university of groningen

faculty of science and engineering

Influence of input methods on retention for vocabulary learning systems; Comparing keyboard and scribble input

Master Thesis

Annika Schnabel (s3826007)

Faculty of Science and Engineering, University of Groningen

Master's in Human-Machine Communication

Internal Supervisor: dr. Jelmer P. Borst (Artificial Intelligence, University of Groningen)

External Supervisor: prof. dr. Hedderik van Rijn (Faculty of Behavioural and Social Sciences, University of Groningen)

August 28, 2021

Abstract

It is often hypothesized that word retention is better when students are learning vocabulary by writing instead of typing the answer. Informal reports of teachers say that writing items while studying leads to better retention than typing the items. However, most research compares word retention in writing on paper with typing on a keyboard. This may lead to confounds since one compares different input methods and different technologies. The development of smartpens and handwriting recognition makes it possible to compare input methods with reduced confounds. This study compares whether word retention is better when using either a smartpen or a keyboard on a tablet. For that, a vocabulary learning app was created, which allows both smartpen input and keyboard input. The app notes the reaction times of the user, and the word retention of the participants is checked during learning and two testing sessions. One of those tests is administered right after the session, the other on the next day. It is hypothesized that using a smartpen improves retention compared to using a keyboard. However, the results show that participants have similar word retention in both conditions. This suggests that, at least with the current technologies digital writing does not outperform typing.

Keywords: Tablet, second language learning, smart pen, typing, word retention

Contents

Influence of input methods on retention for vocabulary learning systems; Comparing keyboard and scribble input

When learning a new language it is important to memorize new words and know what they mean. The goal of vocabulary learning is to retain words in a new language quickly but what is the best way to achieve this? Different strategies have been developed throughout the years.

A learner can create a chart in a notebook with three columns, one for the word, one for its translation and one for notes. Afterward, the notebook can be used to study the words by going through the chart in regular intervals. It is also possible to use a method called semantic mapping, which allows the learner to create a map that shows the connections of the different words. This may help the learner to associate words with a similar meaning. Another popular method is to create word cards. On each card the word is written on one side and on the other side its translation is noted. This allows the learner to easily test themselves. Those cards can be put in a box and sorted by the difficulty level of the words. This way a learner can decide how often a word needs to be repeated. All those strategies have in common that one writes down the word and its translation with a pen on paper. However, for those strategies it is only necessary to write down a word and its translation once. Therefore, it is common to have a separate piece of paper on which the word will be written down during repetition. This way it is only read but also written down during each repetition.

Research shows that more technology is used in classrooms. This results in more self-guided instructions for students and more notes are taken on a computer instead off with a pen (Smoker et al., 2009). With the advancement of technology new vocabulary learning methods, online fact learning systems have been developed, which can either be used on a computer or on a cell phone. Both websites and apps use the same learning strategies. One thing almost all online fact learning systems have in common is the way the user interacts with the system. Usually, the answer is either presented in a multiple choice format or the user has to type the answer. The question formats can filling blanks, showing answer options, or asking for a translation to a word. All of those can be answered either by multiple choice or typing.

One of the biggest differences between traditional and online fact learning systems is the way a learner gives the translation. In offline systems the learner generally writes the answer and in online systems the learner types the answer. The question is if input methods affect word retention. Past research shows that writing leads to better retention compared to typing (Mueller and Oppenheimer, 2014; Smoker et al., 2009).One limitation of the research is that they are based on more complicated tasks like memorizing notes from a lecture and therefore may not hold for factual knowledge. Newer technology, smart pens, make it possible to write on the tablet. It is now even possible to convert handwritten text to standard text, which makes it possible to check the learners answer on the tablet. Taking those research results and the advances in technology into account one has to ask whether it would be beneficial for online fact learning systems to allow input via smart pen.

Research Questions

The idea of of the present study is to use two different input methods (keyboard and smart pen) in a factual learning task to determine whether the input method influences word retention. The goal is to find the best input method, to increase retention on a test in a factual learning task. An online vocabulary learning system is used for the factual learning task. The result may show that fact learning systems should rather implement input via smart pen instead of keyboard to improve word retention.

To summarize this thesis focuses on the following research questions:

Q1. Can similar about the retrieval process be obtained with pen input and keyboard input?

Q2. Does writing the answers instead of typing the answers improve word retention?

The first research question determines whether the measurement for the onset of the pen is comparable to the measurement of the onset of typing. For this the time point at which people start moving the pencil on the text-field will be compared to the time point at which learners start typing. If the onset is comparable it can be proceeded to the second research question. For answering this question research will be conducted testing word retention using different input methods. It is hypothesized that using a smart pen for writing will improve retention compared to using the keyboard in a factual learning task.

Thesis Outline

In this thesis first some background literature will be reviewed. First, the importance of writing will be evaluated and how it differs from typing. Afterward, previous studies comparing writing with typing will be highlighted. After that the idea behind the app will be explained and how the app was build. Next, the study itself will be explained and its results will be shown. At the end the result will be discussed in context with other studies and indications of input methods for future studies will be given.

Background Literature

Different cognitive processes underlie writing and typing due to the differences in operating a keyboard or writing with a pen. The first difference to note is that writing is done using one hand while typing is usually done using two hands. Here it is assumed that the user is efficient when using a keyboard and can use the 10-finger system. Scientifically speaking handwriting requires unimanual movement while typing requires bimanual movement. Another difference is that writing is a slower process than typing. During handwriting the eyes closely follow the movement of the pen, this means that the visual attention is concentrated. This is not the case for typing. During typing the visual attention is detached from the process of typing. Usually one types blindly and follows the formation of the letters on the screen. This means that typing is divided into two separate spaces (motor and visual spaces), which are distinct and spatiotemporally separated.

Another difference concerns the production of the characters. In writing the writer has to form each letter as close to the standard as possible so readers can recognize the character (graphomotor component). This is not the case in typing. In typing a writer has to spatially locate the character and press a key.

Those differences might have an impact on cerebral representation and therefore on letter memorization. One fMRI study by Longcamp et al. (2008) in which learners had to either write or type new characters compared the activation in different brain regions. They discovered that the same areas are active during writing and typing. However, the fMRI data showed difference in recognition performance in related neural pathways. it showed that some brain areas (left Broca's area, bilaeral AIP, left dorsal premotor area, and left postcentral regions) were activated during visual processing of newly written and over-learned letters. This means that a reactivation of motor knowledge took place during visual processing if the letter was written but not if it was typed. This conclusion was supported by the fact that only the left-side of the brain was activated (writers were right-handed) (Longcamp et al., 2008). Another study using Japanese writing supported those results. The fMRI scans showed that some brain regions were activated during both writing and typing (left superior parietal lobule, left supramarginal gyrus, and left premotor cortex close to Exner's area). However, they discovered that some regions were more active during typing (posteromedial intraparietal cortex activation, rostral activity in the left premotor cortex). They concluded that the biggest differences between the two input methods happen in the transition process (retrieving the word from working memory and planning the necessary motion to write or type the word). The results also showed a difference in the motoric process, hence the movement of the hand(s) (Higashiyama et al., 2015).

Those studies were conducted using single characters and not words. Studies comparing word retention between writing and typing usually have people either recall whole words or the content of a lecture. In one of those studies a training session was conducted with children. They had to write down or type the word in order to learn how to spell it. Afterward, they were tested on how well they could spell the word. The study discovered that the children performed better if the word was written by hand compared to being written with a keyboard (Cunningham and Stanovich, 1990). However, another study tried to replicate those results but could not find any spelling differences between writing and typing. Subjects in the study by Cunningham have a lot higher socioeconomic status (SES) compared to the participants in the study by Vaughn. Students in a higher SES population usually have more exposure to books and writing. This difference may account for the superior performance of student in a high SES population in the writing condition. Therefore, SES may partially be the reason for the different results in the two studies (Vaughn et al., 1992).

Other studies confirm the results of Cunningham and Stanovich (1990). In one study children were presented with a list of words and needed to either write them down or type them. Afterward, they were given a recall task to test how many words the children memorized in each task. The results show that more words were recalled in the writing condition than in the typing condition. This means that retention seemed to be better writing the words instead of typing the words (Smoker et al., 2009).

Another study went a step further and tested differences in word retention after taking notes during a lecture. The study examined the recall of information after taking notes during a lecture. In the study it was discovered that participants could recall more information if they took notes with the computer instead of with the pen. The study also showed that how the notes are taken influences the effectiveness of note taking. They discovered that writing by hand it did not matter what method was used to take the notes. However, when taking notes on a computer, performance was better then transcribing the notes instead of taking organized notes (Bui et al., 2013). Therefore, it can be concluded that taking a lot of notes is beneficial for memorization.

Two different hypothesis explain how note taking can affect retention, the encoding hypothesis and the external storage hypothesis. The encoding hypothesis suggests that writing improves retention due to the slower process, which allows a more deep processing of the information. On the other hand, the external-storage hypothesis highlights the benefits of being able to review notes, even from somebody else. According to that hypothesis having more notes is beneficial. Usually, students have more notes when typing, than when writing. That is why this hypothesis suggests that typing improves retention. The benefits of the second hypothesis could be seen in the study by Bui et al. (2013). In another study by Mueller and Oppenheimer (2014) students either took notes by hand or on a laptop. Afterward, students had to respond to both factual-recall questions and conceptual-application questions about the lecture. This means they had to be able to recall facts and be able to give longer answers to questions using the knowledge from the lectures. The results show that participants scored better in the writing condition than in the typing condition. This would validate the encoding hypothesis and show that the slower process of writing helps to memorize content. However, no difference could be observed in the factual knowledge part in the immediate test. Writing only scored better in the delayed test (Mueller and Oppenheimer, 2014). This shows that there may be a difference between learning factual knowledge and more complex tasks.

All of the previously mentioned studies have one thing in common. They compare writing with a pen on paper with typing on a keyboard. This way not only the input methods are compared but also technologies. This could lead to confounds. Participants may simply not be as familiar with computer entry as with using pen and paper (Van Hove et al., 2017). The smart pen gives the option to test both writing and typing on a tablet. This way technology is no longer be a confounding factor. One study by Van Hove et al. (2017) studied word retention in a factual-learning task using keyboard, smart pen, and tapping (multiple-choice questions) as input option. In their study children had to learn French vocabulary. During the study participants had to first memorize French words and were than asked to do exercises using one of the three input option. Afterward, the participants were tested on their vocabulary retention in an immediate test and a delayed post test (10 days later). The results show that each of the input modalities were effective. However, the two groups, which used writing and typing for input, showed better vocabulary recall than the multiple-choice group. No difference between the writing and the typing condition could be found. Those results indicate that processing information at a higher processing level is beneficial. Therefore, it may be necessary to conduct more research using smart pens to verify those results without having the confounding factor of technology.

Method

Participants were tested in two conditions. Both of those were done using a vocabulary learning app on an iPad. In one condition participants had to type the answer and in the other one, they had to write the answer with a smart pen. In each condition, participants learned Swahili-English word pairs. The reaction times and whether they gave the correct answer was recorded.

Vocabulary Learning

In this research, participants are asked to memorize Swahili-English word pairs by responding with the English word to a Swahili word cue. 50 words were chosen from a list compiled by van den Broek et al. (2014). Those were divided into two lists containing 25 words each. All chosen words were nouns.

Language

Swahili is a language commonly spoken in countries along the east coast of Africa. It is a Bantu language. Swahili was influenced by Arabic but uses the Roman alphabet (Encyclopædia Britannica inc., 2014). One of the advantages of using Swahili is that participants from a country with an Indo-European language (Jasanoff and Cowgill, 2020) are normally not familiar with it but the language uses the same letters and phonemes.

This means that it can be assumed that students have the same level of understanding of the language and cannot infer a translation by using similarities between Swahili and another language. Another advantage is that different vocabulary learning studies have been published using this language therefore it is well documented and easy to choose adequate words for one's study (Bangert and Heydarian, 2017; Carpenter et al., 2008; Nelson and Dunlosky, 1994; Sense et al., 2016).

Spacing and Testing Effect

Declarative fact material, like the vocabulary in this study, have to be memorized by learners and will then be tested later. The systems, which help learners accomplish this are called adaptive fact learning systems. Research has shown that two effects can be used to enhance retention (Pavlik and Anderson, 2008). Those are the spacing (Greene, 1989) and the testing effect (Roediger III and Karpicke, 2006).

The deficient-processing subtheory is used to explain the spacing effect in cued-memory tasks. It says that a learning strategy is used in which a learner judges how well (s)he thinks a word is memorized. If a word is judged to be less well learned it will be more extensively processed. Generally, during massed learning words are judged to be better learned than during spaced learning. This is often a wrongly drawn conclusion. As a result, learners learn better when the spacing effect is used since they will more extensively process the words (Greene, 1989). Furthermore, research has shown that it is also beneficial to introduce additional items before a repetition (Green et al., 2014).

Research has shown that testing can improve one's memory for the studied material (Carpenter et al., 2008; McDaniel et al., 2007). It can even enhance retention more than additional study time of the material (Roediger III and Karpicke, 2006). More specifically, successful retrieval of studied material improves word retention (Carrier and Pashler, 1992). If retrieval is not successful the testing effect practically disappears for those items. This even holds if no feedback is given during the testing session (Jang et al., 2012). It has been shown that retrieval practice is important for vocabulary retention. In one study researchers tested whether repeated vocabulary retrieval affected vocabulary retention once a word had been retrieved successfully. It showed that repeated retrieval during tests improved vocabulary retention more than additional encoding or studying once an item has been successfully reproduced (Karpicke and Roediger, 2008).

Tutoring systems can become much more efficient if both of those effects are optimized and adjusted to the individual learner. The spacing of vocabulary depends on the number of words to be studied and on how much time is available before the test (Sense et al., 2016). This means one has to balance those effects. The goal is to maximize the distance between repetitions while still making sure that the word can be successfully retrieved.

Flashcard method

The flashcard method was used to determine the order in which the words would be shown. In the flashcard method, the word list is divided into small subsets. In this case, the word list of 25 words was divided into five sets consisting of five words each. A set is shown to the participants, one word at a time. Afterward, all the words that were answered incorrectly are shown to the participant again until they answer them correctly. This procedure is repeated with the other four sets. After every word was answered correctly, the word order within each set will be randomized and the first set is shown to the participant again. The first time a word is shown its corresponding translation is displayed as well. This set-up is similar to the one described in Van Rijn et al. (2009).

Creating the app

An iOS app on an iPad Air (iOS 14.5) was used for testing. For the writing part of the app, an Apple Pencil 2^{nd} generation was used. For the typing part of the app, the integrated keyboard of the iPad was used. The app was developed in Xcode 12.5. It presents as a basic vocabulary learning app by showing the Swahili word to the user and after the user inputs his or her answer feedback is given to the user and the user can click next to see the next word. Feedback is given by coloring the user input either green or red depending on whether the answer was correct and displaying the correct answer (Figure 1).

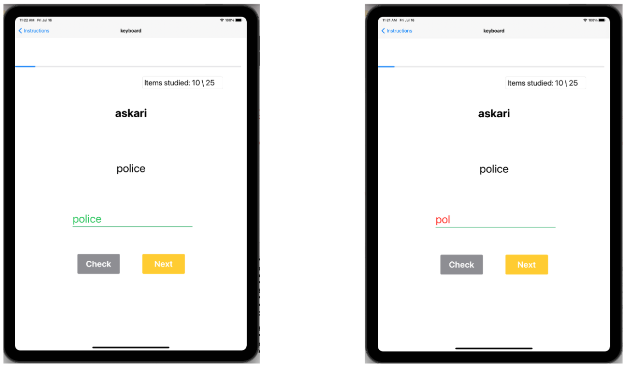


Figure 1

The figure shows how the display looks like once an answer has been given by the user. On the left side if the answer was correct and on the right side if the answer given was wrong.

At the end of the session, the results are sent via email to the researcher and saved on the device itself. Furthermore, the resulting array can be copied from Xcode on the display. This way it was made sure that data could be saved even if exporting the .csv file failed.

For the writing portion of the app, Scribble was used. Scribble allows a user to write the answer with an apple pencil in a text field. It will then automatically convert the handwritten text to standard text. It is also possible to delete words or letters with a smart pencil or insert text the same way you would do on a piece of paper. Scribble interactions allow the developer to individualize the experience of scribble. It makes it possible for the user to write their answer and then have the app check whether the answer is correct. Also, the API *UIScribbleInteraction* lets the developer track when a user starts writing. This allows the recording of reaction times. Furthermore, this API lets the developer disable auto-complete for handwriting, which means one can be sure the user knew the whole word. All the functions just described were included in the app. For typing text, auto-complete can be disabled in the storyboard (Apple Inc., n.d.). The order of the words shown was determined by the previously described flashcard method.

First pilot study

Experimental Setup

Participants

Five participants took part in the first pilot study. All five participants were German with a mean age of 37.5 ($range_{age}$ = [17;26]). No participant had any prior knowledge of Swahili. All participants fulfilled all requirements. Therefore, no participant had to be excluded from the study.

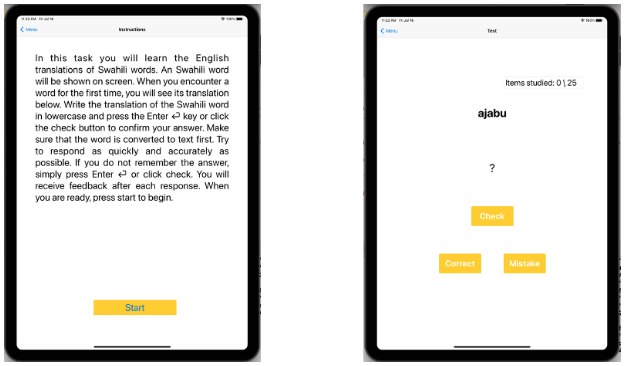
Tools and Technologies

For the 1st session an iPad air (iOS 14.5), an apple MacBook and an Apple Pencil 2nd generation were used. Two lists with 25 words each were used for the study. Which list was shown in what condition depended on the participants' identification number. For the second session, a laptop with a microphone was used.

Procedure

The procedure was based on the language condition of the research by Sense et al., 2016. Each participant took part in two sessions. The two sessions took part on two consecutive days. The session on the first day consisted of two blocks. Each block consisted of a twenty-minute study session, followed by a five-minute distraction task and a test of the studied declarative fact material. The test took about five minutes. The distraction task consisted of the puzzle game Tetris, which could be played until the researcher indicated that five minutes were over. This means one block took approximately 30 minutes. In one study session participants had to write the answer with a smart pen and in the other block, they had to type the answer.

Before the start of the study session, participants asked to read the instructions of the task (Figure 2). During the study session, the Swahili words were shown to the participant on the tablet and they had to write the corresponding English translation. If a word was shown for the first time the English translation was shown as well and had to be copied. Feedback for the answer was provided in both cases. The participant could decide when to show the next word.



In this task you will learn the English translations of Swahili words. An Swahili word will be shown on screen. When you encounter a word for the first time, you will see its translation below. Write the translation of the Swahili word in lowercase and press the Enter ↩ key or click the check button to confirm your answer. Make sure that the word is converted to text first. Try to respond as quickly and accurately as possible. If you do not remember the answer, simply press Enter ↩ or click check. You will receive feedback after each response. When you are ready, press start to begin.

Figure 2

Shows how the interface looks like for the instructions (left image) and the testing (right image).

The test at the end of each block was also conducted in the vocabulary learning app. The Swahili word was shown on the screen and the participant had to let the researcher know the correct English translation orally (Figure 2). After checking whether the correct reply was given the next word was shown. Every word was asked once. The reason for making the testing session orally is that either typing or writing the answer could possibly confound the results by improving the word recall in the condition which uses the same type of input.

The second session was done via google meet and lasted around ten minutes. Participants were shown the Swahili word on the screen and had to let the researcher know the English translation. The researcher then noted down whether the response was correct.

Results

The results show that the app was working. However, participants needed some time to figure out how the pen worked and it could be seen that participants who have used a pen before would have an advantage. Therefore,it was decided to add a practice session for using the pen before the start of the experiment. Furthermore, the results showed that the flashcard method was not implemented as intended and all words were shown in a random order, which could lead to wrong results. That way it could happen that some words were not shown at all to participants or that a word was repeated consecutively. It was concluded that the flashcard method needed to be improved before the next pilot study.

Second pilot study

Experimental Setup

In the following part, the differences to the previous pilot study will be described. The flashcard method was implemented and a scribble practice session was added.

Participants

In this pilot study, six participants took part. They all were German and had no prior knowledge of Swahili. One participant had to be excluded from the analysis due to technical difficulties while doing the study.

Tools and Technologies

The scribble practice session was conducted using the build-in tutorial of the iPad and the app pages.

Procedure

During the scribble tutorial the functions of the pen (writing, insert, delete, join, select) were demonstrated and participants could try them out. Afterward, the app was opened and participants had to write their name and delete it to get a feeling for the function of the pen. Afterward, they had to write ten words (cat, house, mountain, word, food, garage, movie, cinema, fish, weight) and check whether they were converted correctly. The words included all letters, that were used during the study. This way participants could make sure that scribble recognizes all of their letters. After the practice session, the actual session in the vocabulary learning app was started, in which the previously explained flashcard method was implemented.

Results

The feedback by the participants showed that they liked having an introduction to scribble at the beginning and also the data at the beginning of the experiment seemed to indicate that. I also did not receive any questions about the pen anymore after the start of the study session. On the other hand, the implementation of the flashcard method was not completely successful. Instead of only repeating the words that were answered wrongly the whole set was repeated until all words in the set were answered correctly.

Study

Experimental Setup

Before the study started the flashcard method was implemented correctly. This means that only the wrong answers were repeated after a set finishes and not the whole set. Also, a questionnaire at the beginning was added

Exclusion Criteria

Three exclusion criteria were defined to make sure that complete data sets could be analysed and that participants were actively engaged in the task. If the data of a participant fit one of the exclusion criteria, the complete data set from that participant was removed from the analysis. First, participants had to complete both sessions, because otherwise the data set would not be complete. Second, participants had to be shown all 50 words during the training session at least once, this made sure that the data collection worked and participants understood the task. Third, participants needed to have at least 25% of items correct during the delayed recall task. This is a common criteria to use to make sure that participants were actively engaged in the research. One more aspect of data I looked at was how many times a participant wrote the answer within a trial. Trials, in which participants corrected their writing by deleting the whole answer were excluded from the analysis when looking at the reaction times, because the recorded reaction time did not reflect the reaction time, associated witht he word. Often participants had to rewrite an answer due to wrong conversions of scribble. This way it is unlikely that data sets were included in the analysis that could potentially distort the data.

Participants

In total 24 participants took part in the study. One participant was excluded from the study, because (s)he did not see all words. Three more participants were excluded, because they scored less than 25% on the delayed post-test. The remaining 20 participants had an average age of 23.05 years ($SD = 3.97$). 70% of the participants were female. They were all familiar with using a tablet. Most participants (80%) have used some kind of smart pen before. Nearly no participants (5%) have used an Apple Pencil 2nd generation before (the pen used in this study). Participants were not familiar with the language Swahili. The participants had 16 different native languages and all had a good understanding of the English language.

Tools and Technologies

The questionnaire was filled out at the beginning of the study on the MacBook using Google Forms.

Procedure

Before the actual study started a short questionnaire was administered. In the questionnaire general demographic questions and questions concerning prior usage of tablets and smart pens were asked. This questionnaire took around two minutes to fill out. The questionnaire was added to have basic information about the participant and to be able to judge their experience with such devices.

Results

The main analysis is based on two statistical analysis. In the first one the reaction times of participants over time were compared to assess whether a difference in reaction times between the conditions could be detected. For that the average reaction time per minute was calculated for each condition. In a second analysis it was assessed whether participants gave the correct answers during the session and also afterwards during the vocabulary tests to see whether using one of the modes of input (keyboard or smart pen) lead to better vocabulary retention during the tests. All analysis was conducted using Rstudio version 1.4.1717 (RStudio Team, 2021). The assumption of normalcy was tested in each part of the analysis using the Shapiro-Wilk test (Shapiro and Wilk, 1965), which came back significant. Since n $<$ 30 and normalcy could not be assumed a t-test could not be conducted. Non-parametric tests were used, because they do not assume any kind of distribution. However, they are more prone to type two errors. Wilcoxon tests are one kind of non-parametric test, which yield similar results to t-tests and can therefore be used in place of them if normally cannot be assumed (Navarro, 2018). In the following analysis either a Wilcoxon rank sum test or a Wilcoxon signed rank test was conducted depending on whether the samples were independent or paired.

Results Training session

Onset. First, I looked at the reactions times in both conditions. For that I substracted the time at which the participants were shown the word from the time they started typing.The overall mean indicates that the handwriting condition (M = 2.86, SD = 2.99) seems to have longer reaction time compared to the keyboard condition (M = .98, SD = 1.32). This means that on average it took participants longer in the handwriting condition to answer a cue than in the keyboard condition. Afterwards, a more detailed analysis was conducted. I was interested to see how the reaction times changed throughout the 20min block and if a difference between the two conditions could be observed. For that the mean reaction time per condition for each minute was calculated and the corresponding standard error. Afterwards a point graph was created (Figure 3), which displays the calculated means and error bars.

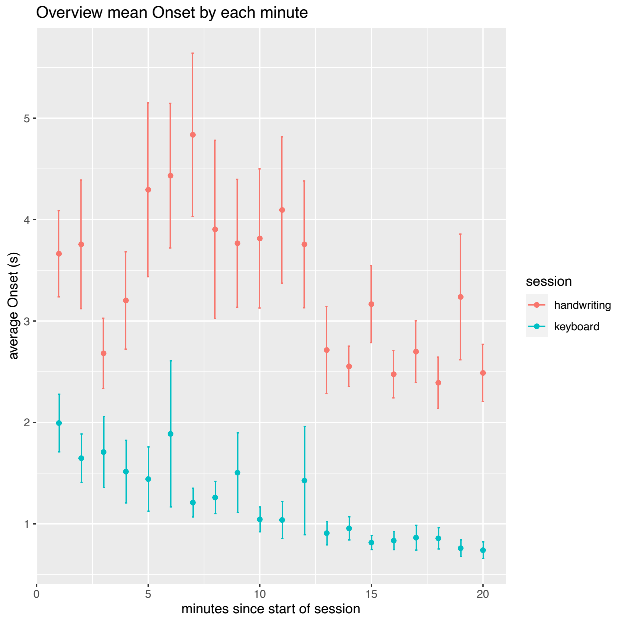


Figure 3

Overview of mean reactions times per minute with error bars. The blue points correspond to the keyboard condition and the red points to the handwriting condition. On the y-axis the average reaction times can be seen and on the x-axis the time that has passed since the start of the block.

In figure 3 it can be seen that the reaction times of the two conditions show differences. First thing to notice is that the reaction times in the handwriting condition are consistently slower than in the keyboard condition. Participants in the keyboard condition seem to, on average, get faster consistently throughout the condition. Opposed to this in the handwriting condition the reaction times of participants slow down a bit at first. Afterwards the reaction times get faster until they become more stable towards the end of the session. It can be said that it took participants longer to start answering in the handwriting condition than in the keyboard condition, hence they thought longer about their answers before replying. They also seem to variate more in the time they think before replying in the handwriting condition. This also explains why less trials were conducted in the handwriting (M = 110.4, SD = 31.45992) condition compared to the keyboard condition (M = 207.2, SD = 65.54677).

Accuracy scores. First the accuracy scores were visually inspected to see the distribution of the accuracy scores. Figure 4 shows violin plots of the accuracy scores (Hintze and Nelson, 1998). The violin plots depict the distribution of the accuracy scores of the two blocks during the training session. The black horizontal line depicts the mean accuracy and the white box with the black whiskers depicts the standard bloxplot. Overall it seems that participants scored rather well on the studied material, which means that they were shown the next word in a sufficient time frame. It can be seen that participants differed in their accuracy scores more during the handwriting condition compared to the keyboard condition. However, the mean accuracy's in the handwriting (M = .80, SD = .13) and the keyboard (M = .83, SD = .11) condition seem to be rather similar.

Next, the Wilcoxon rank sum test with continuity correction was conducted. At .05 significance level, it can be conclude that the accuracy scores of the handwriting and keyboard condition cannot be identified as nonidentical populations (p = .35, W = 165). This means that no statistical difference can be detected between the two conditions and participants seem to retain similar numbers of words in both conditions.

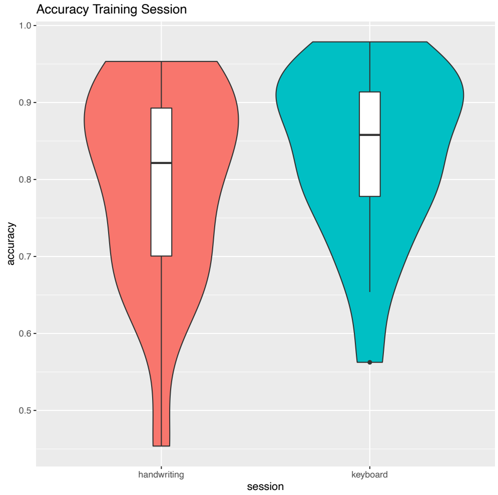


Figure 4

Overview of accuracy scores during the training session. On the left the mean accuracy scores of the handwriting condition can be seen (in red) and on the right the mean accuracy scores of the keyboard condition can be seen (in blue).

Vocabulary tests

Tests right after the corresponding training sessions. After calculating accuracy scores for the tests on the first day the scores are visualized in violin plots to better compare the two conditions (Figure 5). It can be seen that the two plots differ a lot. The accuracy scores in the handwriting condition seem to have a way wider distribution compared to the keyboard condition. It can also be seen that in the keyboard condition a lot of participants reached a ceiling effect by answering all Swahili words with the correct English translation. This means that they had very good word retention. Only few participants in the handwriting condition seemed to get all words correct. It can also be seen that the mean of the handwriting (M = .87, SD = .14) and the keyboard condition (M = .94, SD = .08) differ.

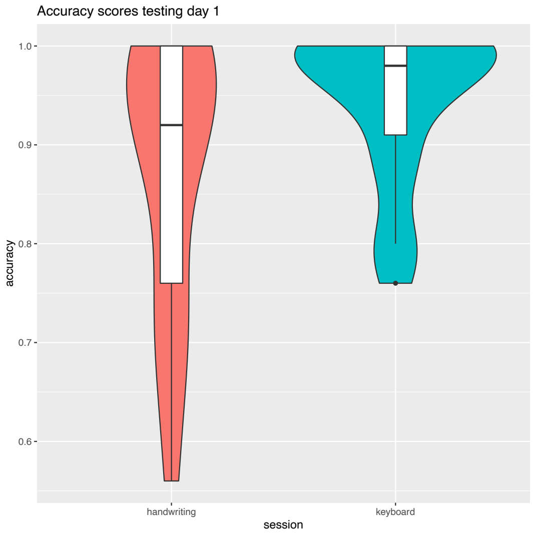


Figure 5

Violin plots of accuracy scores from the tests conducted right after the training sessions. The test from the handwriting condition can be seen on the left (red plot) and the test from the keyboard condition on the right side (blue plot).

To statistically analyze the results a Wilcoxon rank sum test with continuity correction was conducted. Using a significance level of 0.05 it can be concluded that the mean accuracy from the keyboard mean accuracy is not significantly different from the handwriting's mean accuracy with a p = .084 (W = 138). This means that from a statistical point of view participants retained a similar amount of words in both conditions.

Test on the next day.

After calculating accuracy scores for the follow-up session the scores were visualized in a violin plot (Figure 6). The first thing one notices when looking at the plot is that the two plots look different. First, participants seem to have scored better in the keyboard condition than in the handwriting condition. The mean of the keyboard condition (M = .89, SD = .16) is higher than the one of the handwriting condition (M = .82, SD = .16). This can also be seen by looking at the two boxplots and at the overall shape of the plots. Most of the accuracy scores of participants in the keyboard condition are above 0.80, while this is not the case in the handwriting condition. This means that a lot more participants in the keyboard condition reached a high accuracy score than in the handwriting condition. More participants in the keyboard condition reached 100% on the test. This can be seen by the wide blue part at the top of the plot. It can be concluded that in the keyboard condition more participants experienced a ceiling effect on the test. This tells one that they were exposed enough to the material in order to retain the Swahili-English word pairs.

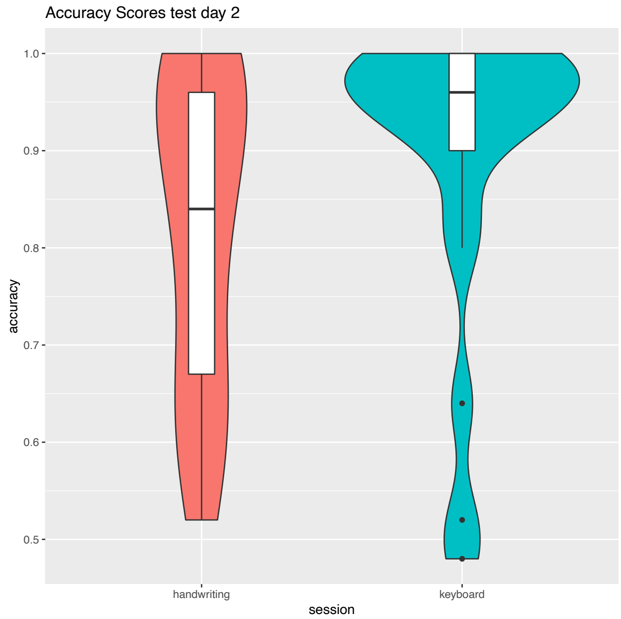


Figure 6

Violin plot of the accuracy scores of day two. On the left side the accuracy scores of the handwriting condition are displayed (red plot) and on the right the accuracy scores of the keyboard condition are displayed (blue plot).

Next I looked at the results of the statistical analysis. By using a significance level of 0.05 it can be concluded that there is no significant difference between the two conditions ($p = .15$, $W = 147.5$). This means though a difference can be seen in the graphs this difference is not significant and overall participants seem to perform similar in both conditions. This means that no difference in word retention between the two conditions can be observed.

Comparison Accuracy scores keyboard sessions on day 1 and day 2. Next, the two test sessions for the keyboard condition were compared. For that the accuracy scores were plotted again in a violin plot (Figure 7). The overall mean accuracy of the test on day one ($M = .94$, $SD = .08$) is different to the mean accuracy of day two ($M = .89$, $SD = 16$). The mean on day two is lower. This can also be seen in the violin plots. Overall it can be observed that participants scored rather well on both tests with an accuracy score of over

80%. However, some participants seemed to have forgotten some words when comparing the first session to the second session, since less participants scored 100% and some participants scored below 70%. Overall it can be said that participants seem to forget some words over night but they still remember over half of the word-pairs.

Since, the retention over time of the same participants are checked the groups can no longer be considered independent and a paired Wilcoxon signed rank test with continuity correction was used. When using a significance level of 0.05 it can be concluded that the accuracy on day one is significantly different from the accuracy on day two with p = .011 (V = 71.5). This means that participants performed different on the test on day one than on the test on day two.

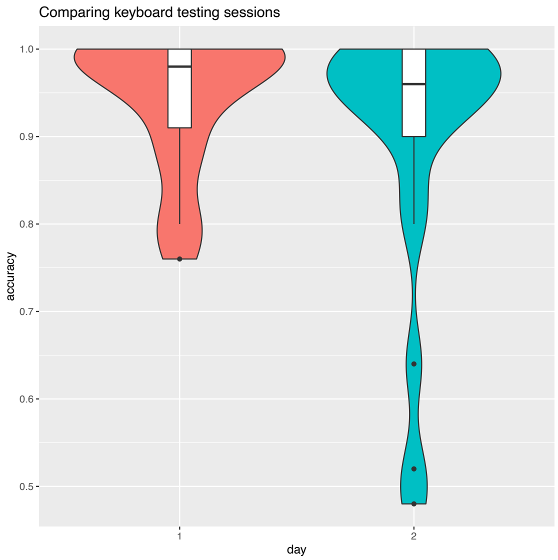


Figure 7

In the plot the comparison between the accuracy scores of the two test sessions can be seen. On the left the accuracy of the test right after the keyboard training session can be seen (red plot) and on the right the scores of the next day (blue plot) can be seen.

Comparison of accuracy scores writing conditions on day 1 and day 2. For the last comparison the mean accuracy scores of the two tests in the writing condition were plotted in a violin plot (Figure 8). In the figure it can be observed that the overall shape of the two plots seem to be fairly similar. The accuracy scores in the second session seem to be a bit lower than in the first session. This can especially be seen in two places of the graph figure. Though a ceiling effect can be observed in both conditions it is more pronounced in the first session, since more participants scored 100% on the test. Furthermore, some participants seem to score lower on day two than any participant does on day one. The mean of the accuracy scores of day one ($M = .87$, $SD = .14$) are a bit higher than on day two ($M = .82$, $SD = .16$). The boxplot also shows where most of the scores lie, which is a bit lower on day two than on day one. All of those observations together lead to the conclusions that participants seemed to score a bit better on day one than on day two.

Next, a paired Wilcoxon signed rank test with continuity correction was conducted. When using a 0.05% confidence interval the null hypothesis cannot be rejected. it can be concluded that the accuracy of day one is not significantly different from the accuracy on day two with a $p = .10$ ($V = 100.5$). This means participants most likely did not perform different on the tests on day one and day two, which means that they retained the words as well for the first test as for the second test.

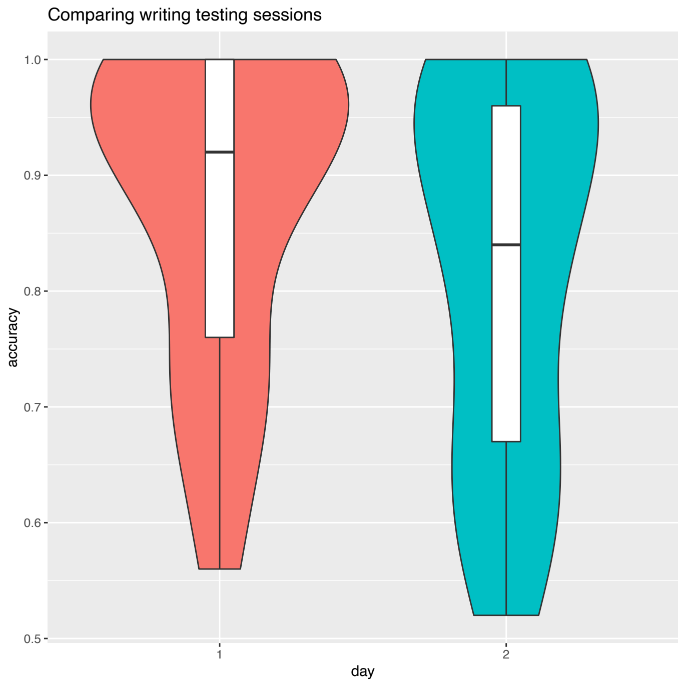


Figure 8

On the left side of the of the plot (in red) the mean accuracy scores of the test right after the training session can be seen. On the right (in blue) the mean accuracy scores of the test the next day can be seen.

Qualitative observations

After the experiment some participants gave informal feedback. They all agreed in their feedback. They told me that it was fun to learn vocabulary with the app. I got the feedback that participants preferred the keyboard condition over the writing condition. The main reason for that was the word recognition program scribble, which did not always convert the words correctly. It sometimes took multiple tries to actually convert the word and one had to write very precise to convert the word correctly. Participants told me that this was very frustrating. This is reflected in the number of trials participants had in the writing and the keyboard conditions. Most participants had more trials in the keyboard condition than in the writing condition. It is also reflected in the longer reaction times in the writing condition. What is interesting here is that though participants preferred the keyboard condition and had more trials in the keyboard condition they did not significantly score better in one of the two conditions. In fact during the statistical analysis no difference between the two conditions could be observed other than comparing the corresponding sessions of day one and day two. When comparing the differences in word retention between testing on day 1 and day 2 participants seemed to have a difference in vocabulary retention in the keyboard condition but not in the writing condition. This leads one to conclude that maybe in the writing condition less trials are necessary to reach the same level of word retention as in the keyboard condition and that maybe by having to be more careful with writing the answer correctly they actually retain the words better the next day.

Discussion

In this study, it was investigated whether the input device (smartpen or keyboard) has an influence on word retention in an online fact learning task. The two research questions that were investigated are: "Can similar retrieval tendency be obtained by pen input as with keyboard input?" and "Does writing the answers instead of typing the answers improve word retention?". Knowing whether using the Pencil improves word retention helps to determine whether it is useful to implement the option to use a Pencil in future vocabulary learning apps.

The results of the analysis show that a smartpen can be used for input in a vocabulary learning app. This is supported by the reaction times, which were recorded in the app. However, the reaction times in the writing condition were a lot slower than in the keyboard condition. This means that participants needed longer to answer using the smartpen compared to using the keyboard. This can also be seen in the number of repetitions. Participants in the keyboard condition had a lot more repetitions compared to participants in the writing condition. It can be concluded that the current speed in the writing condition is too slow, since it is beneficial to have a lot of repetitions.

The difference in reaction times could be due to the experience users have with using a smartpen. Most users were familiar with a tablet and typing on a tablet. However, most of them had only used a smartpen a few times and only one person had used an Apple Pencil before. Therefore, participants may have had to focus more on the usage of the Pencil instead of focusing on writing the answer. Also, the system gets used to the writing style of the user over time and converts the word correctly more often. It was not possible for the system to get used to the writing style of the user due to the short time frame in which the study took place. Another limitation, which needs to be considered is that the technology may not be good enough yet. The technology is still in development therefore the Pencil and the corresponding software may substantially improve and will be easier to use in the future. Handwriting recognition was introcuded by Apple in 2020 with the introduction of iOS 14 and is continuously improved. Therefore, it would be interesting to replicate the study with participants familiar with the Apple Pencil and once Apple develops the software further. This may lead to a smaller difference in reaction times between the two conditions. However, it can also be hypothesized that the difference will never completely disappear due to the differences in writing and typing. Research shows that writing on paper is also slower than typing on a keyboard (Mangen and Velay, 2010). Since differences in brain activation can be seen comparing writing and typing it would interesting to see whether any differences can be observed comparing those with writing with a smartpen. It could be the case that the benefit of the cognitive effort necessary to write is outweight by the time necessary to get used to an advanced writing system such as the Pencil.

Furthermore, this app was developed specifically for the purpose of this study with only a basic vocabulary learning system. The results may also differ when using a more advanced vocabulary system like in Sense et al. (2016). Research has shown that the learning system used influences the learning outcome (Sense et al., 2016). In the future the study should be replicated using a more advanced app and the newest writing software available.

For answering the second research question one had to compare the accuracy scores in the two conditions. The results show no improvement in word retention using the smartpen compared to using the keyboard. Therefore, the hypothesis that using the smartpen for writing will improve retention could not be supported by the results. According to the results, it is possible to use the Pencil but using the keyboard is just as effective. One may need more or more difficult words for a future study, because a ceiling effect happened for a lot of the participants. This way no difference between the two conditions could be detected. It would also be interesting to see if this result is the same for different languages. One study has already shown that French accents are better retained when using the keyboard compared to writing them, which is most likely due to the way accents must be indicated on the keyboard (Van Hove et al., 2017). Another limitation of the study is the number of participants. Since only twenty participants could be considered for the analysis, the statistical power is low and in the future the study should be replicated using more participants.

In conclusion it can be said that writing recognition works with a vocabulary learning app. However, the word recognition software needs to be as good as the user themselves at recognizing letters in order to see a possible benefit of including writing in a vocabulary learning app. With the results of the current study no decision can be made on whether writing improves word retention compared to typing, but one can conclude that at the current level writing with a smartpen is not smooth enough to replace the keyboard, since writing with the Pencil takes too long. This means that the technology needs to be improved or the writer has to adapt his or her writing style. This will take a lot of time since users due not just change their handwriting and technology needs time to advance.

References

Apple Inc. (n.d.). Meet scribble for ipad - wwdc20 - videos. https://developer.apple.com/videos/play/wwdc2020/10106/

Bangert, A. S., & Heydarian, N. M. (2017). Recall and response time norms for english–swahili word pairs and facts about kenya. *Behavior research methods*, *49*(1), 124–171.

Bui, D. C., Myerson, J., & Hale, S. (2013). Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology*, *105*(2), 299.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448.

Carrier, M., & Pashler, H. (1992). The influence of retrievall on retention. *Memory & Cognition*, *20*(6), 633–642.

Cunningham, A. E., & Stanovich, K. E. (1990). Early spelling acquisition: Writing beats the computer. *Journal of Educational Psychology*, *82*(1), 159.

Encyclopædia Britannica inc. (2014). Swahili language | African language. Retrieved July 17, 2021, from https://www.britannica.com/topic/Swahili-language

Green, J. L., Weston, T., Wiseheart, M., & Rosenbaum, R. S. (2014). Long-term spacing effect benefits in developmental amnesia: Case experiments in rehabilitation. *Neuropsychology*, *28*(5), 685.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 371.

Higashiyama, Y., Takeda, K., Someya, Y., Kuroiwa, Y., & Tanaka, F. (2015). The neural basis of typewriting: A functional mri study. *PLOS ONE*, *10*(7), 1–20. https://doi.org/10.1371/journal.pone.0134131

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184.

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology*, *65*(5), 962–975.

Jasanoff, J. H., & Cowgill, W. (2020). Indo-european languages. https://www.britannica.com/topic/Indo-European-languages

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *science*, *319*(5865), 966–968.

Longcamp, M., Boucard, C., Gilhodes, J.-C., Anton, J.-L., Roth, M., Nazarian, B., & Velay, J.-L. (2008). Learning through hand- or typewriting influences visual recognition of new graphic shapes: Behavioral and functional imaging evidence. *Journal of Cognitive Neuroscience*, *20*(5), 802–815. https://doi.org/10.1162/jocn.2008.20504

Mangen, A., & Velay, j.-l. (2010). Digitizing literacy: Reflections on the haptics of writing. https://doi.org/10.5772/8710

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic bulletin & review*, *14*(2), 200–206.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking [PMID: 24760141]. *Psychological Science*, *25*(6), 1159–1168. https://doi.org/10.1177/0956797614524581

Navarro, D. (2018). *Learning statistics with r: A tutorial for psychology students and other beginners: (version 0.6)*. compcogscisydney. https://learningstatisticswithr.com/

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of swahili-english translation equivalents. *Memory*, *2*(3), 325–335.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101.

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, *1*(3), 181–210.

RStudio Team. (2021). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

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

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Smoker, T. J., Murphy, C. E., & Rockwell, A. K. (2009). Comparing memory for handwriting versus typing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*, 1744–1747.

van den Broek, G. S., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812.

Van Hove, S., Vanderhoven, E., & Cornillie, F. (2017). The tablet for second language vocabulary learning: Keyboard, stylus or multiple choice. *Comunicar*, *25*. https://doi.org/10.3916/C50-2017-05

Van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. *Proceedings of the 9th International Conference of Cognitive Modeling*, *2*(1), 7–6.

Vaughn, S., Schumm, J. S., & Gordon, J. (1992). Early spelling acquisition: Does writing really beat the computer? *Learning Disability Quarterly*, *15*(3), 223–228.