Master thesis on Cognitive Systems and Interactive Media

Universitat Pompeu Fabra

Effects of distributed learning patterns on elementary student learning of computational thinking

Lydia Casanova

Supervisors: Patricia Santos, Marc Beardsley

Host Research Group: TIDE

July 2020

Contents

Acknowledgement

Firstly, I would like to thank Marc Beardsley and Patricia Santos for helping me to create this project. They have been excellent tutors, finding the time to meet me and review my steps even on the difficult times. I know these have been strange months which force us to find new ways to work, but I could not be more satisfied with your tutoring. I would also like to thank Davinia Hernández-Leo for trusting me and talking to me about this project, I really enjoyed it. Thanks to all of them I felt part of TIDE research group and I maintained my motivation high.

I also appreciate that I was allowed to run this project within the Makers a les Aules program - which is co-financed by Barcelona Activa, the UPF-ETIC Maria de Maeztu program and TIDE -. I am also thankful to the participating schools educators and students who help me to collect the necessary data for this project.

Finally I would like to thank my family, friends and boyfriend to emotional support me during this project. Without your patience and optimism this project would not be possible.

Abstract

Applied cognitive psychology has been a center topic for a number of research works in last decades. In particular, there are studies conducted on memory, cognition and on the science of learning. The latter are meant to find new methods in order to improve the process of learning. The use of the Retrieval Practice (RP) and Distributed Learning (DL) has been proved to be improving the e□ciency of learning when compared to Massive Learning (ML) practices (which is the most used method in formal education). Even though there are studies which proved the advantages of using DL, di□erent patterns can be found and it is not clear which one is the best to apply to get the best students' results.

This study was conducted in the last 3 courses of primary education and the learning topic was Computational Thinking (CT). CT is one of the 21st century skills that gained more attention and it is progressively being incorporated in formal learning since the early stages of education. CT refers to understanding how to develop step-by-step solutions of problems, helping students to use and improve logical thinking, pattern recognition and decomposition skills. New generations are likely to live in a technologically integrated society, hence, CT might become an essential skill that will enable them to understand and manipulate the technology that surrounds them.

Thus, the main goal of this project is to find out how the learning process is a□ected by di□erent patterns of distributed practice and discover which is the distribution that leads to a better performance in the latter, as well as being suitable to apply in formal learning contexts. In order to do that, two di□erent patterns of DL and a ML group (as a control group) will be compared with the aim to find which one of the two reaches the best performance using RP as a constant in all groups. As a result, I expect that DL groups will perform better than the ML groups, and I forsee to find out whether there is a significant di□erence in performance between the two DL patterns tested.

Keywords: Distributed Learning; Computational Thinking; Massive Learning; Retrieval Practice; Formal Education.

Chapter 1

Introduction

This project aims to figure out which Distributed Learning (DL) pattern leads to better results for a long-term memory retention. The learning topic will be Computational Thinking (CT) adapted for primary school students. The experiment consists in analysing two groups that will be learning through two different DL patterns and one control group that will not be using DL method[1]. In all three groups there will be used Retrieval Practice (RP) as a constant.

In order to explain each piece of this experimental design, the state of the art covers 3 topics. Firstly, it will be explained what DL is and its basis. Secondly, there will be the definition of what RP is and why it is used as a constant in all experimental groups. Thirdly, it will be explained what CT is, and why it has been chosen as the learning topic.

1.1 Problem Statement

The traditional teaching method used in formal education is based in the Massive Learning (ML) approach: presenting the information in large periods of time but not revisiting it. Several researches have shown that DL is more effective for long-term memory formation than ML. This effect is known as "the spacing effect" or

[1]Due the Covid-19 situation, the control group could not be part of the experimental design.

“distributed practice e□ect” [1]). The amount of studies testing the spacing e□ect cover a large variety of learning topics such as vocabulary learning [2] [3], syntax learning [4], text learning [5], mathematics [6] and biology [7].

Even though research studies have proved the e□ciency of DL, it is not clear which is the pattern that should be used to revisit the information. On one hand, we can find the distributed pattern used in the study made by Fishman [3], in which the information was retrieved every two days in a period of 6 days, which means retrieving each piece of information a total amount of 3 times. On the other hand, we can find the study by Rawson [5] where students reread a text only one week after the first reading, which means reading the text a total amount of 2 times. The frequency could be the key issue in the guidance of teachers on how to apply this practice in their lessons. For example, too high frequency (e.g. every 2 days) might be counterproductive, as the weeks pass, the amount of material to review increases in a linear manner, which eventually could result in just simple review of the past information sessions. At the other extreme, too low frequency (e.g. each month) could result in students already forgetting given information.

The first step to introduce DL in formal education must be very clear in settling its basis, and giving teachers the proper parameters to optimize the performance of students in the learning process. In order to do that, it is needed to compare di□erent patterns of DL to find out which is the most suitable pattern for revisiting the learning material.

1.2 State of the Art

1.2.1 Distributed Learning (DL)

Distributed Learning (DL) is one of the research topics that has been most used for studying memory e□ects in cognitive psychology. In order to understand this concept we need to go back to 1885.

At 1885 Hermann Ebbinghaus ran his first studies on the advantages of distributed

practice over massive practice [8]. In his experiment, he used himself as the only participant. Reviewing the results he indicated: “It makes the assumption probable that with any considerable number of repetitions a suitable distribution of them over a space of time is decidedly more advantageous than the massing of them at a single time". His findings were the basis for successive research on memory and learning in di□erent domains in which the beneficial e□ect for long-term retention have been demonstrated [2] [3] [4] .

In order to run a DL experimental design there are two parameters that need to be taken into account: the Interestudy Interval (ISI) and the Retention Interval (RI). The ISI is the time between the first time a topic is introduced and the next time that this topic is revisited. The RI is measured from the last learning session to the final session. Therefore, in DL ISI > 0 while in ML ISI = 0 [9].

The e□ciency of the “Distributed Practice e□ect” or the “spacing e□ect” [1] can be explained mainly by the forgetting curve in Figure 1 [8] . This curve indicates the decrease of the brain’s ability to retain memory over time. This theory defends that human loose the memory of learned knowledge over time unless it is reviewed several times. The more times the learning is reviewed, the flatter the forgetting curve is. In the case of ML, the content is provided just one time so the forgetting curve of that knowledge could be represented by Figure 1. In case of DL, the information is reviewed and therefore the forgetting curve is modified with each review, as it can be seen in Figure 2.

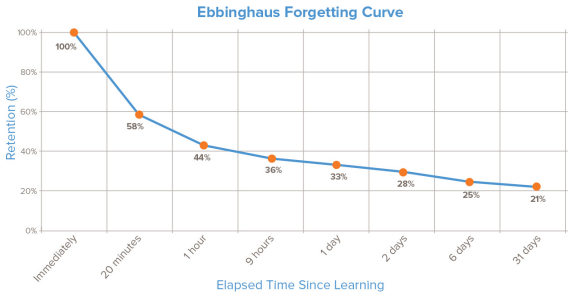

Figure 1: Forgetting curve by Ebbinghaus

Even though DL have been reported to improve LTM formation, we need to consider

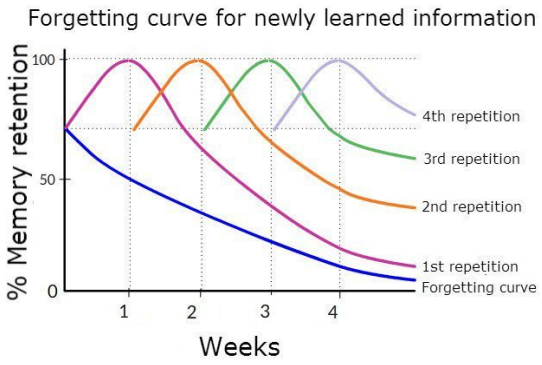

Figure 2: Forgetting curve by Ebbinghaus applying review

that it seems to lead to worse performance on immediate tests. This phenomena has been observed on the study by Rawson & Kintsch [5]. During their study, they were testing the rereading e□ects. According to their results on an immediate test performance was greater after massed versus distributed rereading. On a delayed test, performance was greater after distributed versus massed rereading. Results can be observed on Figure 3. Therefore, even though DL can lead to best long term results it can lead to worse short term performance.

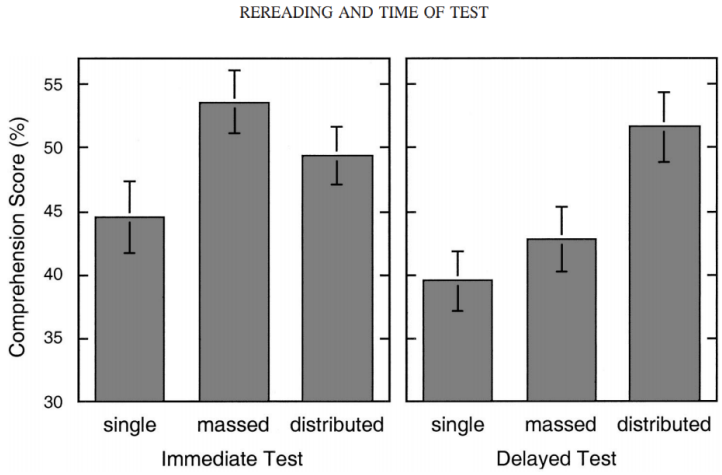

Figure 3: Results by Rawson & Kintsch on immediate and delayed tests.

There have been studies related to optimizing the distributed practice. On the study of Optimizing Distributed Practice: Theoretical Analysis and Practical Implications[10] it has been reported the importance of the gap between the retrieval sessions in order to optimize the LTM formation. Specifically, an optimal gap improved final recall by up to 150 per cent. According to these results, the gap increment caused

test accuracy to initially sharply increase and then gradually decline. The data presented on the mentioned study suggested that very substantial temporal gaps between learning sessions should be introduced-gaps on order of months, rather than days or weeks. The authors of the mentioned study concluded that "If these findings generalize to a classroom settings they suggest that a considerable redesign of a conventional instructional practices may be in order." Even so, it might be difficult for a teacher to apply these changes on their everyday practice since they would need to redesign their conventional practices. Therefore, the optimal time gap between reviews might be balanced with actual teaching practices. In other studies, DL is related with RP and the conclusions are quite different. On the study of Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention [11] is reported that according to recent findings "expanding retrieval practice may be inferior to uniform-interval retrieval practice when memory is tested after a long retention interval". The results of this study suggest that the extent, to which learners benefit from expanding RP, depends on the degree to which the to-be learned information is vulnerable to be forgotten. Therefore, according to these results the most suitable DL pattern might include uniform and enough spaced gaps between reviews. The accuracy of the retrieval depends on the provided information.

Despite the benefits that DL has reported in different studies, there still some challenges on its application in formal education. On the article by Son [12], he describes some of the challenges that exist for the practitioners and learners of DL. These challenges include the lack of awareness of the benefits of spacing, and the undesirable difficulties of spacing here and now. According to the author, "these challenges exists especially for young children, who, nevertheless, are required to spend significant amounts of time studying on their own outside the classroom, and unfortunately, may be studying in ineffective and inefficient ways".

From practitioners part, Son remarks that the first step should be to continue to make a conscientious effort to connect with practitioners regarding the spacing literature. In this context, there is the IlluminatED project by the Research Group of

Interactive and Distributed Technologies for Education (TIDE) of Universitat Pompeu Fabra [13].According to its official website, "IlluminatED brings together experts in the education technology, teacher development and cognitive neuroscience to support teachers in designing more effective lessons by making use of knowledge from neuroscience". This project proposes to create a bridge between teaching practice and cognitive neuroscience to inform teachers about proven practices that promote learning for all students. These practices include the spacing effect and DL, among others.

A second challenge that Son mentioned on his research is the vast amount of information that needs to be covered and the lack of time in a course. According to Son, because these factors "it is reasonable to assume that spacing strategies may not be consistently incorporated into the classroom". For this challenge he proposes to use a more effective strategy to implement spacing strategies such as "to focus on how the individual can think about ways in which to space their study". So he proposes, as solution, to incorporate DL on homework to cover mixed topics, not only a single one per day. Son also recommends to practitioners "to have students take explicit meta-cognitive control of their strategies and to state why they choose particular strategies." Even so, he pointed out that "The challenge for teachers, though, is to find the balance between teaching subject content and teaching meta-cognitive strategies".

There is a lack of research using DL in computer science. There are studies using DL for vocabulary learning [2] [3], syntax learning [4], text learning [5], mathematics [6]) and biology [7]). Nevertheless, I only found one research relating DL and computer science [14] which didn't reach a significant conclusion. Therefore, there are lots of unexplored possibilities on this line of research that could be investigated.

1.2.2 Retrieval Practice (RP)

Retrieval Practice (RP) is a strategy in which recalling information enhances and boosts learning. Often, students listen to a lesson and think they understood the main concepts, but struggle when trying to explain them. This phenomena can

be explained from the neuroscientific perspective: “Human memory is altered in a significant way by the act of retrieval. The retrieved information will become more retrievable in the future than it would have been without such an act of retrieval, and certain related items of information in memory may become less retrievable” [15].

There have been several studies that support the e□ciency of RP for long-term memory retention. One example is shown in the study conveyed by Roediger [16]. In his study it is denied the common thought of learning occurring during studying whereas retrieval of information on testing only serves to assess what was learned. Furthermore, he defended with his results that RP is e□ective without the need of giving feedback to the retrieval exercise, but he remarked that feedback enhances the benefits of testing.

RP can be applied in class with di□erent types of exercises, so the teacher could select which one best suits the topic or the structure of the course. The only rule is that students need to take out the information they have encoded in their minds. More often, RP is applied by making questions to students about already visited topics. Even so, there are several exercises that could be done such as mental maps, writing down all the information they can remember about a certain topic and true or false or filling the gap activities.

This is not the first project that combines RP with DL. These methods have been combined in previous studies in order to blend the benefits from both methods. DL and RP were combined in the experimental design by Pick [17] in order to evaluate the e□ciency of various DL patterns in the study of Swahili-English vocabulary word pairs using RP as the review methodology. Another example of the combination of these strategies can be found on the experiment by Gossens [18] in which DL was combined with restudy and RP as review methods.

1.2.3 Computational Thinking (CT)

According to the European Commission (EC) “Computational thinking (CT) is a shorthand for “thinking as a computer scientist”, i.e. the ability to use the concepts of computer science to formulate and solve problems” [19].

The study “An analysis of educational approaches to developing Computational Thinking (CompuThink)” designed and funded by the Joint Research Centre of the European Commission aims to provide grass root, policy initiatives and an overview of recent research findings in order to develop CT as a 21st century competence among primary and secondary students. On this context, there have been lines of research regarding the assessment of CT teaching on elementary school students like the research lead by Guanhua [20] in which is assessed fifth grade students' CT using their own developed instrument; or the framework proposed by Seiter for understanding and assessing CT in the primary grades (Grades 1 to 6) [21]. Therefore, finding ways to teach and assess CT is influencing some educational lines of research.

CT can be implemented in schools from di□erent perspectives. From a theoretical perspective, teachers can introduce the CT as a method to solve complex problems by dividing them in a set of simple problems that need to be completed in a certain order. They can introduce the importance of finding a solution by following an ordered list of steps. Furthermore, CT can be applied to several learning topics, not only to technology. In fact, CT can be applied as a method to solve mathematical problem as well as a method to organize the steps to do a project about any topic. From a practical perspective, teachers can introduce CT in various hands-on activities such as programming [22], robotics [23] or any activity that requires to solve a problem.

In this context, the TIDE group from Universitat Pompeu Fabra in collaboration with Barcelona Activa and the program María de Maeztu had created the “Makers a les aules” project. This project aims to help teachers to apply CT and Design Thinking with a hands-on approach in any topic to the classes they teach. The

program is intended for primary teachers that want to learn how to apply these methodologies in class while applying technologies as Makey Makey, Scratch or 3D printing. In fact, the different distributed patterns of this work will be tested in "Makers a les aules" sessions in different schools while using Makey Makey and Scratch.

Scratch have been used before as a practical tool to teach CT. On the study by Ruthmann [24] Scratch was used to teach CT through musical live coding in Scratch. Moreover, the study by Shin [25] studied the improvement effectiveness of CT through Scratch education. According to Shin results, software education -such as Scratch- can improve the ability of CT.

This project is not the first study conveyed in "Makers a les aules" sessions. Previous studies have already shown promising results. The study by Martinez-Moreno [26] showed that "few teachers and nearly all students had prior experience using maker tools and teachers participating in the project were willing to learn how to introduce this methodology in their classroom to innovate in their lessons". The results also showed that "students increased their interest and self-perceived efficacy in technology, and their level of autonomy doing maker activities". There is another study made in Makers a les aules sessions by Theophilou [14]. The study by Theophilou concludes that "Teaching computational thinking to children has been proven to benefit analytical and logical skills and it is therefore important to find techniques that can benefit its delivery. The technique of DL has been proven to improve the memory of students in the long term and should be promoted more in the educational system". The experimental design by Theophilou aimed to combine these two findings to create a strong case to support it, but it was not possible due to external factors. The present study pretends to be an extension of this work by Theophilou with the aim of not only to prove that DL groups perform better than the ML group but also to shed light on which DL pattern leads to a better long-term memory retention.

Chapter 2

Methods

2.1 Important Note: COVID-19 Implications

Due to the COVID-19 pandemic, the state of alarm was established in Spain. Therefore, schools were forced to close and the "Makers a les aules" project was unfinished. This situation caused that the original plan of the project changed. After discussing with the supervisors of this work, we concluded that part of the value of the present work was the original experimental design. Therefore, this project includes the proposed experimental design and obtained results according to the original plan. All the parts that could not be completed will be indicated with a footnote on the text, and this situation will be took into account in the Discussion section. This situation compromised, specifically, the control group and the B pattern of the experimental design:

> This situation emerged when the control group completed the 4th session. As a result, this project lacks the immediate and delayed test data of the control group.
>
> From the B pattern the delayed test results could not be obtained in school. Therefore, they have done the delayed test as homework without supervision and in an asynchronous way around 15 and 25 days after the immediate test.

2.2 Research Questions

Even though the e□ciency of using DL has been reported to improve the learning process, it is unclear which pattern leads to better results. It needs to be taken into account that previous studies have used di□erent topics to test DL e□ciency in di□erent learning contexts. In this case, CT will be taught to elementary school students through 6 week workshops. The research question of this work can be divided in two parts:

> Is DL improving results in LTM formation when compared to massive learning?[1]
>
> Is there a DL pattern that leads to better results in LTM formation?

2.3 Hypotheses

Based on the research question exposed in the above section, the hypotheses can be divided in two parts:

> The overall performance of DL groups will be better than the control group on delayed test but worse in the immediate. It is expected to perform better on the delayed test than in immediate test on DL groups. The intuition is based on the literature presented on the State of the Art section. According to the study by Rawson & Kintsch, DL leads to better performance on LTM retention but to a worse perforamance on immediate test results when compared to ML [5]. Therefore, it is expected that this phenomena will be observed on this experiment, so delayed test scores will be greater than immediate test scores on DL groups. [2]
>
> There is a particular DL pattern that leads to better results compared to the rest: pattern A. The perception is -as it has been reported in the 'State of the

[1] This research question could not be checked because there was not control group due the Covid-19 situation.

[2] Due the COVID-19 situation there was not control group, so the first part of this hypothesis could not be checked.

Art' section-, the gap between the retrieval sessions is an important factor to optimize the LTM formation. According to the presented literature, the most suitable DL pattern might include enough spaced gaps of review and those gaps might be on the order of months. As the workshop only lasts 6 weeks the most possible spaced gaps are presented on pattern A, which corresponds to reviewing the content the same day, 1 week after and 3 weeks after.

2.4 General Methodology Applied

The general methodology preserved the same structure for all experimental groups. The workshops 'Makers a les aules' consisted of 5 student's sessions and 2 co-design sessions with teachers. The schools participating in this study agreed to do 6 sessions with students in order to do a delayed test.

In the co-design sessions, I discussed with the teachers the projects that students would do at the workshop. In these workshops, the tools used for developing these projects are 'Makey Makey' and 'Scratch'. In those sessions, I discussed with the teachers the content that was about to be provided as CT theory to the students. After discussing with the schools the CT content slides were created and distributed in 5 sessions (in the 6th session no content was provided).

In the sessions made with students the followed structure was:

Session 1. At the beginning of the class a pre-questionnaire was provided to measure the prior knowledge of students about CT. Then, the content of session 1 was given. At the last 15 minutes of the session users did a retrieval exercise using Kahoot! software.

Session 2,3 and 4. At the first 15 minutes of the class, new content was provided. Then, students had time to work on their projects. At the last 15 minutes of the session users did a retrieval exercise using Kahoot! software.

Session 5. At the beginning of the class, new content was provided to students. Then, students had time to finish their projects. At the last 20 minutes of class,

students took the immediate test.

Session 6. Students needed to present their projects in front of the rest of the class. At the last 20 minutes of class, students took the delayed test.

The immediate test on session 5 was used to test students' performance on a short term. The delayed test on session 6 was used to test their LTM formation, as in the 6th session no content was provided.

2.5 Design & Development Criteria and Strategies of Artifact

Pre-questionnaire

The pre-questionnaire was created in order to be aware of the prior students' knowledge with Scratch, Makey Makey and CT. The pre-questionnaire consisted in two sections: the first section asked about previous experience using the tools Scratch and Makey Makey and the second section asked about CT with a pseudo-code condition example and a final exercise related to this example.

Theoretical Content

It was challenging to elaborate the materials for the groups. Before the sessions with students, there were a minimum of two co-design sessions with teachers. The purpose of these sessions was to plan out with the teacher the content of the project which students were about to develop with 'Scratch' and 'Makey Makey'. It is important to remark that each school worked on different project topics so it was a challenge to elaborate a useful theoretical material for all experimental groups. After discussing with all schools about the content and its order we decided the following content for each session:

Session 1. Introduction to 'Makey Makey' tool and algorithm definition.

Session 2. Inequality symbols and variables.

Session 3. Identify and understand conditions.

Session 4. Identify and understand loops.

Session 5. Definition and identification of inputs and outputs.

Review Questions

According to the experimental design that will be detailed later, all groups saw at least 2 questions of each topic before taking the immediate and delayed test -except for the content of session 5 which was evaluated directly on immediate and delayed tests. Therefore, I decided to make questions about two topics for each session but with a di□erent perspective: one with a multi choice answer and the other one with a True or False answer. Here there is an example of di□erent questions around the same topic:

Select the correct answer. The algorithms are:

a. Computer images.

b. Computer loops.

c. Instructions for the computer.

d. None of the above.

Indicate if the following statement is True or False: An algorithm are step by step solutions to complete a certain task.

After creating the list of all questions (which can be found in First Appendix) I selected Kahoot! as the tool to introduce the retrieval questions. Kahoot! is a free online software that allows users to create quizzes as a contest. The questions were provided to students by the projector and they answered using their computers, tablets or phones. I chose this software as some studies that applied this specific software to increase the student's engagement, some examples of these studies have been done by Alamanda [27]. Also, Kahoot! provides immediate feedback and it has

been reported that “providing feedback after RP further strengthens the benefits of RP” [28]. 4 Kahoots! were created for the first 4 sessions, each of them containing 5 questions and an extra one. Each Kahoot! showed the extra question at the very beginning of the session. This extra question was related to a fun topic and not to CT making it easier for students to understand how to use this tool and capture their attention at the same time.

Immediate and delayed test

Immediate test was created using the Theoretical content section topics. The test contained 10 questions (2 questions from each session). Each question had 3 possible answers: ‘True’, ‘False’ and ‘I don’t know’ which were scored as +1, -1 and 0 respectively. This scoring method is known as formula scoring. Formula scoring is a method that gives students the opportunity to acknowledge that they do not know the correct answer instead of forcing them to guess[29]. The general method to score tests is the number-right scoring which implies that only the number of correct answers is taken into account when calculating the total score and the incorrect answers are not subtracted from the total score. In the case of formula scoring, each question is scored as +1 if the answer is correct, -1 if it is incorrect and 0 if the "I don’t know" option is selected. Formula scoring o□ers an individualized way of correction for guessing and may reduce random guessing to as low as 2% of the items [30].

I asked the students only to answer ‘True’ or ‘False’ only if they were really sure of it before giving them the test. If they were not sure about the answer they had to choose the ‘I don’t know’ option. It was reported to them the scoring method so they could have it in mind while doing the test.

The delayed test structure and the topics were the same as the immediate test. Even so, the perspective and some details of the query changed so the questions were not exactly the same as the previous test.

2.6 Methodology

2.6.1 Participants

The participants for the proposed experimental design were elementary school students from different schools in Barcelona. During the experiment, each school represented one experimental group. The group was assigned randomly for each school before knowing the students prior knowledge or age, in order to not bias the selection. It would not be possible to mix different experimental groups on same school as all students from each of the schools had the same schedule so the sessions were done at the same time with all students. Therefore, doing different reviews within the same school group would not be possible.

All schools needed to be participating in Scratch and Makey Makey workshops delivered by TIDE research group of the Universitat Pompeu Fabra. The regular duration of these workshops is 5 weeks, but it was negotiated with participating schools to have a 6th session to run the delayed test.

2.6.2 Experimental Design and Set-up

In this section I detail the proposed experimental design, the challenges and obstacles emerged during the process and the revised experimental design.

During the planning phase, several experimental designs were discussed until the proposed experimental design was constructed. The workshop consisted in 6 sessions. This is the reason why both experimental designs are constructed considering only 6 sessions.

Proposed Experimental Design

The proposed design is a between subjects experiment using 4 experimental groups. The independent variable is the distribution of review sessions over time using a DL pattern. Dependent variables are the results on immediate and delayed tests.

The first experimental design consists in 4 groups:

3 groups using 3 different DL patterns and RP.

1 control group using RP but no DL.

The distribution of review questions for each group can be seen in Figures 4,5,6,7.

A pattern	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions S1	3 Questions S2 2 Questions S1	3 Questions S3 2 Questions S2	3 Questions S4 1 Question S3 1 Question S1	Immediate test	Delayed test

Figure 4: A DL Pattern

B pattern	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions S1	3 Questions S2 2 Questions S1	3 Questions S3 1 Question S2 1 Question S1	3 Questions S4 1 Question S3 1 Question S2	Immediate test	Delayed test

Figure 5: B DL Pattern

C pattern	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions S1	3 Questions S2 2 Questions S1	3 Questions S3 1 Question S2 1 Question S3	2 Questions S4 1 Question S3 1 Question S2 1 Question S1	Immediate test	Delayed test

Figure 6: C DL Pattern

Unfortunately, one of the four schools was not available for this experimental design because the sessions were conducted by another person of the TIDE research group with different experimental objectives. Therefore, one of the groups was left out of the experimental design. I decided to leave out C by considering several factors.

Control group	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions T1	5 Questions S2	5 Question S3	5 Questions S4	Immediate test	Delayed test

Figure 7: Control group

I did not want to leave out the control group as it is an important factor to compare (DL Vs not DL).

Also I did not want to leave out A for its potential benefits if it worked well (as the reviews are the most spaced of all patterns, which could be very practical for schools) and also B and C were actually very similar to each other but A was quite different.

I could have left out B or C -as they only differ in session 4, C includes topic 1 review questions and B not. Finally, I thought it was better to leave out C, as the proposed pattern kept increasing the topics of questions and results when reviewing all the previous content at the during session, which could have been exhausting for teachers and students. Also in practical terms it would not be helpful as the number of questions would keep increasing during a term being that difficult to apply in a school course.

Therefore, the revised experimental design contains groups A, B and the control group.

Revised Experimental Design

The revised experimental design is a between subjects experiment using 3 groups. The independent variable is the distribution of review sessions over time using DL patterns. Also, a control group is included doing daily reviews on a day-to-day basis.

Dependent variables are the results on immediate and delayed tests. RP questions were made through 5 questions at the end of the class. The same content was

delivered to all groups -through slides and oral explanation made by the same person- during the 5 first sessions and with the same immediate and delayed test.

Some studies reported that prior knowledge is a key factor that must be taken into account in order to evaluate the learning process [20].In order to control the prior knowledge of students, a pre-questionnaire was made. When analysing the results, the prior knowledge score will be taken into account to compare the results on immediate and delayed tests.

The revised experimental design consists in 3 groups:

2 groups using 2 di□erent DL patterns and RP.

1 control group using RP but no DL. [3]

Therefore, all the materials were constructed having in mind this experimental design. The groups used in the experimental design can be seen in Figures 8, 9, 10.

A pattern	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions S1	3 Questions S2 2 Questions S1	3 Questions S3 2 Questions S2	3 Questions S4 1 Question S3 1 Question S1	Immediate test	Delayed test

Figure 8: A DL Pattern

B pattern	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions S1	3 Questions S2 2 Questions S1	3 Questions S3 1 Question S2 1 Question S1	3 Questions S4 1 Question S3 1 Question S2	Immediate test	Delayed test

Figure 9: B DL Pattern

[3]Due the Covid-19 situation, this group didn't finish the workshop and its data was not recollected.

Control group	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
Content provided	Content Topic 1	Content Topic 2	Content Topic 3	Content Topic 4	Content Topic 5	No content provided
Experimental Activities	Pre-questionnaire 5 Questions T1	5 Questions S2	5 Question S3	5 Questions S4	Immediate test	Delayed test

Figure 10: Control group

2.6.3 Procedures Used to Obtain Data and Results

In order to obtain data and results three resources were created.

Pre-questionnaire

The first section of this questionnaire consisted of yes/no questions including: *"Have you used Scratch?" Have you used Makey Makey?"*". Then, I asked students how many times they used these tools. They could select between 4 options from 'Never' to 'A lot of times'. Then, they had an open question so they could explain their experience in detail. This section will be used to test the familiarity of students with the used tools. Their binary questions will be quantified in order to average the overall experience of each group with the tools.

The second section of this questionnaire had a CT related example. Then, students needed to answer a programming question that included a condition. Students needed to understand the example and complete the exercise according to it. This question was quantified as 1 -if it was answered correctly - or 0 - if the answer was wrong. The results of this question will be averaged to have an indicator of the overall performance on CT of each group.

Immediate and Delayed Test

Both tests consisted of a 10 questions questionnaire. The scores were computed by adding 1 if the answer was correct, subtracting 1 if the answer was wrong or adding 0 if the answer was 'I don't know'. Following these scoring rules, the grade of each student will be computed. Then the average score of each group will be computed

as well as the standard deviation. These indicators will be used to compare the overall performance on the immediate and delayed test in each group.

2.6.4 Analysis

In the case of immediate and delayed test, the averaged scores on a 1-10 scale of all groups will be compared between the three groups. The aim is to check if there is a significant difference on their performance on both tests [4]. In order to do it, a one-way ANOVA analysis will be made for each pair of groups - in case data is normally distributed - to test both hypotheses of this work. In order to test if the data is normally distributed a normality test will be done with all scores from both tests.

In the case of the pre-questionnaire, the students prior knowledge will be determined according to their results on this tests. Then, students will be divided on high and low-prior knowledge groups. Then, hypothesis of this work can be analyzed using prior knowledge as an independent variable in order to analyze if this could be a relevant factor on the results.

[4] Only data from 2 groups was analyzed because the control group data was missing due the COVID-19 situation

Chapter 3

Results

3.1 Sample Description

The participants in the proposed experimental design were 65 elementary school students between 9 and 12 years old from three schools in Barcelona. As it was mentioned before, due the COVID-19 situation 1 of the 3 schools was not able to finish the workshops. Therefore the final sample was composed by 42 elementary students from two di□erent Barcelona public schools. Specifically, on school A - which corresponds to Pattern A group - have participated 25 students and on school B - which corresponds to Pattern B- 17 students.

Pattern A students were older than ones in Pattern B. Students on Pattern A were between 11 and 12 years old - in average = 11.24 years old- while students on Pattern B were between 9 and 10 years old -in average 9.12 years old-. On both patterns there were more boys than girls: Pattern A was composed by 15 boys and 10 girls and Pattern B by 12 boys and 5 girls.

The participation on all the tests were not the same on both schools. All students from Pattern A have done the pre-questionnaire: 25 and almost all students from Pattern B: 16 - because one student did not attend to the first session-. All students from Pattern B have done the immediate test: 17 but only 23 from Pattern A - because two students did not attend to class that day-. The biggest di□erence on

participation is on the delayed test. Almost all participants from Pattern A have done the delayed test: 24. But only 9 from Pattern B. The reason is that due the COVID-19 situation it was send as homework to students by e-mail and only 9 replied.

3.2 Performance on Immediate and Delayed Tests

The first hypothesis of this project was: "The overall performance of DL groups will be better than the control group on delayed test but worse in the immediate. It is expected to perform better on the delayed test than in immediate test on DL groups.".

As it was mentioned before, the first part of the first hypothesis can not be tested - as the control group is missing due the COVID-19 situation-. Even so, the second part can be tested: "It is expected to perform better on delayed than in immediate test on DL groups".

In order to check the second part of the hypothesis an ANOVA test was computed for both test - delayed and immediate -. ANOVA test requires data to be normally distributed, so first the Kolmogorov- Smirnov normality test was conducted for both pattern.

The normality test concluded in all 4 cases that data does not di□er significantly from that which is normally distributed. The results from this Normality test can be seen on Figure 11 and 12. The test statistic (D), which can be seen below, provides a measurement of the divergence of the sample distribution from the normal distribution. The higher the value of D, the less probable it is that the data is normally distributed. The p-value quantifies this probability, with a low probability indicating that the sample diverges from a normal distribution to an extent unlikely to arise merely by chance [31].

After concluding that all data does not di□er significantly from that which is normally distributed, ANOVA test can be conducted. Two one way ANOVA test were conducted with two Null hypotheses, one for each pattern:

Pattern A	Immediate test	Delayed test
D value	0,20	0,21
p value	0,29	0,23

Figure 11: Kolmogorov-Smirnov Normality Test Results - Pattern A

Pattern B	Immediate test	Delayed
D value	0,25	0,28
p value	0,2	0,39

Figure 12: Kolmogorov-Smirnov Normality Test Results - Pattern B

- Null Hypothesis 1. Average scores from immediate test and delayed test are equal for Pattern A.

- Null Hypothesis 2. Average scores from immediate test and delayed test are equal for Pattern B.

Pattern A students obtained an average score of 6,22/10 on the immediate test and 7,21/10 on the delayed one. The results for null hypothesis 1 were: f-ratio = 2,82 and p-value = 0,099881. Therefore, the result is not significant at $p < 0,05$ but it is at $p < 0,10$. So there is a significant di□erence between the immediate and delayed test scores on Pattern A - with a significance level of 0,10-. According to this result, average scores from delayed test are significantly better than scores from the immediate test for Pattern A with a significance level of 0,10.

Pattern B students obtained an average score of 4,71/10 on the immediate test and 7,56 on the delayed one. The results for null hypothesis 2 were: f-ratio = 6,24 and p-value = 0,019729. Therefore, the result is significant at $p < 0,05$. So there is a significant di□erence between the immediate and delayed tests scores on Pattern B - with a significance level of 0,05-. According to this result, average scores from delayed test are significantly better than scores from the immediate test for Pattern B with a significance level of 0,10.

According to the obtained results, average scores from the delayed test were significantly better than on the immediate test. Therefore, students had perform better on the delayed than on immediate test on both DL groups. So the second part of the first hypothesis "It is expected to perform better on

delayed than in immediate test on DL groups" is accepted with the signi cance levels mentioned above. All results from this second part of rst hypothesis will be further discussed in the Discussion section.

3.3 Comparison of Performance Between DL Groups

The second hypothesis was: "there is a particular DL pattern that leads to better results compared to the rest: Pattern A". In order to check this hypothesis, a one-way ANOVA test was conducted. There was no need to check if the data was normally distributed as it was already done to test the rst hypothesis.

- Null Hypothesis 1. Means from immediate test from Pattern A and Pattern B are equal.

- Null Hypothesis 2.Means from delayed test from Pattern A and Pattern B are equal.

The comparison of average scores on both test between the two patterns can be observed on Figure 13.

Figure 13: Average scores on immediate and delayed test from Pattern A and B

Performance on Immediate Test

Null hypothesis 1 is related to the immediate test. The immediate test was delivered to students on week 5, the same day in which content from topic 5 was teach. It was scored on a 0-10 scale.

On Figure 14 it can be observed that the average score on Pattern B is under 5 - 4.71 - while on pattern A is above 5 - 6.22 . It is also remarkable that the variance is bigger on pattern B -10.10- than on pattern A - 4.63-. The percentage of failures on Pattern B over a quarter part of the students -35.39%- while its under a quarter part on Pattern A -35.29%. The percentage of scores between 5 and 6 is quite similar on

Figure 14: Immediate test results

(a) Pattern A scores distribution (b) Pattern B scores distribution

Figure 15: Immediate test scores distribution

both groups - 43.48% on Pattern A and 41.18% on Pattern B-. On scores between 7 and 8 there is a bigger di erence between both patterns - 26.09% on pattern A and 17.65% on Pattern B-. On top scores between 9 and 10 there is even a bigger di erence between both patterns - 13.04% on Pattern A and 5.88% on Pattern B-. A graphical distribution of the scores percentages from both groups can be seen on Figure 15. Taking into account all this statistics, it can be considered that Pattern A performed better than Pattern B on the immediate test, but in order to test it an ANOVA test have been done.

Results for the null hypothesis 1 on the ANOVA test were: f-ratio = 3,22 and p-value = 0,08. Therefore, the result is not signi cant at $p < 0,05$ but it is signi cant at $p < 0,10$. So there is a signi cant di erence between the immediate test scores on Pattern A and Pattern B - with a signi cance level of 0,10 -. According to this result,

average scores from immediate test are signi cantly better for Pattern A than for Pattern B with a signi cance level of 0,10.

Performance on Delayed Test

Null hypothesis 2 is related to the delayed test. The delayed test was delivered on week 6 for pattern A and between week 7 and week 8 for Pattern B - this di erence is explained on Discussion & Conclusion section: Limitations Due the COVID-19 -. It was scored on a 0-10 scale.

Figure 16: Delayed test results

On Figure 16 it can be observed that the average score on Pattern B and Pattern A are quite similar - 7.21 on Pattern A and 7.56 on Pattern B -. It is also remarkable that the variance is also quite similar on both patterns - 1.89 on Pattern A and 1.67 on Pattern B-. The percentage of failures on Pattern A is under 10% while on Pattern B there is a 0% of failures. There is a big di erence on the percentage of scores between 5 and 6 on both group- 16.67% on Pattern A and 44.44% on Pattern B-. On scores between 7 and 8 there is also a big di erence between both groups - 58.33% on pattern A and 11.11% on Pattern B-. On top scores between 9 and 10 there is also a big di erence between both patterns - 16.67% on Pattern A and 44.44% on Pattern B-. A graphical distribution of the scores percentages from both groups can be seen on Figure 17. Observing this distributions, we can see that the

(a) Pattern A scores distribution (b) Pattern B scores distribution

Figure 17: Delayed test scores distribution

mean is quite similar from both groups but distributions are not. On Pattern B there are no failures and most of scores are distributed between pass and excellent scores. On contrast, on Pattern A the percentage of failures is not 0 but quite low and the majority of scores are remarkable scores. To see if the di erence between the performance of both groups is signi cant, a one-way ANOVA test have been done.

Results of the one-way ANOVA test for the null hypothesis 2 were: f-ratio = 0.97 and p-value = 0.63. Therefore, the result is not signi cant at $p < 0,05$ or $p < 0,10$. So there is not a signi cant di erence between the delayed test scores on Pattern A and Pattern B. According to this result, average scores from Pattern A on the delayed test are not signi cantly better compared to scores from Pattern B.

Comparison of Performance Between DL Groups: Results

According to the obtained results, the second hypothesis of this work is rejected. There is not a signi cant di erence on the performance of Pattern A and Pattern B on the immediate and delayed test. Pattern A does not signi cantly lead to better results than Pattern B on the delayed test. Even so, it seems that Pattern A leads to signi cant better results on the immediate test compared to Pattern B with a signi cance level of 0,10. All results from second hypothesis will be further discussed in the Discussion section.

3.4 Additional Analysis

3.4.1 Analysis Based on Prior Knowledge

The pre-questionnaire was delivered at the beginning of session 1, before delivering any class to students. It contained questions about previous experience with Scratch, Makey Makey and one CT problem which was scored as correct or incorrect.

We can observe that there is a di erence on students' prior knowledge about Scratch on the two patterns. As we can see on Figure 18, all students in pattern A have used Scratch before the workshop. In contrast, less than a 15 per cent of pattern B students have previous experience with this tool -only 12.5 % have used Scratch-. In contrast, we can see that the previous experience on Makey Makey is quite similar in the two models. In this case, Pattern B has a higher percentage of prior knowledge about Makey Makey with 12.5%, while Pattern A has only 4%. As for the CT exercise, the percentage of correct answers is quite high in both patterns compared to the prior knowledge on the mentioned tools. In the case of Pattern A there is a 88% of correct answers and in Pattern B a 68.75%.

Figure 18: Students previous knowledge

There are di erences on the previous knowledge of students on Scratch, Makey Makey and CT. This could be an important factor when analysing the performance

of students. Therefore, the first and second hypothesis can be checked grouping the student with low-prior knowledge and high-prior knowledge. Students with high-prior knowledge are the ones that had previous experience with at least one of the used tools - Scratch and/or Makey Makey - and had answered to the CT problem correctly. Rest of students will be considered part of the low-prior knowledge group.

One-way ANOVA tests have been done in order to check the second part of the first hypothesis and the second hypothesis dividing student on low-prior knowledge and high-prior knowledge groups.

Performance on Immediate and Delayed Tests Based on Prior Knowledge

In this case, I have checked if there is a significant difference between the immediate and delayed test performance based on the prior knowledge. This corresponds to testing second part of the first hypothesis with prior knowledge as an independent variable. I would like to run this analysis dividing students with high prior knowledge from Pattern A, low-prior knowledge from Pattern A, high prior knowledge from Pattern B and low-prior knowledge from Pattern B. That was not possible as there were not enough samples in each group in order to run an ANOVA test. Therefore, there have been tested differences on scores from immediate and delayed test for all students -from Pattern A and B- with low-prior knowledge and with high-prior knowledge. The aim is to check if the prior knowledge was a relevant variable for determining the results of immediate and delayed test.

- Null Hypothesis 1. Average scores from immediate test and delayed test are equal for the high-prior knowledge group.

- Null Hypothesis 2. Average scores from immediate test and delayed test are equal for the low-prior knowledge group.

Average scores from immediate and delayed test for both prior knowledge groups can be observed at Figure 19.

Results of the one-way ANOVA test for the null hypothesis 1 were: f-ratio = 3,91 and p-value = 0,05. Therefore, the result is not significant at $p < 0,05$ but it is at

Figure 19: Average test scores depending on prior knowledge.

$p < 0,10$. So there is a signi cant di erence between immediate and delayed test scores on high-prior knowledge group. High-prior knowledge students have obtained better scores on the delayed test than on the immediate test with a signi cance level of $p < 0,1$.

Results of the one-way ANOVA test for the null hypothesis 2 were: f-ratio = 4,77 and p-value = 0.04. So there is a signi cant di erence between immediate and delayed test scores on low-prior knowledge group. Low-prior knowledge students have obtained better scores on the delayed test than on the immediate test with a signi cance level of $p < 0,05$.

According to the obtained results, we can observe that both groups have performed signi cantly better on the delayed test compared to the immediate test at least with a signi cance level of $p < 0,1$.

Comparison of Performance Between Groups Based on Prior Knowledge

In this case, I have checked if there is a signi cant di erence between low-prior and high-prior knowledge performance on the immediate and on the delayed test. As in the previous section, I would like analyse second hypothesis introducing prior knowledge as an independent variable. I would like to di erentiate between Pattern A and B but that was not possible as sample sizes were not big enough. Therefore, the aim of this analysis is to check if there was a signi cant di erence on the performance of the immediate and delayed test depending on the student's prior knowledge level, without taking into account DL patterns.

- Null Hypothesis 1. Means from immediate test from high-prior knowledge and low-prior knowledge students are equal.

- Null Hypothesis 2. Means from delayed test from high-prior knowledge and low-

prior knowledge students are equal.

Results of the one-way ANOVA test for the null hypothesis 1 were: f-ratio = 1,21 and p-value = 0,28. Therefore, the result is not signi cant at $p < 0,05$ or at $p < 0,10$. So there is not a signi cant di erence on the performance of the immediate test between low and high-prior knowledge students.

Results of the one-way ANOVA test for the null hypothesis 2 were: f-ratio = 0.1 and p-value = 0,76. Therefore, the result is not signi cant at $p < 0,05$ or at $p < 0,10$. So there is not a signi cant di erence on the performance of the delayed test between low and high prior knowledge students.

According to the obtained results, prior knowledge level is not a relevant variable on the performance of students on any of the tests. All results from this analysis based on prior knowledge level will be further discussed in the Discussion section.

3.4.2 Analysis of Gain Between Immediate and Delayed Test

On the proposed hypotheses of this work it was compared performance of DL groups on the results of immediate and delayed test, but not the evolution from one test to the other. It could be that one of the DL pattern leads to a signi cant greater gain than the other. The average gains of both topics can be seen on Figure 20. Positive values mean that the performance on the delayed test was better than on the immediate. Averages were computed subtracting immediate from delayed test scores for all participants and then computing the average of those gains for each of the patterns. Therefore, the range of averages could go from -10 to 10. An ANOVA test has been conducted to compare both average gains. The null hypothesis is:

- Null hypothesis: The average gain from immediate to delayed test is the same for both DL patterns.

Figure 20: Gains between immediate and delayed test for both DL patterns.

The f-ratio value is 2,04. The p-value is 0,16. The result is not signi cant at p < 0,5 or p < 0,10. Therefore, it seems thatthere is not a signi cant di erence between the two DL patterns on the gain from immediate to delayed test.
This result will be further discussed on Discussion section.

3.4.3 Analysis of Gain of Topic 1 and 2 Between Immediate and Delayed Test

The proposed DL patterns have similar review structure but they de er on reviews of topics 1 and 2. Both patterns have answered a total amount of 8 questions about topic 1 and 5 questions about topic 2. Even so, the distribution of these reviews are di erent.

In the case of Pattern A review questions about topic 1 were asked on sessions 1, 2 and 4 and questions about topic 2 on sessions 2 and 3. In the case of Pattern B topic 1 was reviewed on sessions 1, 2 and 3 and 2 on sessions 2,3 and 4. Therefore, as both have exactly the same structure on the rest of review sessions, makes sense to compare the gain of both patterns only on the results of questions about topic 1 and 2. The average gains of both topics can be seen on Figure 21. Negative values mean that the performance on the immediate test was better than on the delayed. Positive values mean the opposite. The average was computed summing scores earned on both questions on immediate test and subtracting the sum of scores of both questions on delayed test. Therefore, the gain scale is between -2 and 2. In order analyze if the di erence in the gains is signi cant, ANOVA tests were computed with following null hypotheses:

- Null hypothesis 1. The gain between immediate and delayed test questions about topic 1 is the same for both DL groups. - Null hypothesis 2. The gain between immediate and delayed test questions about topic 2 is the same for both DL groups.

In the case of topic 1 - null hypothesis 1- the f-ratio= 0,21 and the p-value = 0,65. Therefore the result is not signi cant at p < 0,05 or p < 0,1. It seems thatthere is not a signi cant di erence between the gain of both DL groups on topic

Figure 21: Gains on topic 1 and topic 2 for both DL patterns.

1 questions.

In the case of topic 2 - null hypothesis 2- the f-ratio = 1,74 and the p-value = 0,2. Therefore, the result is not signi cant at $p < 0,05$ or $p < 0,1$. It seems that there is not a signi cant di erence between the gain of both DL groups on topic 2 questions.

The results obtained in this section will be further discussed on Discussion section.

3.4.4 Comparison of DL Groups Performance on Topic 1 and 2

On the section above, it was analyzed the gain between the immediate and delayed tests on those questions from topic 1 and 2. In this section, I analyzed the di erence on the performance of both DL groups on those topics using a one-way ANOVA test. Null hypotheses are the following:

-Null hypothesis 1. Average scores from immediate test questions about topic 1 and 2 are the same for both patterns.

-Null hypothesis 2. Average scores from delayed test questions about topic 1 and 2 are the same for both patterns.

In this case, the average of scores from questions about session 1 and 2 was computed and compared for both patterns - for each of the tests-. Therefore, the average scores can go from -1 to 1 -remember all questions were scored with 1 if correct, -1 if incorrect and 0 if the I don't know option was selected-. Average scores can be seen on Figure 21.

Results of the one-way ANOVA test for null hypothesis 1 were: f-ratio = 0,92 and p-value = 0,34. Therefore, the result is not signi cant at $p < 0,05$ or $p < 0,1$.

Figure 22: Average scores of DL groups on questions about topic 1 and 2

The null hypothesis 1 is not rejected as there is not a signi cant di erence on average scores from immediate test questions about topic 1 and 2 from both patterns.

In the case of null hypothesis 2, the average score were exactly the same -as it can be seen on Figure 22-. Therefore, it did not make sense to run the ANOVA test. Null hypothesis 2 was acceptedaverage scores from delayed test questions about topic 1 and 2 are the same for both patterns.

Results on this section will be further discussed on Discussion section.

Chapter 4

Discussion & Conclusion

4.1 Interpretation of Results

4.1.1 Performance on Immediate and Delayed Tests

The first part of the first hypothesis: "The overall performance of DL groups will be better than the control group on delayed test but worse in the immediate" could not be checked because the control group data was missing due the COVID-19 situation.

The second part of the first hypothesis: "It is expected to perform better on the delayed test than on immediate test on DL groups" was analysed on results section. The results concluded that students have performed significantly better on the delayed test compared to the immediate test. This can be explained by DL observed results mentioned on the State of the Art. According to the results on the study by Rawson & Kintsch [5] DL groups perform better on the delayed test than on immediate test. Therefore, DL seems to be useful to form durable learning, so the second part of the first hypothesis is not rejected. It seems that DL groups perform significantly better on the delayed than on the immediate test.

4.1.2 Comparison of Performance Between DL Groups

The second hypothesis of the present project was: "there is a particular DL pattern that leads to better results compared to the rest: Pattern A". This hypothesis has been rejected because the di erence between both group scores on the delayed test was not signi cant.

Even so, the di erence between the scores from both patterns on the immediate test was signi cant. In particular, Pattern A seems to lead to better immediate test results than Pattern B. Therefore, it could be that the di erent DL patterns lead to di erent short term results. This is an interesting result as one of the "weak" aspects of DL is observed worse results on immediate performance. Finding a DL pattern that optimizes both immediate and delayed performance could have great implications on education. As the control group -which corresponded to ML group- was missing on this study, its performance on immediate and delayed test could not be analysed. According to the results of this work Pattern A leads to signi cant better results on the immediate test when compared to Pattern B. Therefore, comparing Pattern B with ML could give the opportunity to compare the immediate test performance of both groups. According to the obtained results, this di erence could be smaller than ML Vs Pattern A. Testing this results with other groups could help to consolidate that Pattern A leads to better results on immediate retention than Pattern B which could help to gradually decrease the gap between the immediate test performance of ML and DL.

4.1.3 Additional Analysis

Analysis Based on Prior Knowledge

According to the rst hypothesis analysis based on students' prior knowledge level, we can observe that results are quite similar to the ones obtained testing hypothesis 1 - comparing scores from immediate and delayed test from both patterns-. Therefore, seems logical to think that the better performance on the delayed test does not depend on the pattern or the prior knowledge - as all had signi cant better results

on the delayed test -. It might be that the di erence is the DL instead of any other variable, as it has been said on literature before.

According to the second hypothesis analysis based on prior knowledge, results showed that there is not a signi cant di erence on the performance of the immediate or delayed test between low and high-prior knowledge students.Therefore, it seems that the better or worse performance of students on the tests does not depend on their prior knowledge level. This is a useful result as it has been seen that Pattern A have more prior-knowledge than Pattern B. If results have shown that prior-knowledge introduced a signi cant di erence on test performance, it would be di cult to extract conclusions from the data - as it would be impossible to distinct if the good results were because the DL pattern or because pattern A had greater levels of prior knowledge -.

Analysis of Gain Between Immediate and Delayed Test

According to the analysis of the results, there is not a signi cant di erence between the two DL patterns on the gain from immediate to delayed test. Nevertheless, the di erence on the average gains -Pattern A: 0,91 and Pattern B: 2,56 - are both in favor of delayed test and Pattern B has reported a greater improvement even if the result was not signi cant. It could be that the short sample size of Pattern B on the delayed test did in uence the results or the fact that it was not supervised. On future studies, these gains should be further analysed as a larger gain on the delayed test from one pattern to another could be indicating that one pattern leads to better performance on LTM formation. Or it could be that one of the two patterns leads to better performance on short-term memory formation, as it has been discussed on the discussion of the section "Comparison of Performance Between DL groups". On that section it was discussed that Pattern A seems to lead to better performance on immediate test when compared to Pattern B. This fact combined with the result of Pattern A having less gain that Pattern B could be indicating that Pattern A leads to a memory formation that keeps more balanced over time. In other words, Pattern A could be helping to create a learning on a shorter amount of time and

maintaining these knowledge on long-term memory.

Analysis of Gain of Topic 1 and 2 Between Immediate and Delayed Test

According to the results, there is not a signi cant di erence between the DL groups on the gain of topic 1 and topic 2 correctly answered questions between immediate and delayed test. Therefore, it seems that DL the di erence on reviews distribution is not signi cantly in uencing the evolution of memory retention from one test to the the other. This possibility will be further discussed on next paragraph.

Comparison of DL Groups Performance on Topic 1 and 2

According to the results, there is not a signi cant di erence on the performance of both DL groups on topic 1 and 2 questions. This result is not expected because according to the results of the section "Comparison of Performance Between DL Groups" Pattern A leads to better immediate tests results. But the only di erence between patterns were topics 1 and 2 reviews distribution. If the performance of both DL groups was not signi cantly di erent on topics 1 and 2 - that were the only di erence between the 2 patterns - which was the factor that introduced that di erence?

On the limitations section will be detailed all the constraints of this project. One of those was that Pattern B did the delayed test without supervision. It is suspicious that performance on immediate test was signi cant di erent for both DL groups with Pattern A having a signi cant better average score- average score of Pattern A = 6,22 and for Pattern B = 4,71 - but it was not signi cant di erent on delayed test average scores with Pattern B having better average score - average score of Pattern A = 7, 21 and for Pattern B = 7,56-. But these di erences are not on those topics that have been reviewed di erently. This could be for many reasons. Seems logical to think that this di erence could be because the fact that Pattern B group did not have supervision on delayed test as it obtained worse scores than pattern A on immediate -which was supervised- and better scores than Pattern A on delayed test -which was not supervised-. Therefore it could be that Pattern B

students obtained help to do the delayed test or only those that feel con dent about their knowledge decided to send the delayed test - Only 9 students from Pattern B send me the realized test-. Another reason for this di erence could be that the sample size of Pattern B on the delayed test was shorter than on the immediate, so it could be not big enough to represent the overall performance of the DL pattern. Another possible variable could be the di erence on timings between both patterns. Due the COVID-19 situation Pattern B have done the delayed test 2 weeks after the last session while Pattern A have done it 1 week after. Knowing that there were at least the mentioned uncontrolled variables, results of this work do not seem very reliable.

4.2 Limitations

4.2.1 Limitations Due The COVID-19

Due the COVID-19 situation, the proposed methodology of this work could not be followed as it was planned. Here I detail the consequences of this situation on the present work:

One of the two hypothesis of this work could not be checked. The control group was not able to nish the sessions of the Makers a les Aules workshop and therefore could not be evaluated. In later studies the control group should be studied and adequately compared to the two DL patterns in order to check the DL e ciency on CT on elementary school students.

The Pattern B did the delayed test around 2 weeks after the immediate test, while Pattern A did it 1 week after as it was planned. This happened as Pattern B needed to do the delayed test on the same day schools started to close on Spain. Teachers helped me to deliver students the delayed test so they could do it at home. Even so, for organizational reasons they did not send the test to students until 15 days after the immediate test. Therefore, students from Pattern B have done the delayed test on an asynchronous way.

Pattern B students have done the tests without supervision. For the reason mentioned above, Pattern B students have done the tests on their own homes. Therefore, it can not be checked if they received help from their parents, or if they have search information while doing the test. Therefore, the results of the delayed test of Pattern B could not be re ecting the reality of Pattern B students performance.

4.2.2 Other limitations

There were some factors that limited the present study:

Sample size. The sample size was determined by the group size of schools and the number of authorizations from parents. As the sample size was not quite big it could be not representing the reality of the population studied. Furthermore, as it was mentioned on results section some analysis could not be done because the sample size was not big enough to run some tests.

Workshop duration. Ideally, LTM formation should be test on a larger amount of time to check how it evolves with time. Nevertheless, the schedules of tests were limited with school schedules and workshop duration. Makers a les aules project is a 5 week workshop, it was negotiated with participating schools to extended it to 6 weeks as an exception. Even so, more time could be useful to analyse LTM on a greater interval of time.

Population educational level. All participants from Pattern A and B were elementary school students. Even so, Pattern A students were on the 6th grade while Pattern B students were on 4th grade. Therefore, their educational level di er on two entire years. This di erence could be introducing an undesired variable between the groups, which could be hard to detect. Even so, both groups were compared because they are part of the second level of primary education - it was a restriction for participating on Makers a les Aules workshop-. Therefore, even if educational level of both groups was not exactly the same at least both of them were part of 2nd level of primary education.

4.3 Further work

According to the observed results, the second part of the first hypothesis of this work is not rejected: "it is expected to perform better on delayed than on immediate tests on DL groups". This hypothesis supports the spacing effect characteristics. Even so, the significance level is only 0,10 so repeating the study with a larger sample size could help to support it. Also, further studies could include the control group in order to compare the performance of DL groups with an ML group.

The second hypothesis of this work was rejected as any of the groups performed significantly better than the other on the delayed test. Even so, it has been found a significant difference on the performance of both DL groups on the immediate test. Running other tests and comparing it with more DL patterns could help to find if there is a DL pattern that improves not only the LTM formation but also the immediate students performance. Finding this DL pattern could tip the balance in favor of the DL when compared to ML. As it has been mentioned before, ML has reported better results on short-term tests. Finding a DL pattern which give good results on both immediate and delayed test could be a great advance on educational models.

Further studies could also take into account the educational level of students. Finding an homogeneous sample on age, previous knowledge and educational level could help to decrease uncontrolled variables. Even so, according to the additional analysis of this work, prior knowledge does not seem to influence the performance of students. This fact could be deeper studied and with a sample size more analysis could be done. For example: comparing different DL patterns with high-prior knowledge students and low-prior knowledge students. It could be that some DL pattern were more or less effective depending on previous learning. Therefore, deeper analysis on how this variables interact could help to find more personalized DL patterns.

The difference of gains between different DL patterns should be deeper analysed. Analysing the performance of different DL pattern over time could be a key point in order to find the optimal learning method. According to the literature presented on

this project it has been reported that ML leads to better performance on immediate than on delayed tests. Finding a DL pattern with a small gain and good performance on both tests would create a learning method that combines bests characteristics from ML and DL. In other words, that could mean a method that do not need to sacri ce short or long-term memory retention. Instead, a method that could optimize performance of students from short to a long period of time.

The methods used for reviews and evaluations -multi-choice test using Kahoot! or regular paper sheets- have not a proven reliability. As it has been mentioned before on this work, some studies have reported Kahoot! as a useful tool to increase students engagement. Even so, comparing di erent RP methods could help to nd the one that leads to better results when combined with DL. Multi-choice questions used for the delayed and immediate tests could be further analysed and improved to be sure they are enough reliable.

4.4 Conclusion

As conclusion, this work did nd one signi cant result on the second part of the rst hypothesis: "it is expected to perform better on delayed than on immediate tests on DL groups". This supports the spacing e ect characteristics mentioned on the literature. Even so, it has also been observed on the results of this study that there is one pattern that leads to signi cant better results on immediate tests. This result could be indicating that there are DL patterns that lead to better short-term performance than others. If that was the case, is that sacri cing LTM formation? There is always the need to have a balance between immediate results and long-term results? What if there is a particular DL pattern that has same -or similar- short-term results when compared to ML with same long-term results as any other DL pattern? There are still plenty of possibilities to analyze and results of this work could help later studies to decide next steps that could be done in this educational research ambit. Even so, as it has been commented on the results section, the signi cant better performance of Pattern A on immediate test when compared to Pattern B could be caused by other variables. The lack of supervision of Pattern

B on the delayed test and the short amount of data from this group on this test could be undesired variables that a ect the results. We also need to keep in mind that the Pattern A have performed delayed test one week after last session while Pattern B have done it two weeks after. Therefore, limitations of this work could be in uencing obtained results so further studies should avoid those constraints in order to obtain more reliable data to analyze.

Thanks to this project I learned a lot about DL and the studies that have been done about it, which helped me to construct the proposed experimental model. The experimental model created for this work is detailed and each design decision is justi ed with reported literature. Unfortunately, this model was not adequately applied due the COVID-19 situation. But I hope the constructed materials and the proposed experimental model could be useful for later studies related to DL and educational research. The sessions that have been done on schools with DL patterns took place without problems, until COVID-19 situation emerged. Therefore, the DL patterns proposed seem to be suitable for a formal school context, doing review sessions each week. Even thought, the context of this project was doing a workshop not a curricular subject. Therefore, the application of these patterns might need to be modi ed in order to correctly be scheduled on curricular subjects.

While doing the research for the State of the Art of this project I learned about RP through realized studies and I discover the di erent ways in which it can be applied. About CT I was surprised to discover that European Commission has done research around educational approaches related to it. Speci cally, it surprises me the that CT is being teach from early stages of education as I have never studied CT on school.

With this project I learned that working with elementary school students can be challenging, as its hard to keep their attention and help them to maintain their focus. Nevertheless, they are still on an early educational stage and researchers can learn a lot from them: how fast can they learn? What helps them to improve their learning performance? Is there an educational model that can be useful to any student or the focus should be on nding di erent learning methods and help

students to nd the one that suits them best?

List of Figures

Bibliography

[1] Rohrer, D. & Pashler, H. Increasing retention without increasing study time. Current Directions in Psychological Science 16, 183 186 (2007).

[2] Bloom, K. & Shuell, T. J. E ects of massed and distributed practice on the learning and retention of second-language vocabulary. The Journal of Educational Research 74, 245 248 (1981).

[3] Fishman, E. J., Keller, L. & Atkinson, R. C. Massed versus distributed practice in computerized spelling drills. Journal of Educational Psychology 59, 290 296 (1968).

[4] Bird, S. E ects of distributed practice on the acquisition of second language english syntax. Applied Psycholinguistics 31, 635 650 (2010).

[5] Rawson, K. A. & Kintsch, W. Rereading e ects depend on time of test. Journal of Educational Psychology 97, 70 80 (2005).

[6] Rohrer, D. & Taylor, K. The e ects of overlearning and distributed practice on the retention of mathematics knowledge. Applied Cognitive Psychology 20, 1209 1224 (2006).

[7] Reynolds, J. H. & Glaser, R. Ee ects of repetition and spaced review upon retention of a complex learning task. Journal of Educational Psychology 55, 297 308 (1964).

[8] Ebbinghaus, H. Memory: A contribution to experimental psychology. (1885).

[9] Küpper-Tetzel, C. E. Understanding the distributed practice e ect: Strong e ects on weak theoretical grounds. Zeitschrift für Psychologie 222, 71 81 (2014).

[10] Cepeda, N.et al. Optimizing distributed practice: Theoretical analysis and practical implications. Experimental Psychology56 (2009).

[11] Storm, B., Bjork, R. & Storm, J. Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retentionMemory Cognition 38, 244 253 (2010).

[12] Son, L. K. & Simon, D. A. Distributed learning: Data, metacognition, and educational implications. Educational Psychology Review24, 379 399 (2012).

[13] Consortium: Universitat Pompeu Fabra, A. B. t. U. o. H., University of Western Macedonia & UAS., M. illuminated (illuminating e ective teaching strategies with the science of learning). URLhttp://www.illuminatedproject.eu/ .

[14] Theophilou, E. Applying distributed learning to a programming course for children. (2019).

[15] Küpper-Tetzel, C. E. Retrieval practice and the maintenance of knowledge. Practical aspects of memory: Current research and issues1, 396 401 (1988).

[16] Roediger, H. & Butler, A. C. The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences15, 20 27 (2011).

[17] Pyc, M. & Rawson, K. Examining the e ciency of schedules of distributed retrieval practice. Memory Cognition 35, 1917 1927 (2007).

[18] Goossens, Net al. Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study: Distributed practice and retrieval practice. Applied Cognitive Psychology30, 5 (2016).

[19] Comission, E. The computational thinking study. URLhttps://ec.europa.eu/jrc/en/computational-thinking .

[20] Guanhua, J.et al. Assessing elementary students' computational thinking in everyday reasoning and robotics programming.Computers Education.109, 162 175 (2017).

[21] Seiter, L. & Foreman, B. Modeling the learning progressions of computational thinking of primary grade students. Proceedings of the ninth annual international ACM conference on International computing education research (ICER '13) 59 66 (2013).

[22] Yee Lye, S. & Ling Koh, J. Review on teaching and learning of computational thinking through programming: What is next for k-12? Computers in Human Behavior 41, 51 61 (2014).

[23] Atmatzidou, S. & Demetriadis, S. Advancing students' computational thinking skills through educational robotics: A study on age and gender relevant di erences. Robotics and Autonomous Systems75, 661 670 (2016).

[24] Ruthmann, A., Heines, J., G., G. R., Laidler, P. & Saulters, C. Teaching computational thinking through musical live coding in scratch.Proceedings of the 41st ACM technical symposium on Computer science education (SIGCSE '10) (2010).

[25] Shin, S. B. The improvement e ectiveness of computational thinking through scratch education. Journal of the Korea Computer Information Society20, 191 197 (2015).

[26] Martínez-Moreno, J. A study on the development of maker activities with primary education teachers and students: from self-concept change to gender factors. Zenodo (2019).

[27] Alamanda, D., Anggadwita, G., Ramdhani, A., Kriseka Putri, M. & Susilawati, W. Kahoot!: A game-based learning tool as an e ective medium to improve students' achievement in rural areas!Opening Education for Inclusivity Across Digital Economies and Societies191 208 (2019).

[28] McDaniel, M. A., Howard, D. C. & Einstein, G. O. The read-recite-review study strategy: E ective and portable. Psychological Science 20, 516 522 (2009).

[29] Muijtjens, A., Van Mameren, H. R., H., Evers, J. & Van der Vleuten, C. The e ect of a don't know option on test scores: number-right and formula scoring compared. Medical Education 267 275 (1999).

[30] Cecilio-Fernandes, Det al. Comparison of formula and number-right scoring in undergraduate medical training: A rasch model analysis. BMC Medical Education 17, 192 (2017).

[31] Upton, G. & Cook, I. Oxford dictionary of statistics - online version. URL https://www.oxfordreference.com/view/10.1093/acref/9780199541454.001.0001/acref-9780199541454.

Appendix A

First Appendix

Questions about session 1

Multichoice

1. El algorismes/codis són. . .

a) Les imatges del ordinador.

b) Els bucles del ordinador.

c) Les instruccions que donem al ordinador.

d) Ninguna de les anteriors.

2. Què passaria si no contactéssim el cable de terra en el makey makey?

a) El circuit estaria obert i no funcionaria.

b) Que el makey makey explotaria.

c) Que ens podem fer mal.

d) Ninguna de les anteriors.

True or false

3. Un algorisme/codi són instruccions pas a pas per a completar una tasca.

4. Un algorisme/codi són tasques que no segueixen cap ordre.

5. El cable de terra en el makey makey permet que la electricitat tingui un camí per tornar (circuit tancat).

Questions about session 2

Multichoice

1. Quina d'aquestes opcions és correcte?

a) x>y signi ca que x és menor que y.

b) x<y signi ca que x és menor que y.

c) x< y signi ca que x és major que y.

d) Ninguna és correcte.

2. Les variables...

a) Poden canviar de valor.

b) No poden canviar de valor.

c) No poden ser números.

d) Ninguna és correcta.

True or false

3. C > D signi ca que C és major que D.

4. Z < Y signi ca que Z és major que Y.

5. Les variables no poden canviar de valor.

Questions about session 3

Multichoice

1. Quin dels 4 codis és una condició?

a) A

b) B

c) C

d) D

2. Què signi ca aquesta condició a Scratch?

a) Si VALOR és major que 50 el personatge dirà ½Adiós!

b) Si VALOR és major que 50 el personatge dirà ½Hola!

c) Si VALOR és menor que 50 el personatge dirà ½Hola!

d) Ninguna de les anteriors

True or false

3. Digues si és una condició o no:

4. Donada la següent condició:

Digues si la següent a rmació és certa o falsa. Si la velocitat és igual 50 s'iniciarà el so Miau

5. Donada la següent condició:

Digues si la següent a rmació és certa o falsa. Si DINERO és major que 50 el personatge canviarà de disfressa

Questions about session 4

Multichoice

1. Quin dels 4 codis és un bucle?

a) A

b) B

c) C

d) D

2. Al fer click a la bandereta verda, la variable DINERO tindrà el valor...

a) DINERO és igual a 2

b) DINERO és igual a 0

c) DINERO és igual a 6

d) DINERO és igual a 4

True or false

3. La següent imatge és un bucle.

4. Al executar el següent codi:

La variable DINERO tindrà el valor 10.

5. Al executar el següent codi:

La variable PATATAS tindrà el valor 5.

Questions about session 5

Multichoice

1. Què és un input?

a) És la informació que li donem al ordinador.

b) És la resposta del ordinador.

c) Són missatges d'error.

d) Son disfresses que podem posar als nostres personatges

2. Al fer clic al botó espai el personatge salta. Quina a rmació és certa?

a) El botó espai és el output.

b) El personatge salta és el input.

c) El botó espai és el input.

d) Ninguna de les anteriors.

True or false

3. Els inputs i els outputs són tipus de bucles.

4. El input és la resposta del ordinador.

5. Si al fer clic al botó espai el personatge salta. Digues si la següent a rmació és certa o falsa: El personatge salta és el output