, the input box turned green for 1 second. When the students had entered the wrong answer, the input box turned red and the correct translation was given for 5 seconds. When there was a small spelling error, the answer was registered as correct, the input box turned green and a warning ('Let op de spelling') with the correct answer was given for 5 seconds. The program ignored errors in the use of accents and capitalization.

Points condition. For every correct answer the score increased with 10 points, for every incorrect answer the score decreased with 1 point. The points were displayed above the timer in the top right corner (Figure 1).

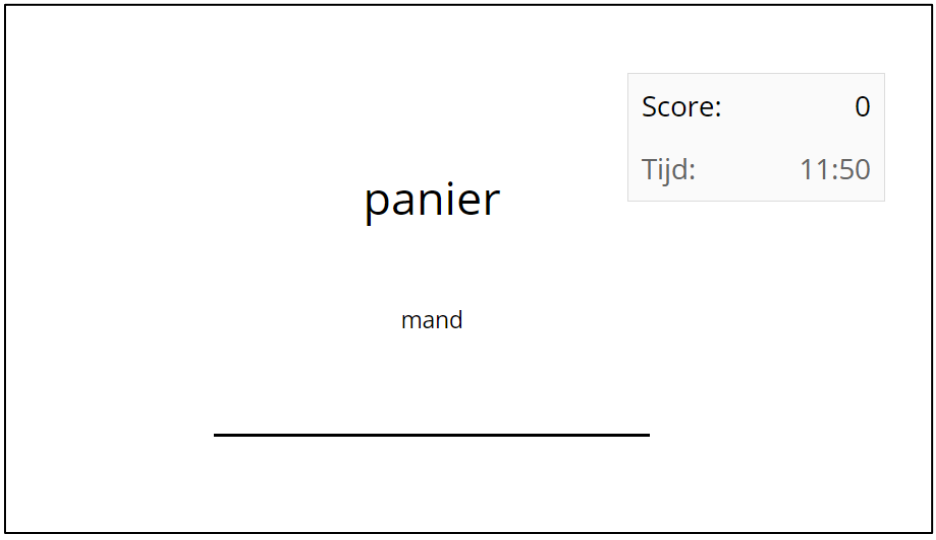


Figure 1. Display of the points condition.

Pilots and Manipulation checks. The experiment was piloted by five participants, to see whether it was user-friendly. Based on the feedback of the participants multiple adaptations were made. First, the explanation about what was expected of the participant was

simplified by adding a graphical representation. Second, the possibility to return to a previous questionnaire page was turned off. Third, in the beginning of the second session a sentence was added about the possibility that participants could encounter words that they did not practice in the first session. A first manipulation check was done by adding a question to the questionnaire to check whether the participants were distracted during the twelve minutes of word learning, namely 'Heb je tijdens het oefenen een van deze dingen gedaan?'.

Intrinsic motivation. Intrinsic motivation was measured with three subscales of the Intrinsic Motivation Inventory (IMI) (n.d) (Appendix F). Previous studies support the validity of the IMI (McAuley, Duncan, & Tammen, 1989). For the present study, the items of the subscales were rated on a 7-point Likert scale, ranging from 1 '*Helemaal niet waar'* [Not at all true] to 7 '*Helemaal waar*' [Very true]. The English items of the IMI were translated to Dutch and is a shortened version of the original survey.

The subscale Interest/enjoyment (five items) is considered as a self-reported measure of intrinsic motivation ("Intrinsic Motivation Inventory (IMI)",n.d). An item example of the Interest/enjoyment construct is, '*Ik had veel plezier tijdens deze oefening*' [Engl. I enjoyed doing this activity very much].

The subscale Perceived Competence (four items) measures how effective participants feel when they do a task (Monteiro, Mata, & Peixoto, 2015). An item example of the Perceived Competence construct is, '*Ik ben tevreden over mijn prestaties tijdens deze oefening*'[Engl. I am satisfied with my performance at this task].

The subscale Value/usefulness (two items) measures whether people experience an activity as valuable or useful for themselves. If they do so, they are more likely to become more self-regulating concerning the activity ("Intrinsic Motivation Inventory (IMI)",n.d). An item example of the Value/usefulness construct is, '*Ik denk dat deze oefening nuttig is om te leren voor een toets*' [Engl. I think this exercise is useful to learn for a test].

The reliability of the three subscales of the IMI indicated by Cronbach's alpha ranges from .79 to .93, which is satisfactory (Kline, 1999). For the subscale Interest/Enjoyment, Cronbach's α =.89 in the No Points and Cronbach's α =.91 in the Points condition. For the subscale Perceived Competence, Cronbach's α =.93 in the No Points and Cronbach's α =.93 in the Points condition. For the subscale Value/Usefulness, Cronbach's α =.88 in the No Points and Cronbach's α =.79 in the Points condition.

Practice performance. During the first session the performance during practice was measured as the number of correct responses and the average reaction time. A response was assessed as correct when the Levenshtein distance (Levenshtein, 1966) between the response and its corresponding stimuli was either 0 or 1.

Posttest Retention. A recall test was used to measure the retention of words of participants. Participants took the final recall test two or three days after the practice session. Due to a technical error, 9 words were not questioned in the test that may have been practiced in the first session. 69 items were presented in a randomized order, giving the written French word, to which participants had to provide the corresponding Dutch translation (as during practice). The Posttest Retention score was calculated as the percentage of correct responses in the second session that have been practiced in the first session, with a maximum score of 1.00. A response was assessed as correct when there was no difference between the answer and its corresponding stimuli. Unpracticed words were not included in the calculation of Posttest Retention.

Procedure

The experiment was performed online with a survey tool Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018) and consisted of two parts. Participants clicked on the link to go to the first session of the experiment. During the first session, a short instruction was given about what was expected of the participant and participants had to

accept the informed consent (Appendix A). The participants worked on the retrieval practice task in the two conditions, each for twelve minutes on a computer. After each practice block, participants filled in the intrinsic motivation questionnaire (see Appendix B to E for questionnaires and instructions in the first session). Two days or three days (when participants had not responded to the first reminder) after the first session, the second session took place, which consisted of a posttest, participants had one minute to translate a French word (Appendix F & G).

Data Analysis

The data was analysed in IBM SPSS Statistics (Version 26). For all analyses a two-tailed test with a significance level of $\alpha = .05$ was used. Practice Performance (number of correct items and average reaction time in the first session), and Posttest Retention (percentage of correctly recalled items on the final recall test out of those items that were practiced in the first session, and the average reaction time for these items in the second session). For the first and second hypothesis paired samples *t*-tests were used, to test whether there was a significant difference between the Points and No Points condition for Practice Performance and Posttest Retention. To test the third hypothesis, Scoring was entered as a within-subject factor (Points or No Points) in a one-way repeated measures multivariate analysis of variance (MANOVA) with three measures of Intrinsic Motivation: Interest/enjoyment, Perceived Competence, Value/usefulness.The fourth hypothesis, whether the motivational effect of points on Intrinsic Motivation is larger for people with a high practice performance than for people with a low practice performance, was tested by using the one-way repeated measures MANOVA of the second hypothesis and adding 'Relative Practice Performance' as a between-subject factor. Relative Practice Performance was based on participants' practice performance (specifically the total number of correctly translated words during the first session). Participants were divided into three equal groups (low,

medium or high performance) based on their ranking of the total of correct translated words during practice.

Cohen's *d* and partial η^2 were used as measures of effect sizes for significant results. A Cohen's d is classified as small when *d* = .2, medium when *d* = .5, large when *d* = .7, the partial η^2 is classified small when partial η^2 = .01, medium when partial η^2 = .06, and large when partial η^2 = .14 (Cohen, 1988). For both paired samples *t*-tests the assumption of normality and normality of the difference scores were checked based on histograms, skewness, and kurtosis. Some variables did not meet the assumption of normality, however the paired samples *t* test has fair robustness to moderate violations of normality (Jennings, Zumbo & Joula, 2002).

For the one-way repeated measure MANOVAs, histograms and *z*-scores were checked for outliers. If outliers were extreme (i.e., $z > 3.29$; Field, 2013) and had implausible values, they were removed. Probable outliers (i.e., $z > 2.58$ but $z < 3.29$; Field, 2013) were retained if they had plausible values. The assumption of normality and multicollinearity were checked and whether there were multivariate outliers. Also, the linearity between related variables was tested by examining corresponding scatterplots. For the one-way repeated measure MANOVA with Relative Practice Performance as between-subject factor, the assumption of homogeneity of variance was checked.

Results

Manipulation checks

Exclusion of participants. Four participants indicated on the questionnaire that they were distracted during word learning (2 participants had made notes; 1 participant was distracted from the experiment for longer than 1 minute; 1 participant had made notes and was distracted from the experiment for longer than 1 minute). However, these participants were included in the experiment because their practice data was plausible.

Controlling for possible order effects and prior knowledge. A paired samples *t*-test was performed to compare the number of correctly translated words in the first retrieval practice block (M = 126.54, SD = 31.99) and the second retrieval practice block (M = 133.09, SD = 31.64). On average, the participants answered 6.54 more words, 95% CI [2.12, 10.97], correctly in block 2 with a small effect size, $t(89) = 2.94$, $p = .004$, $d = .20$. Complete counterbalancing was used, however, it is assumed that this order effect did not confound the condition manipulation and results are aggregated across blocks (cf. Field, 2013, p. 18). Participants answered on average 4.5% of the unpracticed words correct on the test (as opposed to 50.0% of the practiced words); the participants had thus only limited prior knowledge of the French words.

Performance

Research Question 1. Do points in retrieval practice lead to better performance during practice? Descriptive statistics of the number of correctly translated words during the practice session in the Points and No Points condition are listed in Table 1. The difference between conditions was not statistically significant, $t(89) = 0.33$, $p = .743$. Furthermore, the mean reaction times (in milliseconds) during practice in the Points and No Points condition had no statistically significant difference, $t(89) = -0.48$, $p = .627$.

An independent samples *t*-test was performed to explore whether younger participants (25 and younger, n = 59) had a higher practice performance during practice than older participants (26 and older, n = 31). The *t* test was statistically significant with a medium effect size, $t(88) = 2.33$, $p = .022$, $d = 0.52$. Younger participants translated more words correctly during practice (M = 135.17, SD = 31.04) in the points condition than older participants (M = 118,52, SD = 34.56). However, the same test for the control condition was not significant.

Research Question 2. Do points in retrieval practice lead to better retention on a delayed posttest? Participants' retention on the posttest was also not significantly higher for the Points condition $t(83) = .451$, $p = .654$. In addition, the difference of the mean reaction times (in milliseconds) for Posttest Retention in the Points and No Points was not statistically significant, $t(83) = 1.215$, $p = .228$. See Table 1 for the descriptive statistics.

Table 1

Descriptive Statistics for Practice Performance (N = 90) and Learning Performance (N = 84) in the two experimental conditions

	No Points		Points	
	M	SD	M	SD
Practice Performance (number of correct responses)	129.43	33.07	130.20	30.86
Practice Performance (reaction time)	3770.59	997.65	3746.99	909.06
Posttest Retention (percentage of correct responses)	.50	.19	.50	.22
Posttest Retention (reaction time)	6570.78	2814.41	6235.19	342.47

Note. Practice Performance (reaction time) is measured in milliseconds, Learning Performance ranged from 0-1.

Intrinsic Motivation

Hypothesis 3. Points in retrieval practice lead to higher intrinsic motivation. The one-way repeated measures MANOVA was conducted to examine whether Intrinsic Motivation was higher for the Points condition than for the No Points condition. Before conducting the analyses, histogram and z-scores were checked for outliers. For all the dependent variables, a small number of potential outliers (i.e., $z > 1.96$) were detected, but

these were less than 5% of all observations, and thus in the to-be-expected range (Field, 2013). Probable outliers (i.e., $z > 2.58$ but $z < 3.29$; Field, 2013) were detected for two variables, but the cumulative percentage of the potential and probable outliers were still less than 5% of all observations. For one variable an extreme case (i.e. > $z > 3.29$; Field, 2013) was detected. The probable outliers and the extreme case were retained because the values were plausible, the data of the respective participants was not otherwise unusual. Four variables violated the assumption of normality but the repeated measures MANOVA is robust to small departures of normality (Stevens, 2009). The correlations between the dependent variables did not indicate multicollinearity (all $r < .90$).

The multivariate test indicated that there was a significant difference between the Points and No Points condition on Intrinsic Motivation, with a large effect size, see Table 2. Specifically, Interest/Enjoyment and Value/Usefulness differed between the Points and No Points condition, with respectively a large and medium effect size. Perceived Competence did differ in the two conditions with a small effect size. The means in Table 3 indicate that the Interest/Enjoyment score was higher for the Points condition than for the No Points condition ($p < .001$), and participants rated the points condition as significantly more valuable or useful for themselves than the No Points condition, $p = .024$. Also, participants rated their perceived competence in the points condition significantly higher than in the condition without points, $p = .050$.

Table 2

Results for the repeated measures MANOVA of the Influence of Points on Intrinsic Motivation (N =90)

Univariate

	Multivariate	Interest/Enjoyment	Perceived Competence	Value/Usefulness
Wilks Λ	0.84			
F	5.68	16.20	3.93	5.26
p	.001	<.001	.050	.024
partial η^2	**.16**	**.15**	**.04**	**.06**

Note. Multivariate *df* = 3, 87, Univariate *df* = 1, 89.

Table 3

Descriptive Statistics for Intrinsic Motivation (N = 90)

Measure	No Points		Points	
	M	*SD*	*M*	*SD*
Interest/Enjoyment	4.66	1.16	4.98	1.10
Perceived Competence	4.11	1.42	4.38	1.41
Value/Usefulness	5.67	1.08	5.85	0.93

Note. Interest/Enjoyment, Perceived Competence and Value/Usefulness ranged from 1-7.

After the participants had practiced word learning in the Points and No Points condition, 57.6 percent said they would prefer to 'practice with timer and points', 27.2 percent preferred to 'practice with timer' and 15 percent answered 'I have no preference'.

Participants' opinion regarding the points was also asked for with an open question 'Wat vond je van de punten tijdens het oefenen en hoe ben je met de punten omgegaan?' [Engl.*What did you think of the points during the practice and how did you deal with the points?*]. The answers to this question were thematically analysed in Nvivo (2019). A hierarchical tree map was created after open coding of the answer (see Appendix K). The themes were divided into three categories, namely positive (63 reactions), neutral (53 reactions) and negative experiences (34 reactions).

A positive aspect of points was that it was experienced as motivating (21 reactions) '*Ik vond het wel fijn het gaf me meer motivatie om de woorden te onthouden en zoveel mogelijk punten te halen*.' Some people indicated that they were focused on collecting a high score (15 reactions) and that points made it possible to set a goal during practice (9 reactions). For four people the points made them feel more competent '*De punten gaven je meer het gevoel dat je ook veel woorden goed had gedaan.*' Four people said the points would be more valuable for them if they practiced repeatedly with points so that they could compare their scores between sessions.

There were 53 neutral comments. It was often mentioned that people did not pay attention to the points. 22 participants said that they ignored the points on purpose and 3 participants said that they have not seen the points at all. Nine people did not see points as an added value when learning words. In eight reactions it was made clear that the balance between the amount of points you could earn and lose was unequal.

There were also 34 negative experiences of learning words with points. 15 people found the points confusing because the value of the points was unclear and no standard was given about what constituted a 'good' score, '*Het aantal punten geeft niet aan of je alle woorden onder controle hebt. Je hebt geen idee wat goed is en niet qua punten*'. Eleven people felt that the points distracted them from learning words, '*Irritant, stressvol. Je let er*

op terwijl je dit eigenlijk niet wil. Het gaat om woorden leren, niet om punten scoren mijn inziens. The addition of points to word learning was also experienced as "stressful" (4 reactions) and "annoying" (3 reactions).

Research question 4. Is the effect of points on intrinsic motivation more negative for students with a low practice performance than for students with a high practice performance? To examine whether practice performance moderated the effect of points on motivation, participants were split into three groups (low, medium, high). The group of low performers had an average performance of 193.83 correctly translated words during practice; medium performers 262.87; high performers 322.20. A one-way repeated measures MANOVA was conducted to examine differences in Intrinsic Motivation, with condition as within-subject factor and relative practice performance as between-subjects factor. The assumption of homogeneity was violated, but because the group sizes are equal this assumption can be ignored (Field, 2013).

The multivariate test indicated that there was a significant difference between conditions on Intrinsic Motivation with a large effect size, $F(3, 85) = 5.55$, $p = .002$; Wilk's $\Lambda = 0.836$, partial $\eta^2 = .16$. For the Interest/Enjoyment aspect of motivation, the condition effect reported in the previous section was replicated. That is, participants in de points condition reported to have higher enjoyment, $F = (1, 87) = 15.85$, $p < .001$, partial $\eta^2 = .15$ (see Figure 2). For Perceived Competence, the condition effect was not significant. For Value/Usefulness, there was an effect of condition with a small effect size. Participants reported higher value/usefulness scores in the points condition, $F(1, 87) = 5.20$, $p = .025$, partial $\eta^2 = .05$ (see Figure 3).

The multivariate test indicated that there was also a significant difference in the between-subject factor Relative Practice Performance with a medium effect size, $F(6, 170) = 3.57$, $p = .002$; Wilk's $\Lambda = 0.789$, partial $\eta^2 = .11$. There was no effect of performance for

Interest/Enjoyment. However, there was a main effect of performance for Perceived Competence with a large effect size, F (2,87) = 10.86, p < .001, partial η^2 = .20. Medium (M = 4.29) and high performers (M = 4.91) rated their perceived competence significantly higher compared to low performers (M = 3.53) (see Figure 4). For Value/Usefulness, there was no effect of performance. There was no interaction effect between Intrinsic Motivation and Relative Practice Performance.

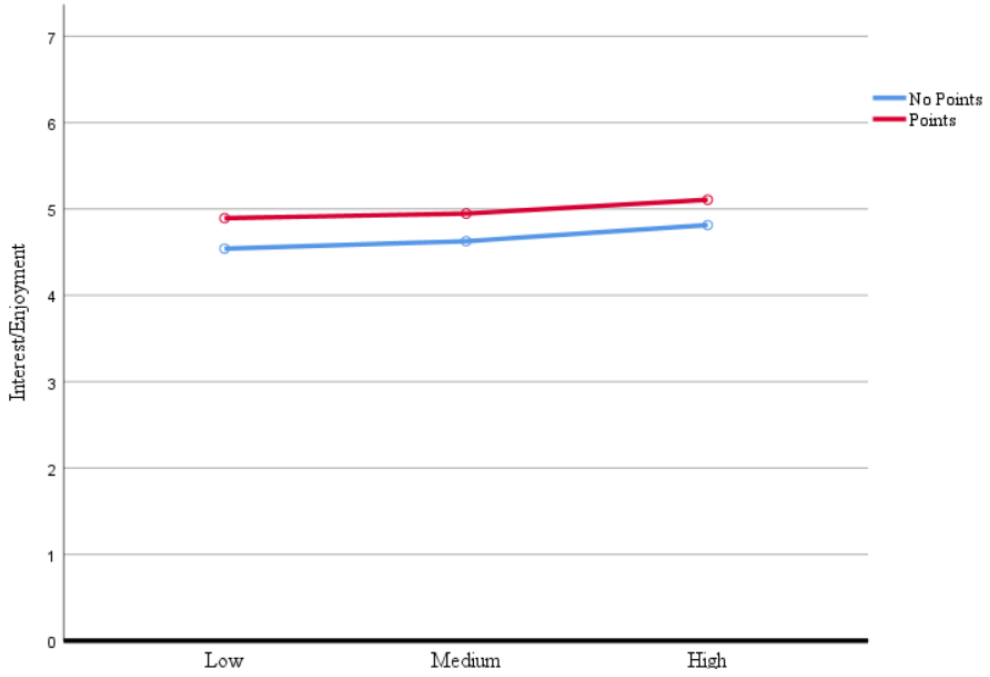


Figure 2. Interest/Enjoyment for low, medium and high Practice Performance.

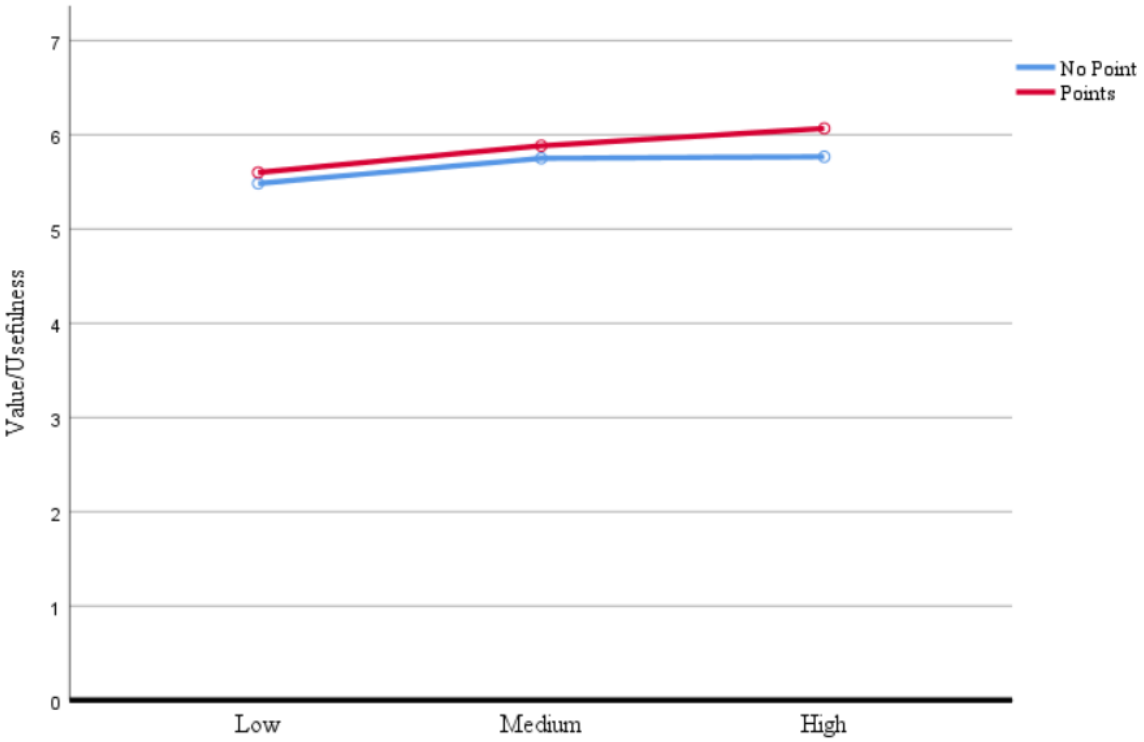


*Figure 3.*Value/Usefulness for low, medium and high Practice Performance.

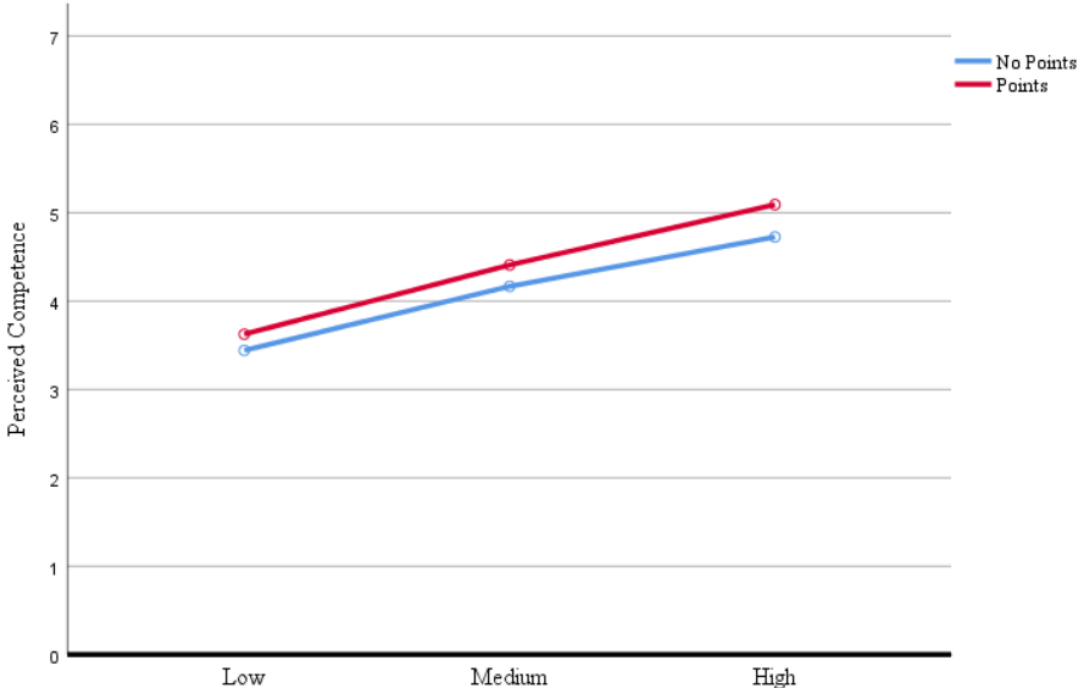


Figure 4. Perceived Competence for low, medium and high Practice Performance.

A Pearson's Chi-square test of contingencies was used to evaluate whether relative practice performance during practice was related to the preferred practice style (with timer and points, with points or no preference). The test was not significant, so participants with high performance were not more likely to prefer practice with timer and points than participants with low performance.

Discussion

The present study investigated whether performance during retrieval practice and retention on a posttest after retrieval practice improved when practice included points as a simple gamification feature. Also, it was tested whether points in a retrieval practice program led to higher intrinsic motivation and if this effect was different for people with low or high performance during practice. Based on literature, it was expected that people's performance during the task would be better in the points condition than in the control condition due to higher task quantity (Mekler et al., 2017; Mekler et al., 2013). However, contrary to this hypothesis, no difference was found in performance during practice in the two conditions.

Effects of points in retrieval practice on practice performance and retention

A possible explanation for the fact that no effect was found on performance during practice could be related to the negative experiences some participants had with points as a single gamification feature in the retrieval practice program. Participants indicated that they were distracted by the points during practice. Possibly, points did not direct attention of participants to the task, but rather let participants focus on demonstrating high ability or masking low ability relative to others. An indication of this could be that in the qualitative analyses of the present study a number of participants said to be mainly focused on getting a high score. This might have distracted participants during the task and hereby interfered with task performance, because the locus of interest can influence the effect of performance feedback (Butler, 1987).

Some participants mentioned that they experienced the points as confusing. This might be a consequence of the fact that it is not yet clear what a good points system should consist of (Arieli-Attali, 2015). Arieli-Attali (2015) demonstrated that adaptive weighting scores and a monotonically increasing score can function as positive properties of points in assessment. However, points can have many different properties, therefore, future studies should explore variations of points systems in different learning settings. More knowledge about whether or not to use points and the type of points can have an important added value for practice. Learning environments can be designed more effectively when the different effects that gamification elements could have on learning and motivation are known (Nacke & Deterding, 2017).

It was expected that points would lead to improved retention compared to the achievement of participants in the control condition (Nebel et al., 2017). Nonetheless, in the present study, the addition of points in a retrieval practice program did not lead to improved retention. This might have to do with the fact that no improvement was found in the practice performance of participants. Consequently, participants did not practice more retrieval compared to the control condition and therefore their ability to retrieve it again in the future did not improve (Karpicke, 2012).

Effects of points in retrieval practice on intrinsic motivation

Higher intrinsic motivation for retrieval practice with points compared to retrieval practice without points showed that the manipulation with points as a single gamification feature was successful, specifically, for interest/enjoyment a large effect size was found. This finding is in line with previous studies that showed an increase of intrinsic motivation for a task when points were added (Mekler et al. 2013). In retrieval practice students find it difficult to make effective study choices (Dunlosky & Rawson, 2012), which could result in shorter practice sessions (Karpicke & Roediger, 2007), higher intrinsic motivation might help

students overcome this dislike of difficult practice (Cameron et al., 2005). This is especially meaningful for retrieval practice, because multiple retrieval practice sessions lead to ideal retention (Karpicke & Roediger, 2007). This implies that points, added as a single gamification feature, could lead to more retrieval practice sessions because it is experienced as more motivating than practice without points.

However, research in gamification implemented over a longer period of time found that the motivational effects disappeared over time (Attali & Arieli-Attali, 2015; Farzan et al., 2008). Therefore it is important to investigate how the positive effect of a single gamification element on motivation for retrieval practice develops over a longer period of time.

Based on literature about grading (Klapp, 2015), it was expected that the effect of points on intrinsic motivation would be different for low and high performers, however no significant difference was found. Unlike grading, points in this experiment were only visible for the individual and not for peers. Therefore points as implemented in the present study might not have led to negative consequences such as, focus on performance goals and the desire to be better than others (Pulfrey, Butera, & Buchs, 2011) which are known to be antecedents of grading. However, it should be taken into account that in the present study low performers had significantly lower scores on perceived competence independent of the condition they were in. It is therefore important to further investigate if points, implemented as a gamification feature, also influence this effect in a classroom setting. In competitive situations, social comparison seems most likely to have negative effects on the feeling of competence of low performing students (Marshall & Weinstein, 1984) and therefore adding points in a learning program used in classrooms might have a negative effect on the perceived competence of low performing students.

Future research

Even though participants had significantly higher intrinsic motivation in the points condition than in the no points condition, this motivation did not result in better performance during practice or improved retention. Still it would be interesting to further examine this relation since motivation is known to be of influence on persistence in learning (Vollmeyer & Rheinberg, 2000). Possibly, higher motivation in a condition with points results in more persistence resulting in more practice sessions leading to optimal retention (Karpicke & Roediger, 2007). Therefore, it would be interesting to execute the experiment again in a more authentic situation where participants can freely choose how much and when they want to use the online retrieval practice program. So, the relation between motivation, persistence and achievement can be investigated.

Another reason why higher intrinsic motivation did not lead to better performance might be because only one gamification feature was implemented in the retrieval practice program. Other gamification features were not tested in the present study. Despite the fact that it is important to look at gamification elements separately, it might be that simple gamification, such as the implementation of points or badges, is only effective until its novelty wears off (Xu, 2012). Additionally, Deterding (2015) states that gamified systems hardly add game design elements, although involvement in games emerges from dynamic interaction of users with all system components. Therefore, it is important that further research focuses on how to make gamification sustainable by looking at the intrinsic reward systems of game design.

In this study, all participants were older than 16, although retrieval practice is especially relevant for students in secondary education. It is possible that gamification in retrieval does not work the same way for teenagers since Attali and Arieli-Attali (2015) found that teenage students had higher likeability and perceived effort in the points condition. Furthermore, the ease of use of gamification is shown to decline with age (Koivisto &

Hamari, 2014). In an exploratory analysis in the present study, younger participants practiced significantly more words in the points condition than older participants. However, for the control condition there was no significant difference between younger and older participants, this implies that points have more positive effects on younger people. Therefore, the motivational effects of points, implemented in retrieval practice, might be even higher for teenagers. Future research needs to establish to what extent the effects found in this study can be generalized for teenagers.

Conclusion

To conclude, this study has contributed to current literature by investigating the effect of points, used as a single gamification feature, on the performance, retention and intrinsic motivation of people. Next to that, the attention to the role of relative performance within these relationships added new insights regarding points and motivation. While no improvement was found regarding the effect of points on performance and retention we can conclude that the use of points in retrieval practice does have an effect on people's motivation, which is promising. Higher motivation might be effective in promoting the use of effective study strategies such as retrieval practice. The present study demonstrated the complexity of implementing gamification in educational settings.

Reference list

Arieli-Attali, M. (2015). *Self adapted testing as formative assessment: Effects of feedback and scoring on engagement and performance* (Doctoral dissertation). Retrieved from ProQuest Dissertation & Theses Global database. (UMI No. 10013437)

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance? *Computers & Education, 83,* 57–63. doi: 10.1016/j.compedu.2014.12.012

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474–482. doi:10.1037/0022-0663.79.4.474

Cameron, J., Pierce, W. D., Banko, K. M., & Gear, A. (2005). Achievement-Based Rewards and Intrinsic Motivation: A Test of Cognitive Mediators. *Journal of Educational Psychology, 97*(4), 641–655. doi: 10.1037/0022-0663.97.4.641

Chapman, J. R., & Rich, P. J. (2018). Does educational gamification improve students' motivation? If so, which game elements work best? *Journal of Education for Business, 93*(7), 315–322. doi: 10.1080/08832323.2018.1490687

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York, United States: Academic Press.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research, 71*(1), 1–27. doi: 10.3102/00346543071001001

Deci, E. L., & Ryan, R. M. (2002). *Handbook of self-determination research*. Rochester (NY): The University of Rochester Press.

Deterding, S. (2015). The Lens of Intrinsic Skill Atoms: A Method for Gameful Design. Human–Computer Interaction, 30(3-4), 294-335. doi:10.1080/07370024.2014.993471

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness. *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek 11*. doi: 10.1145/2181037.2181040

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: a systematic mapping study. *Educational Technology & Society, 18*(3), 75–88. Retrieved from https://s3.amazonaws.com/academia.edu.documents

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. doi: 10.1016/j.learninstruc.2011.08.003

Farzan, R., Dimicco, J. M., Millen, D. R., Dugan, C., Geyer, W., & Brownholtz, E. A. (2008). Results from deploying a participation incentive mechanism within the enterprise. *Proceeding of the Twenty-sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*. doi:10.1145/1357054.1357145

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, England: Sage Publications.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & gaming*, *33*(4), 441-467. doi: 10.1177/1046878102238607

Gorilla Experiment Builder (2018) [Computer software]. Retrieved from www.gorilla.sc

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology*, *106*(1), 58–68. doi: 10.1037/a0033208

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - A literature review of empirical studies on gamification. *2014 47th Hawaii International Conference on System Sciences*. doi: 10.1109/hicss.2014.377

IBM SPSS (Version 25) [Computer software]. Retrieved from https://myibm.ibm.com/dashboard/

Intrinsic Motivation Inventory (IMI). (n.d.). Retrieved January 20, 2020, from http://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Jennings, M. J., Zumbo, B. D., & Joula, J. F. (2002). The robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicologica, 23*, 415-450. Retrieved from: https://www.redalyc.org/html/169/16923209/

Karpicke, J. D. (2012). Retrieval-Based Learning: Active Retrieval Promotes Meaningful Learning. *Current Directions in Psychological Science*, *21*(3), 157-163. doi:10.1177/0963721412443552

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. doi: 10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151-162. doi: 10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L. (2008). The Critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. doi: 10.1126/science.1152408

Kelle, S., Klemke, R., & Specht, M. (2013). Effects of game design patterns on basic life support training content. *Journal of Educational Technology & Society*, *16*(1), 275-285. Retrieved from: https://d1wqtxts1xzle7.cloudfront.net/38010015/24.pdf

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 302–323. doi: 10.1080/0969594x.2014.988121

Kline, P. (1999). *The handbook of psychological testing* (2nd ed). London, Engeland: Routledge.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, *10*(8), 707-710.

Marshall, H. H., & Weinstein, R. S. (1984). Classroom Factors Affecting Students' Self-Evaluations: An Interactional Model. *Review of Educational Research, 54*(3), 301-325. doi:10.3102/00346543054003301

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport, 60*(1), 48–58. doi: 10.1080/02701367.1989.10607413

Mckernan, B., Martey, R. M., Stromer-Galley, J., Kenski, K., Clegg, B. A., Folkestad, J. E., … Strzalkowski, T. (2015). We don't need no stinkin' badges: The impact of reward features and feeling rewarded in educational games. *Computers in Human Behavior, 45*, 299–306. doi: 10.1016/j.chb.2014.12.028

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior, 71*, 525–534. doi: 10.1016/j.chb.2015.08.048

Mekler, E. D., Tuch, A. N., Brühlmann, F., & Opwis, K. (2013). Disassembling gamification: The effects of points and meaning on user motivation and performance. *Conference on Human Factors in Computing Systems - Proceedings*, 1137–1142. doi: 10.1145/2468356.2468559

Monteiro, V., Mata, L., & Peixoto, F. (2015). Intrinsic motivation inventory: Psychometric properties in the context of first language and mathematics learning. *Psicologia: Reflexao e Critica, 28*(3), 434–443. doi: 10.1590/1678-7153.201528302

Nacke, L. E., & Deterding, S. (2017). The maturing of gamification research. *Computers in Human Behavior, 71*, 450–454. doi: 10.1016/j.chb.2016.11.062

Nebel, S., Schneider, S., Beege, M., & Rey, G. D. (2017). Leaderboards within educational videogames: The impact of difficulty, effort and gameplay. *Computers & Education, 113*, 28–41. doi: 10.1016/j.compedu.2017.05.011

Nvivo (Version 12.6) [Software]. (2019). Retrieved from https://www.qsrinternational.com

Price, P. C., Jhangiani, R. S., Chiang, I. A., Leighton, D. C., & Cuttler, C. (2017). Experimental design. In *Research Methods in Psychology* (3rd American Edition, pp. 79–82). Retrieved from https://opentext.wsu.edu/carriecuttler/chapter/experimental-design/

Pulfrey, C., Butera, F., & Buchs, C. (2011). Why grades engender performance avoidance goals: The mediating role of autonomous motivation. *PsycEXTRA Dataset, 103*(3), 683-700. doi: 10.1037/e549962013-069

Rijn, H. van (2010). *SlimStampen.nl* [Computer software]. Retrieved from https://onderzoek.kennisnet.nl/app/uploads/2016/12/SlimStampen.pdf

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. doi: 10.1016/j.tics.2010.09.003

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 101860. doi: 10.1016/j.cedpsych.2020.101860

Sailer, M., Hense, J., Mandl, H., & Klevers, M. (2014). Psychological perspectives on motivation through Gamification. *Interaction Design and Architecture(s) Journal,* (19), 28–37. Retrieved from https://mediatum.ub.tum.de

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies, 74*, 14–31. doi: 10.1016/j.ijhcs.2014.09.006

Sense, F., Behrens, F., Meijer, R. R., & Rijn, H. V. (2016). An individuals rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305–321. doi: 10.1111/tops.12183

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, United States: Routledge.

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, *38*(2), 244–253. doi: 10.3758/mc.38.2.244

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442. doi: 10.3758/s13421-012-0274-5

Vollmeyer, R., & Rheinberg, F. (2000). Does motivation affect performance via persistence? *Learning and Instruction, 10*(4), 293-309. doi:10.1016/s0959-4752(99)00031-6

Wang, H., & Sun, C. T. (2011). Game reward Systems: gaming experiences and social meanings. In C. Marinka, K. Helen, & W. Annika (Eds.), *Proceedings of the DiGRA 2011 Conference: Think design play*. Retrieved from http://www.digra.org/dl/db/11310.20247.pdf

Werbach, K., & Hunter, D. (2012). *For the win: How game thinking can revolutionize your business.* Pennsylvania, United States: Wharton Digital Press.

Xu, Y. (2012). *Literature Review on Web Application Gamification and Analytics*. Honolulu, Hawaii: University of Hawaii. http://citeseerx.ist.psu.edu/

Appendices

Appendix A

Informed Consent

Bedankt voor je interesse in dit online onderzoek!

Het doel van dit onderzoek is om oefeningen te verbeteren waarmee nieuwe woorden geleerd kunnen worden via de computer.

Uitvoering

Het onderzoek bestaat uit twee sessies waarin je op je eigen computer Franse woorden leert. Deze sessies vinden 2 dagen na elkaar plaats. De eerste sessie duurt ongeveer 30 minuten, de tweede sessie 15 minuten. Voor deze tweede sessie krijg je een herinneringsmail.

Beloning

Per vijftig deelnemers verloten we een bol.com-bon ter waarde van 25 euro. Om hier kans op te maken dien je beide sessies afgerond te hebben.

Vereisten

Om deel te kunnen nemen moet je 16 jaar of ouder zijn.

Op de volgende pagina vind je informatie over jouw rechten als proefpersoon. Als je akkoord gaat met deze voorwaarden, kun je gelijk van start gaan.

-Page break-

Op deze pagina vind je informatie over de voorwaarden van deelname en jouw rechten als proefpersoon. Wij verzoeken je vriendelijk om deze pagina goed door te lezen alvorens je kiest of je wel of niet mee wilt doen aan het onderzoek.

Voorwaarden voor deelname

Je bent ten minste **16 jaar oud**. Voorkennis van de Franse taal is **niet** nodig. Je mag **niet** Frans zijn opgevoed (tweetalig) of de Franse taal op hoog niveau (C1 of hoger) beheersen.

Wat houdt het onderzoek in?

Het onderzoek bestaat uit twee sessies. In beide sessies doe je oefeningen op de computer om Franse woorden te leren. Je gaat bijvoorbeeld woorden bestuderen en vertalen. Tussendoor vragen we je om vragenlijsten in te vullen waarin je de oefeningen beoordeelt. Deze sessies vinden 2 dagen na elkaar plaats. De eerste sessie duurt ongeveer 30 minuten, de tweede sessie 15 minuten. Voor deze tweede sessie krijg je een herinneringsmail.

Toestemming

Deelname aan dit onderzoek is **vrijwillig**. Je kan op elk moment stoppen als je het onderzoek toch niet af wilt maken. Je kunt dan gewoon je browser afsluiten.

Dataverzameling en Privacy

Tijdens het onderzoek worden gegevens opgeslagen, zoals de antwoorden die je geeft. Deze gegevens worden vertrouwelijk behandeld en alleen voor onderzoek gebruikt.

De data worden geanonimiseerd. Dat wil zeggen dat de dataset geen informatie bevat over jouw identiteit. Je naam staat dus niet in de dataset. Ook je mailadres wordt alleen los van de overige data opgeslagen. Daarnaast zal het in de publicaties over het onderzoek niet herkenbaar zijn welke antwoorden een bepaalde participant gegeven heeft.

Vragen?

Als je nog vragen hebt over het onderzoek, kun deze je mailen naar woordenleren@uu.nl.

Op de volgende pagina kun je bevestigen dat deze voorwaarden van deelname duidelijk zijn en kun je het experiment starten. Wil je niet meedoen, dan kun je deze website gewoon afsluiten.

-Page break-

Toestemmingsverklaring

Door hieronder het vakje aan te vinken, bevestig ik het volgende:

- Ik ben **16 jaar of ouder.**
- Ik ben naar tevredenheid over het onderzoek geïnformeerd.
- Ik heb de informatie over het onderzoek goed gelezen.
- Ik heb de voorwaarden van deelname gelezen.
- Ik heb met de onderzoeker contact opgenomen over eventuele vragen over het onderzoek. Mijn vragen zijn naar tevredenheid beantwoord.
- Ik heb goed over deelname aan dit onderzoek kunnen nadenken.
- Ik begrijp dat deelname aan dit onderzoek vrijwillig is.
- Ik heb het recht mijn toestemming op ieder moment weer in te trekken zonder dat ik daarvoor een reden behoef op te geven.
- Ik ga ermee akkoord dat tijdens dit onderzoeksgegevens voor wetenschappelijke doelen worden verkregen en bewaard.

• Ik stem hierbij toe voor deelname aan het onderzoek.

-Page break-

Appendix B

Start Questionnaire

Om de data van sessie 1 en sessie 2 van het experiment aan elkaar te kunnen koppelen, gebruiken we de antwoorden op de volgende vragen:

Vul hieronder in:

- Twee laatste letters van je voornaam
- Eerste twee getallen van mobiele telefoonnummer na 06
- Je huisnummer zonder toevoeging

Bijv. te4597

-Page break-

Wat is je geslacht?
Selecteer: Man; Vrouw; Anders
Wat is je leeftijd?

Wat is uw hoogst genoten opleiding? (*Dit is de opleiding die u nu volgt of ooit gevolgd heeft, u hoeft deze niet afgerond te hebben*)
Selecteer: Basisschool; Middelbare school VMBO; Middelbare school HAVO; Middelbare school VWO; MBO; HBO; WO Bachelor; WO Master; Gepromoveerd; Anders

Welke omschrijving is passend voor u?
Selecteer: Ik ben middelbare scholier; Ik ben student; Ik ben werkend; Ik ben werkzoekend; Anders, namelijk ...

Hoeveel jaar heb je Frans als middelbare schoolvak gehad?
Selecteer: 0; 1; 2; 3; 4; 5; 6

Heb je, buiten de Franse lessen op de middelbare school, ervaring met de Franse taal?
Selecteer: Nee; Ja (Leg uit)

Appendix C

Instruction Retrieval Practice

Control condition:

Let op! Zorg dat je browser op fullscreen staat. Druk hiervoor op F11 (Fn+F11) of CMD+CTRL+F voor Apple computers.

-Page break-

Instructie

In de komende 12 minuten zal je de Nederlandse vertalingen van een aantal Franse woorden leren.

Het is jouw taak om de Nederlandse vertaling van de woorden in te typen. Het kan zijn dat je een aantal keer hetzelfde woord moet vertalen.

De eerste keer dat je een Frans woord ziet wordt de Nederlandse vertaling eronder getoond, die je moet overtypen.

Druk op **Volgende** om te beginnen.

Points condition:

Let op! Zorg dat je browser op fullscreen staat. Druk hiervoor op F11 (Fn+F11) of CMD+CTRL+F voor Apple computers.

-Page break-

Instructie

In de komende 12 minuten zal je de Nederlandse vertalingen van een aantal Franse woorden leren.

Het is jouw taak om de Nederlandse vertaling van de woorden in te typen. Het kan zijn dat je een aantal keer hetzelfde woord moet vertalen.

De eerste keer dat je een Frans woord ziet wordt de Nederlandse vertaling eronder getoond, die je moet overtypen.

Je kan punten verdienen door woorden te oefenen. Bij een goed antwoord krijg je 10 punten. Bij een fout antwoord -1 punt. Je score wordt weergeven boven de timerklok.

Druk op **Volgende** om te beginnen.

Appendix D

Intrinsic Motivation Inventory

Control condition

Je hebt net een oefening gedaan. Je gaat nu vragen beantwoorden over de oefening die je net hebt gedaan.

Kies per vraag de optie jij passend vindt.

Ik had veel plezier tijdens deze oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Ik denk dat ik vrij goed ben in deze oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Deze oefening was erg leuk.

Helemaal niet waar 1234567 **Helemaal waar**

Tijdens deze oefening voelde ik me bekwaam (=vaardig).

Helemaal niet waar 1234567 **Helemaal waar**

Ik denk dat deze oefening nuttig is om te leren voor een toets.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Beantwoord alsjeblieft ook de volgende vragen over de oefening die je net hebt gedaan.

Kies per vraag de optie jij passend vindt.

Deze oefening hield mijn aandacht vast.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond dit een saaie oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Ik denk dat deze oefening mij helpt bij het leren van nieuwe woorden.

Helemaal niet waar 1234567 **Helemaal waar**

Ik ben tevreden over mijn prestaties tijdens deze oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond deze oefening zeer interessant.

Helemaal niet waar 1234567 **Helemaal waar**

Deze oefening heb ik niet zo goed gedaan.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Beantwoord alsjeblieft ook de volgende vragen over de oefening die je net hebt gedaan.

Kies per vraag de optie die jij passend vindt.

Tijdens het oefenen wilde ik goed presteren.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond het oefenen stressvol.

Helemaal niet waar 1234567 **Helemaal waar**

Het voelde alsof ik tijdens het oefenen toewerkte naar een doel.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Tijdens het oefenen zag je rechts in beeld een timer. De volgende vragen gaan over je waarneming van deze extra informatie op het scherm.

Kies per vraag de optie die jij passend vindt.

Ik vond de timer begrijpelijk.

Helemaal niet waar 1234567 **Helemaal waar**

Tijdens het oefenen werd ik afgeleid door de timer.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond de timer een waardevolle toevoeging aan de oefening.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Kies de optie jij passend vindt.

Hoeveel procent van de vertalingen die je net hebt geoefend, denk je over twee dagen nog te weten?

0 **%**

Points condition

Je hebt net een oefening gedaan. Je gaat nu vragen beantwoorden over de oefening die je net hebt gedaan.

Kies per vraag de optie jij passend vindt.

Ik had veel plezier tijdens deze oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Ik denk dat ik vrij goed ben in deze oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Deze oefening was erg leuk.

Helemaal niet waar 1234567 **Helemaal waar**

Tijdens deze oefening voelde ik me bekwaam (=vaardig).

Helemaal niet waar 1234567 **Helemaal waar**

Ik denk dat deze oefening nuttig is om te leren voor een toets.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Beantwoord alsjeblieft ook de volgende vragen over de oefening die je net hebt gedaan.

Kies per vraag de optie jij passend vindt.

Deze oefening hield mijn aandacht vast.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond dit een saaie oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Ik denk dat deze oefening mij helpt bij het leren van nieuwe woorden.

Helemaal niet waar 1234567 **Helemaal waar**

Ik ben tevreden over mijn prestaties tijdens deze oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond deze oefening zeer interessant.

Helemaal niet waar 1234567 **Helemaal waar**

Deze oefening heb ik niet zo goed gedaan.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Beantwoord alsjeblieft ook de volgende vragen over de oefening die je net hebt gedaan.

Kies per vraag de optie jij passend vindt.

Tijdens het oefenen wilde ik goed presteren.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond het oefenen stressvol.

Helemaal niet waar 1234567 **Helemaal waar**

Het voelde alsof ik tijdens het oefenen toewerkte naar een doel.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Tijdens het oefenen zag je rechts in beeld een timer, je punten (score). De volgende vragen gaan over je waarneming van deze extra informatie op het scherm.

Kies per vraag de optie jij passend vindt.

Ik vond de punten/score begrijpelijk.

Helemaal niet waar 1234567 **Helemaal waar**

Tijdens het oefenen werd ik afgeleid door de punten/score.

Helemaal niet waar 1234567 **Helemaal waar**

Ik vond de punten/score een waardevolle toevoeging aan de oefening.

Helemaal niet waar 1234567 **Helemaal waar**

Tijdens het oefenen probeerde ik een hoge score te behalen.

Helemaal niet waar 1234567 **Helemaal waar**

-Page break-

Kies de optie die jij passend vindt.

Hoeveel procent van de vertalingen die je net hebt geoefend, denk je over twee dagen nog te weten?

0 **%**

Appendix E

End Questionnaire First Session

Heb je tijdens het oefenen een van deze dingen gedaan? Je kan meerdere antwoorden aanvinken.

- Aantekeningen gemaakt (bijvoorbeeld woorden opgeschreven of een foto genomen)
- Woorden online opgezocht
- Meer dan 1 minuut niet naar het onderzoek gekeken door afleiding
- Geen van deze dingen gedaan

-Page break-

Je hebt nu twee verschillende oefeningen gedaan om woorden te leren.

Welke oefening zou jij gebruiken voor een volgende keer om woorden te leren? Kies er één:

Selecteer: Oefening met een timer; Oefening met een timer en punten; Ik heb geen voorkeur

Wat vond je van de punten tijdens het oefenen en hoe ben je met de punten omgegaan?

-Page break-

Appendix F

Introduction second session

Welkom terug bij het experiment!

Bij deze tweede sessie vragen we je een aantal Franse woorden, waaronder degenen die je tijdens de eerste sessie geleerd hebt, te vertalen. Per woord heb je maximaal één minuut om de vertaling in te vullen. Als je geconcentreerd werkt, kun je deze sessie binnen 15 minuten afronden. Het is wederom belangrijk dat je een **rustige werkplek** opzoekt en de sessie niet onderbreekt.

Hiermee willen we graag meten hoeveel Franse woorden je onthouden hebt van de eerste sessie. Probeer daarom de woorden uit je hoofd te vertalen en zoek de betekenis van de woorden **niet** op. Woordenboeken of je telefoon gebruiken is **niet toegestaan**. Als je een vertaling niet weet, kun je het antwoord vrij laten.

Je dient je antwoorden in door op Enter (return) te drukken.

-Page break-

Om de data van sessie 1 en sessie 2 van het onderzoek te koppelen, gebruiken we de antwoorden op de volgende vragen:

Zorg dat je dezelfde gegevens invult als bij de eerste sessie.

Vul hieronder in:

- Twee laatste letters van je voornaam
- Eerste twee getallen van mobiele telefoonnummer na 06
- Je huisnummer zonder toevoeging

Bijv. te4597

Appendix G

Debriefing

Bedankt voor je deelname aan het onderzoek van de Universiteit Utrecht!

Bij interesse kun je hieronder meer lezen over het doel van het onderzoek. We geven ook enkele praktische studietips, gebaseerd op het onderzoek. Bij vragen kun je contact opnemen met de onderzoekers via woordenleren@uu.nl.

1. Waarover gaat het onderzoek?

Het onderzoek waaraan je hebt deelgenomen gaat over een handige manier waarop je kunt leren: overhoren. Een belangrijke eigenschap van het geheugen is dat het oproepen van een herinnering de inhoud en latere toegankelijkheid van de herinnering beïnvloedt.

Dit is relevant voor iedereen die veel informatie wil onthouden, zoals bij het uitbreiden van je woordenschat voor een vreemde taal. Een oproep van een herinnering uit het geheugen wordt ook wel *retrieval* genoemd. Elke *retrieval* maakt het makkelijker om dezelfde herinnering later opnieuw op te van nieuwe woorden en de vertaling, dit voordeel wordt testeffect genoemd.

Dit onderzoek kijkt naar retrieval practice op de computer. In dit onderzoek is gamification toegepast. Bij gamification wordt een onderdeel van game design (zoals punten en levels) toegepast in een niet game omgeving. Het toevoegen van punten en levels aan online retrieval practice kan mogelijk zorgen voor meer motivatie voor het woorden leren en een betere prestatie. Wij zullen de data analyseren en we proberen een antwoord te geven op de roepen. *Retrieval practice* is een effectieve techniek om woorden op lange termijn te onthouden. *Retrieval practice* is effectiever dan het herhaaldelijk doorlezen van nieuwe woorden en de vertaling, dit voordeel wordt *testeffect* genoemd.

Dit onderzoek kijkt naar *retrieval practice* op de computer. In dit onderzoek is *gamification* toegepast. Bij *gamification* wordt een onderdeel van game design (zoals punten en levels) toegepast in een niet game omgeving. Het toevoegen van punten en levels aan online *retrieval practice* kan mogelijk zorgen voor meer motivatie voor het woorden leren en een betere prestatie. Wij zullen de data analyseren en we proberen een antwoord te geven op de vraag of scores en levels in een overhooroefening nuttig en motiverend zijn.

2. Zelf retrieval practice gebruiken?

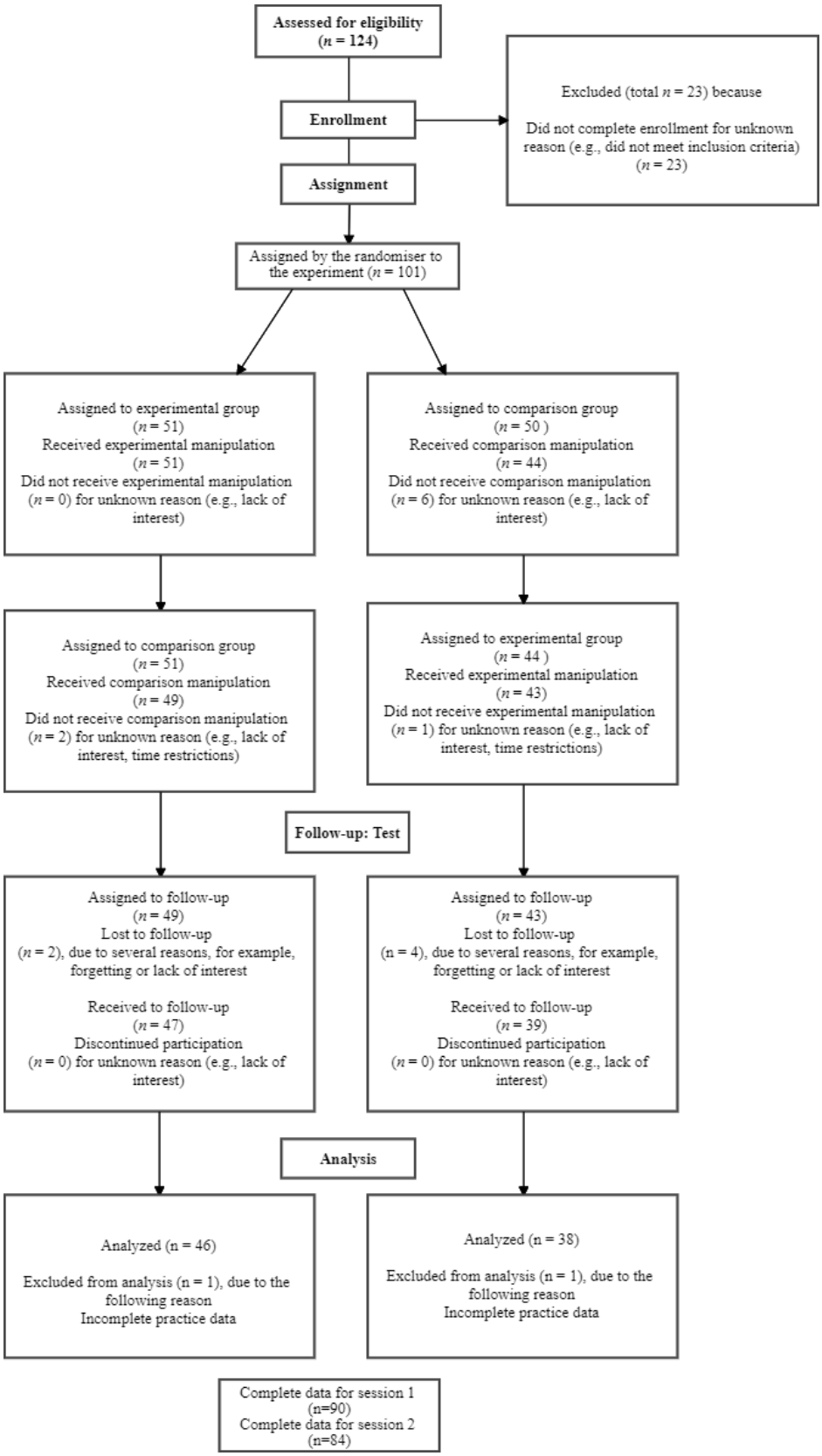
Wil je in het vervolg woorden of feiten leren met een overhoorprogramma ontwikkeld op basis van recente wetenschappelijk kennis over *retrieval practice*? Dat kan via Slimstampen. Slimstampen is adaptief in de volgorde waarin de feiten worden aangeboden. De volgorde van feiten wordt aangepast op de juistheid en de antwoordsnelheid.

3. Franse woorden

Ben je benieuwd naar de woordenlijst waarmee je hebt geoefend stuur dan een mailtje naar woordenleren@uu.nl.

Appendix H

Flow of Participants Through Each Stage of the Experiment



Appendix I

Tree Map of Experience of Points

Name	Files	References
Positief	0	0
Motiverend	1	21
Hoger scoren	1	15
Naar doel toe werken	1	9
Volgende keer nuttig	1	4
Gevoel van competentie	1	4
Handig	1	3
Competitief	1	2
Beter gepresteerd door punten	1	2
Inzicht hoe het gaat	1	1
Voortgang zien	1	1
Feedback	1	1
Neutraal	0	0
Niet mee bezig geweest	1	22
Geen toevoeging	1	9
Punten laat opgevallen	1	6
Timer alleen voorkeur	1	5
Te weinig minpunten	1	5
Niet gezien	1	3
Veel punten bij goed antwoord	1	3
Negatief	0	0
Onduidelijk	1	15
Afleidend	1	11
Stressvol	1	4
Irritant	1	3
Competitief negatief	1	1

Appendix J

FETC form

Section 1: Basic Study Information

Name student:

Bente van Thuijl

Name(s) of the supervisor(s):

Gesa van den Broek

Title of the thesis (plan):

The Effect of Points in Retrieval Practice on Performance and Motivation

Does the study concern a multi-center project, e.g. a collaboration with other organizations, universities, a GGZ mental health care institution, or a university medical center?

Yes / ~~No~~

If yes: Explain.

The project is in collaboration with the Psychology Department of the University of Groningen. The Department will adapt their computer program *SlimStampen* for the different conditions in this study.

Where will the study (data collection) be conducted? If this is abroad, please note that you have to be sure of the local ethical codes of conducts and permissions.

The data collection will be online conducted in the Netherlands.

Section 2: Study Details I

Will you collect data?

Yes ~~/ No~~ Yes □ Continue to question 11 No □ Continue to question 7

Where is the data stored?

Is the data publicly available?

Yes / No If yes: Where?

Can participants be identified by the student? (e.g., does the data contain (indirectly retrievable) personal information, video, or audio data?)

Yes/ No If yes: Explain.

If the data is pseudonymized, who has the key to permit re-identification?

Section 3: Participants

What age group is included in your study?

16 years and older

Will be participants that are recruited be > 16 years? ~~Yes~~/No

Will participants be mentally competent (wilsbekwaam in Dutch)? Yes/~~No~~

Does the participant population contain vulnerable persons?

(e.g., incapacitated, children, mentally challenged, traumatized, pregnant)

~~Yes~~/No

If you answered 'Yes' to any of the three questions above: Please provide reasons to justify why this particular groups of participants is included in your study.

What possible risk could participating hold for your participants?

There are no risks in this study.

What measures are implemented to minimize risks (or burden) for the participants?

What time investment and effort will be requested from participants?

Two sessions, one retrieval practice session and one session for the final test. The first session will take approximately 45 minutes and the second session half an hour. Mental effort is requested from the participants, because they are practicing French words.

1. **Will participants be reimbursed for their efforts? If yes, how? (financial reimbursement, travelling expenses, otherwise). What is the amount? Will this compensation depend on certain conditions, such as the completion of the study?**

There will be no compensation for their efforts.

1. **How does the burden on the participants compare to the study's potential scientific or practical contribution?**

Time and mental effort are the burdens in this study. This study can contribute to the optimization of the learning outcomes and motivation of retrieval practice.

1. **What is the number of participants? Provide a power analysis and/or motivation for the number of participants. The current convention is a power of 0.80. If the study**

deviates from this convention, the FERB would like you to justify why this is necessary.

(Note, you want to include enough participants to be able to answer your research questions adequately, but you do not want to include too many participants and unnecessarily burden participants.)

For the paired samples *t*-test with an effect size of 0.3 (small), a minimum of 71 participants was required to get a power of 0.80. Therefore a maximum 90 participants will participate in this study.

How will the participants be recruited? Explain and attach the information letter to this document.

Participants will be online recruited via social media.

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

They can decide anytime

Please explain the consent procedures. Note, active consent of participants (or their parents) is in principle mandatory. Enclose the consent letters as attachments. You can use the consent forms on Blackboard.

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

Yes, participants can stop at any moment during the research. This is stated in the informed consent and also is mentioned during the explanation of every session.

1. **Will the participants be in a dependent relationship with the researcher?**

~~Yes~~ / No

If yes: Explain.

1. **Is there an independent contact person or a general email address of a complaint officer whom the participant can contact?**

dr. Lisette Hornstra and dr. Anouschka van Leeuwen at edu.acma.thesis@uu.nl.

1. **Is there an independent contact person or a general email address of a complaint officer whom the participant can contact in case of complaints?**

klachtenfunctionaris-fetcsocwet@uu.nl.

Section 4: Data management

1. **Who has access to the data and who will be responsible for managing (access to) the data?**

Researcher Bente van Thuijl will have access to the data and will be responsible for managing the data, supervisor Gesa van den Broek will also have access to the data.

1. **What type of data will you collect or create? Please provide a description of the instruments.**

The age and gender of participants will be collected. With three subscale of the Intrinsic Motivation Inventory data will be collected about the interest/enjoyment, perceived competence and value/usefulness of students on two computer tasks. Also the performance of participants on the final vocabulary test will be measured.

1. **Will you be exchanging (personal) data with organizations/research partners outside the UU?**

~~Yes/~~No

If yes: Explain**.**

1. **If so, will a data processing agreement be made up?**

~~Yes / No~~

If yes: Please attach the agreement.

If no: Please explain.

1. **Where will the data be stored and for how long?**

The data will be stored in a personal folder in YODA (YourData) for 7 years.

1. **Will the data potentially be used for other purposes than the master's thesis? (e.g., publication, reporting back to participants, etc.)**

No

1. **Will the data potentially be used for other purposes than the master's thesis? (e.g., publication, reporting back to participants, etc.)**

~~Yes~~ / No

If yes: Explain.