UTRECHT UNIVERSITY

Governing students in educational practice games: Towards more independent and effective learning

by

Wessel Cordes

Submitted in partial fulfillment of the Master Business Informatics

in the
Faculty of Science
Department of Information and Computing Sciences

July 12, 2018

Abstract

Utilizing online digital educational content has become the norm when teaching young students. Adaptive educational practice systems such as Math Garden allow students to practice their arithmetic abilities in various domains on a preferred difficulty and pace. However, due to the intensification of the teaching profession, students are often left unsupervised, and as a result do not practice domains that are most important. Therefore, this thesis proposed governing as a solution to increase student abilities. Governing is defined as computerized data driven supervision that guides students in practicing domains most important without intervention of a teacher. Since no prior research on this topic was performed, Math Garden was studied to assess ways of governing.

First, governing was conceptualized and exhaustively assessed with examples. Next, A solution governing method was developed in Math Garden. This governing method calculated and selected three domains a student should practice each day. The governing method was evaluated in an A/B test running for two full weeks. A total of 13 578 students participated in the experiment. 6 785 students were in control variant A(default system) and 6 793 were treated with the solution governing method (variant B). The solution governing method was found to have positive effects on both engagement and ability. Students willingly practiced selected domains and a significant increase in domain abilities were found. Therefore, governing was effectively introduced in Math Garden. Concluding, this thesis explored governing and provides the first steps, knowledge and reasoning to introduce governing in educational practice systems.

Acknowledgements

Nine months ago I started my internship at Oefenweb, to conduct the longest research project of my life thus far. I am proud of what has been achieved during this time. Of course, I would not have been able to complete this project alone. Therefore, I would like to thank everyone who helped me in this project. Particularly, I would like to thank my supervisors Matthieu Brinkhuis and Abe Hofman for their guidance and pleasant communication. Finally, I would like to thank my Oefenweb colleagues with whom I had many enjoyable lunch breaks, hacky sack moments, game nights and drinks.

I hope you enjoy reading my Master thesis.

Thank You!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Using the internet as a source of information has become the standard in western culture. Likewise, in education it has become common practice to use online digital course material in classrooms to support teachers in educating students. So-called online, web-based or mobile learning makes it possible to access content and learn anywhere, anytime. Where traditional teaching in primary and secondary education mainly involved pen and paper, books, blackboards and written assignments. Now most elementary students in western culture have access to PC's, laptops or tablets in classrooms and use these devices to utilize course material, complete assignments, finish homework and connect to the internet. This shift in teaching creates opportunities for companies to take advantage of. For example, traditional Dutch learning methods (Malmberg, 2018; Noordhoff, 2018; ThiemeMeulenhoff, 2018) have each digitized their material into online portals connected to an easy-access hub (Basispoort, 2018). The portals include digital assignments, teaching videos, automated assessment, digital exams and individual progression tracking. The hub provides centralized authentication for students. Momento (2018) and ParnasSys (2018) provide a centralized interactive dashboard for teachers connected various learning materials. Others have followed a more innovative approach. Squla (2018) combines gaming and learning to make learning fun and interesting for children. Snappet (2018) offers affordable tablets for schools together with their own learning software, allowing for individual digital devices in classrooms. Dreambox, ALEKS, Knewton and Math Garden (Lemke, 2014; Falmagne, Cosyn, Doignon, & Thiéry, 2006; Wilson & Nichols, 2015; Klinkenberg, Straatemeier, & van der Maas, 2011; Straatemeier et al., 2014; Brinkhuis et al., 2018) took a data driven approach and implemented adaptive learning into their product, enabling students to learn at a preferred difficulty.

This thesis is written with the latter. Math Garden originated in 2007 as a tool to study the dynamics of cognitive development in children, specifically the development of mathematical knowledge and abilities. Brinkhuis et al. (2018) regard Math Garden as a large-scale online learning system, and analyzed more than a decade worth of data. Fundamentally, Math Garden is a computerized adaptive practice (CAP) system aimed towards primary education. The adaptive element is a modification of the Item Response Theory (IRT) approach and is based on the Elo (1978) rating system used in chess combined with the High Speed, High Stakes (HSHS) scoring rule (Maris & Van der Maas, 2012; Coomans, Hofman, Brinkhuis, van der Maas, & Maris, 2016) The model is able to estimate the ability of the student and the difficulty of the item (i.e. question). The estimation is updated after every answered item, allowing for on the fly calibration. In 2009, Math Garden was commercialized as Oefenweb and other systems were introduced with the same practicing and monitoring capabilities using the same adaptive learning model, namely: "Language Sea," a system for learning the Dutch language. "Words & Birds," a system for learning the English language. "Typetuin," a system to learn typing (van den Bergh, Schmittmann, Hofman, & van der Maas, 2015). "Rekenjebeter," a system aimed at nurses for learning medical arithmetic. Each system consists of multiple games (e.g. Math Garden consists of addition, subtraction, division, etc.). Each game trains a specific ability using open and/or multiple choice questions that need answering within a limited amount of time. As of 2017, more than 2000 Dutch primary schools use Oefenweb to practice arithmetic, Dutch and English, completing roughly a million items every day.

1.1 Problem Statement

Adaptive learning systems have some clear advantages over traditional methods. Presumably, they provide adaptive content to enable students to learn at a preferred difficulty without intervention of a teacher or parent. However, in the context of educational practicing systems, a system is divided into many practicing domains or games, each providing adaptive content. The student has the ability to practice a domain of choice. This option of choice can lead to unwanted behavior in a practice system, since in most countries student ability is assessed using high-stakes tests. A high-stakes test is any test used to make important decisions about students, teachers and schools (Earl, 2012). Dutch students are also assessed using high-stakes tests. In primary education this is mostly done with Cito tests (Cito, 2018). Tests are taken at the end of each grade, with the final tests taken in the 6$^{\text{th}}$ grade (age 11-12). Questions are about Language, Arithmetic, Study Skills and World Orientation. Cito test results are the basis of students ability, and future decisions involving secondary education are primarily based on its outcomes (van der Lubben, n.d.). Educational practice systems support students in practicing domains of choice, supposedly providing the necessary knowledge for high-stakes tests.

In Math Garden new domains are released regularly and students are overwhelmed with choice. Math Garden consists of 23 domains. Not all 23 domains are immediately playable, since some domains are unlocked after reaching a certain level in another domains or after reaching a certain age. Still, most students have access to 12+ games. While there is some form of control from the system over what domains are available, there is no supervision over what domains should be practiced. Currently, the only supervision comes from a person with authority (i.e. teacher or parent), who instructs a student what to do, and, checks whether the action is completed. According to Oefenwebs' data scientists this rarely happens. They report that one of the reasons Math Garden is utilized is to provide teachers with time for students who need it most. This means most students are left without supervision, and are therefore able to practice whatever they want. In Math Garden this has led to unwanted behavior. Exploratory Data Analysis (EDA) revealed unexpected results in domain popularity. Around 7% of students age 10-12 still practice domains targeted towards preschoolers (Appendix A.2). These domains involve topics such as numbers, counting and shapes. Also, some domains are barely practiced (Appendix A.1). Furthermore, EDA suggest most students have favorite domains. 20% of students mainly practice three domains. This is not desirable, since students need to practice all domains to acquire necessary knowledge for high-stakes tests. This information indicates there is a strong need for guidance from the system in Math Garden.

This thesis will provide a solution to help guide students in educational practice systems, which we call governing. Since there is no prior research on this topic, Math Garden is studied to assess ways of governing students in practicing domains. This research will explore and provide the initial steps and reasoning helping to introduce governing in other educational practice systems. While Oefenweb has several applicable systems that could benefit from governing, this research is not limited to Oefenweb products.

In Math Garden a governing method or so-called Governer, is developed and implemented. Fundamentally, the governing method selects the best possible domains. It is tailored to the individual and is available for elementary students. The idea is to present a student with a set of domains which should be practiced to increase a student's overall ability. Domains are selected daily. Domains are selected based on how students perform against both, grade specific educational goals (SLO, 2018), peers and Math Garden student population. Also, several visualizations are hypothesized and a feasible visualization is developed. A possible implementation is to present the governing method by simply disabling remaining domains. Yet, the researchers believe this would have a negative effect on motivation. According to Malone and Lepper (1987), control over an activity significantly improves motivation and academic performance. Introducing a governing method would decrease the amount of control of the student, hence decrease his/her motivation. Fortunately, introducing gamification elements could negate the negative effect on motivation (Zichermann & Cunningham, 2011). Concluding, this thesis assesses ways in Math Garden to govern practice and increase students abilities. Therefore, the main research question is as follows:

> *"MRQ. How to increase student abilities, by assessing ways of governing practice in educational systems, applied to Math Garden?"*

1.2 Thesis Outline

This thesis encloses the entirety of a thesis project on ways of governing practice in Math Garden. Therefore, a brief outline is presented in this section.

In Chapter 2, The literature is reviewed explaining the scientific background. Chapter 3 defines research objectives and aims. Moreover, research questions are presented and explained. Chapter 4 explains the research methods chosen for this research. In Chapter 5, governing is conceptualized and the governing solution is described. Chapter 6 explains the experiment conducted to test the governing solution. In Chapter 7, results of the experiment are described. Finally, in Chapter 8, results are discussed and a conclusion is provided.

Chapter 2

Literature Review

In this section, the literature is thoroughly reviewed. First, Intensification and Adaptive learning are explained. Second, Math Garden and its underlying technologies are reviewed. Lastly, Motivation and Gamification are outlined.

2.1 Intensification

During the last decades the teaching profession has undergone multiple changes, often referred to as intensification or depersonalization (Apple, 2013). Teachers are spending more time at work, as well as more time at home working on administrative activities (Bullough, Hall-Kenyon, MacKay, & Marshall, 2014). With increasing intensification, teachers struggle to balance home and work obligations (Kelliher & Anderson, 2010). Ultimately, intensification worsens teaching quality, which in turn has a negative outcome on students and their learning (Williamson & Myhill, 2008). The intensification of teaching is the result of teachers being increasingly subjected to external pressures and demands from policymakers, supervisors, experts and parents (Van Droogenbroeck, Spruyt, & Vanroelen, 2014). For example, in 2014 Dutch policymakers decided primary schools have care duties for children with autism and other handicaps, which in turn led to an increased workload (van Grinsven & van der Woud, 2016). In the Netherlands the ever increasing intensification combined with unfair compensation has led to great dissatisfaction among teachers. Research suggest introducing proper ICT in classrooms could help alleviate teacher workload (Selwood & Pilkington, 2005). In recent years many new technologies have emerged in education and teachers are coming to realize the benefits.

Through the digital transformation the teaching profession has changed. Newman (2017) identifies digital transformation trends in education. Newman's identification of augmented- virtual- and mixed reality has yet to be widely adopted in classrooms. While his others trends are more widely adopted. One of the most important transformations is the increase of digital devices in classrooms. Students no longer have to go to technology labs/rooms to access computers, but have access to a devices in class. Also, learning spaces have been redesigned. For example, classrooms have smart-boards allowing for more interaction and creativity. Furthermore, more educational material use gamification. Which combines learning with gaming elements in classrooms and transforms difficult educational subjects into something more exiting and interactive. Lastly, other transformations in education are the use of Artificial Intelligence (AI) and the introduction of personalized learning, which is umbrella term of education tailored to suit each individual. Adaptive and blended learning are examples of personalized learning.

2.2 Adaptive learning

Adaptive learning is one of the technologies that have emerged in education as mentioned in the previous section. Adaptive learning is not new, it has its origins in the AI movement and began gaining popularity is the 70's (Sleeman & Brown, 1982). In recent years more companies have included adaptive learning into educational systems. Gartner even highlighted adaptive learning to be the number one strategic technology in education in both 2015 Gartner (2015a) and 2016 Gartner (2015b), while assigning the 4th place in 2017 (Gartner, 2016). Yet, what exactly is adaptive learning? It has different meaning in various contexts. In the context of education many definitions in the literature were found. Each definition is different, yet there is a consensus among researchers. A few notable definitions:

> Gartner (2018)): "Adaptive learning in its fundamental form is a learning methodology that changes the pedagogical approach toward a student based on the student's input and a predefined response. Adaptive learning more recently is being associated with a large-scale collection of learning data and statistically based pedagogical responses and can be seen as a subset of personalized learning that includes such approaches as effective and somatic computing."

> Newman, Bryant, Fleming, and Sarkisian (2016): “Solutions that take a sophisticated, data-driven, and in some cases, non-linear approach to instruction and remediation, adjusting to each learner’s interactions and demonstrated performance level and subsequently anticipating what types of content and resources meet the learner’s needs at a specific point in time.”

> Paramythis and Loidl-Reisinger (2003): “A learning environment is considered adaptive if it is capable of: monitoring the activities of its users; interpreting these on the basis of domain-specific models; inferring user requirements and preferences out of the interpreted activities, appropriately representing these in associated models; and, finally, acting upon the available knowledge on its users and the subject matter at hand, to dynamically facilitate the learning process.“

To summarize, adaptive learning is a data driven teaching method that monitors and adjusts educational content for each individual to their unique needs based on responses to previous content. The latter definition is adopted in this thesis, since the notion of monitoring activities and domain-specific models are recognizable in Math Garden.

2.2.1 Adaptive Learning systems

As mentioned, there are many adaptive learning systems. Yet, not all are worth mentioning. Brinkhuis et al. (2018) argue that online learning systems are only sustainable through deliberate research. Thus, in scope of this thesis a few successful systems are explored that have solid scientific background.

DreamBox (Lemke, 2014) is an adaptive system for learning math, aimed at children age 5-14. It utilizes something called Intelligent adaptive learning (IAL) to optimize learning by establishing a digital learning environment that keeps students in their optimized learning zone. It captures every decision a student makes and adjusts the student’s learning trajectory both within and across lessons. With artificial intelligence, specifically deep learning, the IAL system attempts to identify the psychological cause of mistakes. Intelligent feedback induces reflection and rethinking by the student, and, therefore lowers the probability the mistake will happen again (Lemke, 2014). Research shows students using DreamBox scored 2.3 points higher on the Northwest Evaluation Association assessments with gains equivalent to 5.5 percentile point in 16 weeks (Wang & Woodworth, 2011).

ALEKS or Assessment and LEarning in Knowledge Spaces is a Web-based, artificially intelligent assessment and learning system (Falmagne et al., 2006). It focused on math and arithmetic and is aimed at both lower and higher education. ALEKS is the practical realization of knowledge space theory (Doignon & Falmagne, 2012) and uses adaptive questioning to quickly and accurately determine exactly what a student knows and does not know in a course. ALEKS instructs the students on the topics they are most ready to learn. As a student works through a course, topics learned are periodically reassessed to make sure they are retained.

Knewton offers affordable, adaptive course solutions in higher education. Knewton's model is based on modifications of the IRT framework, coupled with a knowledge graph and sufficient student interaction data. This enables both productive interpretation of student interaction data, and tractable real-time inference computation (Wilson & Nichols, 2015). A study assessing the effects of adaptive assignments on student performance concluded better student performance on average in courses that used Knewton compared to those that did not. The improvements increase with more use of Knewton's adaptive assignments. The researchers observed a peak average score difference of four percentage points (Bomasch & Kish, 2015).

Math Garden (Klinkenberg et al., 2011; Straatemeier et al., 2014) originated in 2007 as a tool to study the dynamics of cognitive development in children, specifically the development of mathematical knowledge and abilities. Math Garden is system for practicing and monitoring arithmetic in primary education. Math Garden works using a modified Computerized Adaptive Testing (CAT) model in Item Response Theory (IRT). The modification is a combination of the Elo (1978) rating system used in chess with the High Speed, High Stakes (HSHS) scoring rule (Maris & Van der Maas, 2012). The model is able to estimate the ability of the user and the difficulty of the item (i.e. question). The estimation is updated after every answered item, allowing for on the fly calibration.

2.3 Math Garden

Math Garden is the educational practice system focused on in this thesis, therefore it is important to have an understanding of what Math Garden is, how it works and what underlying technologies establish adaptive learning. Thus, in this section, an overview of Math Garden is presented, followed by its adaptive learning technologies, namely: Item Response Theory & Computerized Adaptive Testing, Elo Rating System and the High Speed, High Stakes scoring rule.

2.3.1 Overview

Math Garden is a practice systems, mainly aimed towards children age 3-12 to practice arithmetic. It is a web application built with JavaScript, accessible via web-browser on both PC and tablet. After logging in students land on a personalized page, containing a garden with distinct plants, each representing a mathematical domain (Figure 2.1). Plants thrive if the domain is practiced, yet wither if not practiced. Every domain is playable as a game, where students are presented with items (Figure 2.2). An item is in open or multiple-choice format. Open format items are usually answered via a digital numeric-keypad, yet multiple-choice items have 4 or 6 answers which need selecting. By pressing the question mark, students are able to skip the item. The system adaptively matches a students with an item so that 75% of items are answered correct. The default setting is 75%, though this setting is adjustable to 60% (hard) or 90% (easy) on the landing page. After every item the score is updated. A students' ability is translated into a Quantile (Q) score intended to increase readability for students and teachers. The Q-score is based on all items that have been answered and is calibrated yearly. The Q-score has a value between 0-1000. For example, a student that can correctly answer 55% of items has a relative score of 550. Further explanation of the adaptive model and Q-score can be found in Section 2.3.2 and 2.3.3. In addition, correctly answered items are rewarded with coins. Faster responses are rewarded with more coins. After a session, comprising of 10 or 15 items depending on the domain, additional coins are rewarded, subsequently returning to the landing page. Explicit details about the scoring rule is explained in Section 2.3.4.

Furthermore, on the landing page a student has the option to navigate towards other pages using click-able buttons: The bonus-garden contains extra domains and is only accessible when primary domains on the landing page are practiced regularly. In the trophy cabinet, students can buy several ribbons, medals and trophies in exchange for acquired coins during practice. The grow chart displays played games in descending order based on domain score. It also displays the number of items ever completed and today. Lastly, there is an info page with instruction videos and explanations.

2.3.2 IRT & CAT

Item Response Theory (IRT), also known as Latent Trait Theory or Modern Mental Test Theory, is a popular approach in psychometrics to analyze responses to tests or questionnaires with the goal of improving measurement accuracy and reliability. First introduced by Lord and Novick (2008), the main difference with classical test theory's is the focus on individual items (i.e. questions), instead of focusing on the test as a whole.

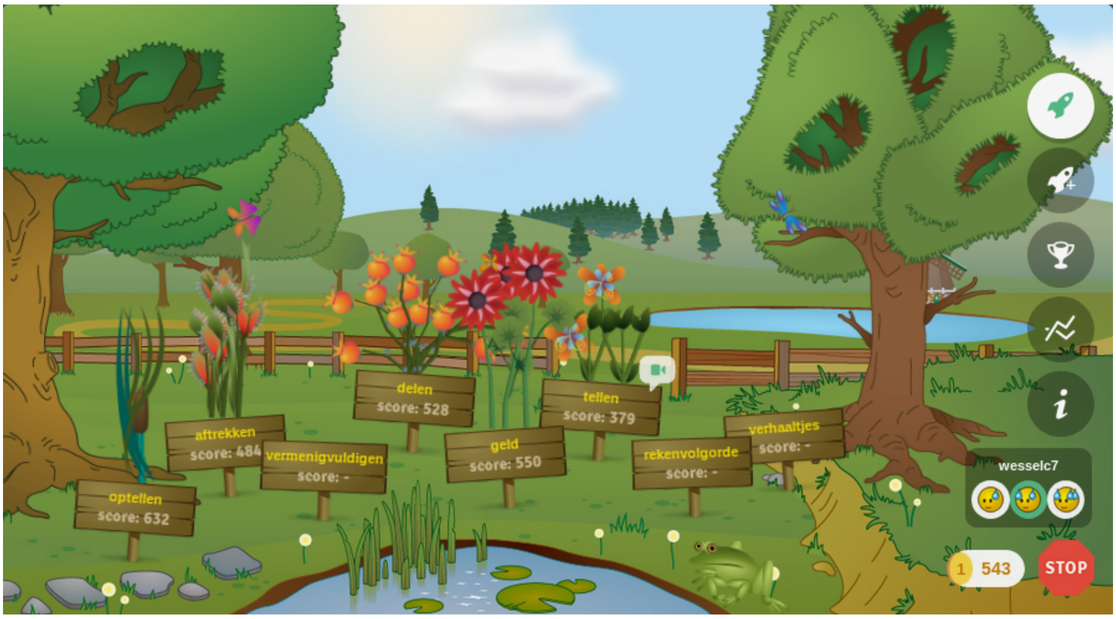


FIGURE 2.1: Math Garden landing page

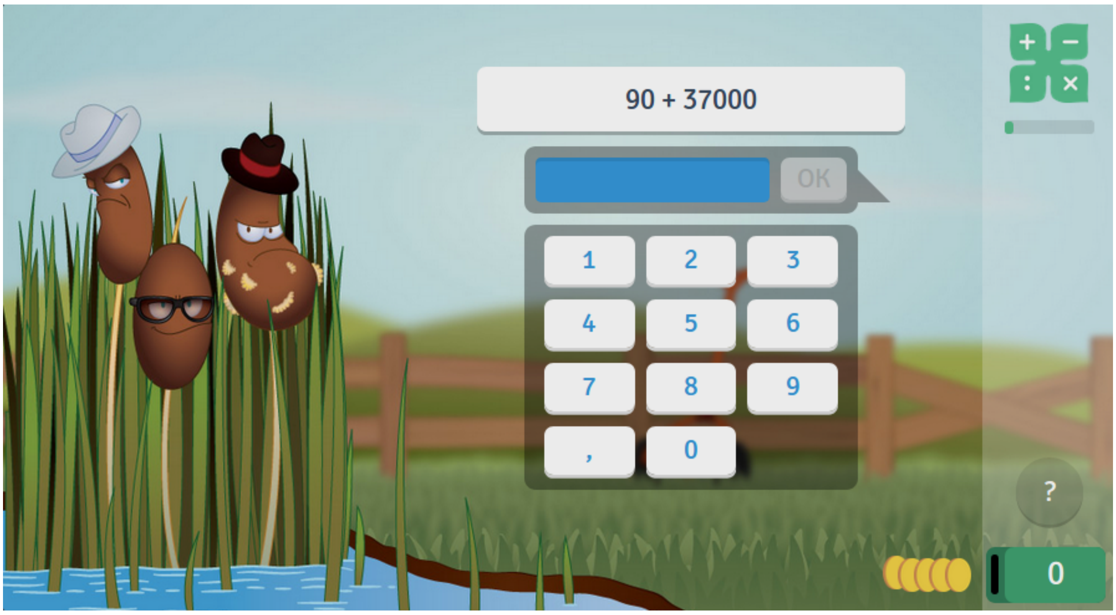


FIGURE 2.2: Typical Math Garden item in open format

The focus on individual items allows for item banking (Choppin, 1968), which means candidates can be given a completely different set of items, without losing test accuracy and reliability. This minimizes the effects of cheating, since there are no fixed tests. The IRT model is defined by item parameters (Embretson & Reise, 2013). The parameters are able to estimate both student ability and item difficulty (van der Linden & Hambleton, 2013). The 1 Parameter Logistic model (1PL), also known as the Rasch model (Rasch, 1960), uses item difficulty as a parameter for calculating a candidate's ability. The 2 Parameter Logistic model (2PL), uses both item difficulty and item discrimination as parameters, where the item discrimination parameter measures underlying psychological constructs. The 3 Parameter Logistic model (3PL) uses both item difficulty and item discrimination parameters, extended with a guessing parameter, which accounts for correctly guessed items. Item response models can be used for equating tests, to detect and study differential item functioning (bias) and to develop Computer Adaptive Tests (CAT) (van der Linden & Hambleton, 2013). CAT dynamically adjusts the difficulty of items to each individual candidate, thus providing tailored tests. CAT presents a candidate with an item from the item bank and, when answered correctly (incorrectly), presents a more (less) difficult item. Originally, CAT was developed for measurement only, and, using CAT could shorten tests lengths up to 50% (Eggen & Verschoor, 2006). Klinkenberg et al. (2011) extended CAT into a Computerized Adaptive Practice (CAP), namely Math Garden. By applying a new estimation method based on the Elo (1978) rating system (ERS) used in chess combined with an explicit scoring rule.

2.3.3 Elo Rating System (ERS)

As mentioned adaptive selection in Math Garden is based on ERS (Elo, 1978). ERS dynamically changes the abilities of chess players expressed in Elo ratings, with higher Elo translating to a better chess player. In ERS, chess players are matched against other chess players and Elo ratings are updated after every match. The winner gains Elo and the defeated loses Elo. Yet, the amount of Elo gained or lost is based on the difference in initial Elo ratings between players. The bigger the difference the higher the Elo rating consequences and vise-versa. The mathematical equation of ERS can be found in Equation 1. θ represents the ability or Elo Rating and $\hat{\theta}$ the updated Elo rating. The updated Elo rating depends on the weighted difference in match result S (1, 0.5 or 0 for win, draw and loss) and expected match result E. E expresses the probability of winning and the K factor, proposed by Elo, is a value based on rating and number of played matches, that determines the maximum θ can change from a single match. Glickman (1995) suggests that the original K factor does not always rate Elo ratings accurately. When players are new or have not played for a long time, their Elo rating might not represent their ability. Glickman proposes a K factor function that incorporates recently played and playing frequency. The K factor is low for recent and frequent players, yet high for new and absent players. An algorithm based on ERS can provide a means to estimate dynamic ratings that involves paired comparisons. Hence, it is suitable for application in an educational context where item responses are regarded as student-item paired comparisons. Furthermore, student and item difficulties are expected to change over time (Brinkhuis, Bakker, & Maris, 2015; Klinkenberg et al., 2011; Pelánek, 2014; Wauters, Desmet, & Van Den Noortgate, 2010).

$$\begin{aligned}\hat{\theta}_j &= \theta_j + K(S_i - E(S_j)) \\ \hat{\theta}_k &= \theta_k + K(S_k - E(S_k))\end{aligned} \tag{2.1}$$

$$E(S_j) = \frac{1}{1 + 10^{(\theta_j - \theta_k)/400}} \tag{2.2}$$

In order for ERS to work in Math Garden it required some modification. A student is not matched against another student (θ), but against an item (β). Responding to an item is considered a match. A student wins the match when correctly answering the item and loses the match when incorrectly answering. Both items and students have Elo ratings and are updated after every match. This means a student is matched to more (less) difficult items when gaining (losing) Elo, and, an item becomes more (less) difficult when answered incorrectly (correctly). Furthermore, a K factor function as described by Glickman (1995) is introduced depending on uncertainty U. It is assumed that after 30 days (D) of not playing a student or item reaches the maximum uncertainty of one. Yet, 40 responses reduces uncertainty to the minimum of zero. The student and item ratings are presented as real numbers. To improve readability for students, student-ratings are transformed into a Quantile score. The score represents the percentage of items a student can correctly answer. Any item that is answered by any student is included. The score visible to students is the percentage multiplied by ten.

$$\begin{aligned}\hat{\theta}_j &= \theta_j + K_j(S_{ij} - E(S_{ij})) \\ \hat{\beta}_i &= \beta_i + K_i(E(S_{ij} - (S_{ij}))\end{aligned} \tag{2.3}$$

$$\begin{aligned}K_j &= K(1 + K_+U_j - K_-U_i) \\ K_i &= K(1 + K_+U_i - K_-U_j)\end{aligned} \tag{2.4}$$

$$\hat{U} = U - \frac{1}{40} + \frac{1}{30}D \tag{2.5}$$

2.3.4 High Speed, High Stakes scoring rule

The other extension of CAT is the appliance of a scoring rule. The so-called High Speed, High Stakes (HSHS) scoring rule used in Math Garden introduced by Maris and Van der Maas (2012). The rule imposes a speed-accuracy trade-off setting on the individual. The mathematical representation of the scoring rule is in Equation 6. x_ij is the response of student j on item i in time t_ij before time limit d_i. The score S_ij is scaled by discrimination parameter a_i. A correct response translated to $x_ij = 1$, and an incorrect response to $x_ij = 0$. If the response is correct, the score equals the remaining time. If the response is incorrect the remaining time is multiplied by -1. Therefore, a fast incorrect response contributes to high negative score (-1 multiplied by the remaining time). Hence, the high speed, high stakes. The updated probability formula E integrating HSHS is in Equation 7. The formulas together (Equation 1, 6, 7) allow for on the fly calibration and are the basis of adaptive learning within Math Garden.

$$S_{ij} = (2x_{ij} - 1)(a_i d_i - a_i t_{ij}) \tag{2.6}$$

$$E(S_{ij}) = a_i d_i \frac{e^{2a_i d_i(\theta j - \beta i)} + 1}{e^{2a_i d_i(\theta j - \beta i)} - 1} - \frac{1}{\theta j - \beta i} \tag{2.7}$$

2.4 Student Motivation

In Section 1.1 decrease in motivation is outlined as a potentially negative side effect of introducing a Governing in Math Garden. In this section the notion of motivation is elaborated as well as the possible (negative) effects introducing a governing method in Math Garden might have. Student motivation is not be confused with student engagement. Some researchers use the terms interchangeably (Martin, 2007). Yet, in this thesis the researcher believe that motivation and engagement are distinct, related constructs wherein motivation represents intention and engagement represents an action (Christenson, Reschly, & Wylie, 2012).

One of the most important psychological concepts in education is certainly that of motivation (Vallerand et al., 1992). According to Malone and Lepper (1987), motivation is a necessary precondition for student involvement in any type of learning activity, and what and how effectively students learn may be influenced by their level of motivation.

For this thesis, Malone and Lepper's taxonomy of intrinsic and extrinsic motivation for learning is used, validated by Ciampa (2014). Their work is focused on what makes games both fun and educational, which is in line with the aim of this thesis. The theory is based on six categories of individual motivations that make an activity both intrinsically and extrinsically motivating, namely: challenge, curiosity, control, cooperation, competition, recognition. Intrinsic motivation means the learner engages in an activity because it is interesting or enjoyable, yet extrinsic motivating means the learner engages in an activity because of a desired outcome and wanting to achieve some instrumental end such as earning a reward (Vallerand et al., 1992).

Intrinsic motivations

Challenge

A learner is more motivated when goals are clearly defined and when challenge is balanced. An activity should neither be too easy and incite boredom, or too difficult that succeeding seems impossible. There are various ways to obtain the optimal level of challenge. Activities should have varying difficulty levels of instruction, provide multiple levels of goals, have varying time constraints, provide incomplete information and make the learner look for missing elements. Furthermore, performance feedback (e.g. score) allows for progression tracking towards objectives. Objectives must be meaningful and tailored to the individual.

Curiosity

Curiosity is the most direct intrinsic motivation for learning. Curiosity can be divided in sensory curiosity and cognitive curiosity. Sensory curiosity involves attention through various sensory stimuli such as light (e.g. video), sound (e.g. music) and touch (e.g. haptic feedback). Cognitive curiosity is induced when learners have the desire to explore and obtain new information and competences when they discover their knowledge is incomplete or inconsistent.

Control

The concept of control is important for intrinsic motivation. Control is associated with motivation when students are given control over their learning. The sheer illusion of control significantly improves motivation and academic performance. Control is best promoted when an activity promotes a sense of personal control and meaningful outcomes and provides independence and versatility. Furthermore, when students are provided with opportunities to make choices about their learning, task engagement increases (Ryan & Deci, 2000).

Extrinsic motivations

Cooperation

Cooperation is generally defined as involving a group of individuals working together to attain a common goal. Cooperation should facilitate performance, especially when individuals have common goals and promotes effort and productivity. Also relationships among individuals and psychological adjustment improves.

Competition
Competition is generally defined as involving two or more people with opposing goals. Competition can be divided in direct competition and indirect competition. Direct competition is comparing one's self against others. It involves individuals or groups competing against each other. Goals are performance oriented (i.e. winning, rather than playing well), thus involve extrinsic motivation. On the other hand, indirect competition involves individuals or groups competing against one's self such as a previous performance to improve or acquire skills. Goals are mastery oriented, thus associated with intrinsic motivation.

Recognition
Generally, recognition is associated with the enjoyment of having efforts and accomplishments recognized and appreciated by others. Recognition can be realized in several ways. One way is to make engagement in an activity visible to others. Another is to make the outcome of an activity visible to others.

2.5 Gamification

The introduction of a governing method into Math Garden or educational practice systems in general might raise unwanted side effects. Gamification is recognized as a possible solution to compensate for these negative effects on motivation. Gamification desires to raise intrinsic motivation by combining it with an extrinsic motivator. "When done well, gamification helps align our interest with the intrinsic motivation of our players" (Zichermann & Cunningham, 2011). Gamification is defined as: the use of game design element in non-game contexts (Deterding, Dixon, Khaled, & Nacke, 2011). Gamification is an old topic in Human Computer Interaction (HCI) and its main goal is to raise user engagement by using game-like techniques such as scoreboards and personalized fast feedback (Flatla, Gutwin, Nacke, Bateman, & Mandryk, 2011). Most prominently, gamification has been commonly associated with point, levels and leaderboards (Hamari, Koivisto, & Sarsa, 2014), yet there are more gamification mechanics and elements useful for specific user types (Figure 2.3).

According to Muntean (2011), engagement is the important metric for success in gamification. In web-applications engagement can be analyzed in analytics such as: page views per visitor, time spent on site, total time per user, frequency of visit, participation

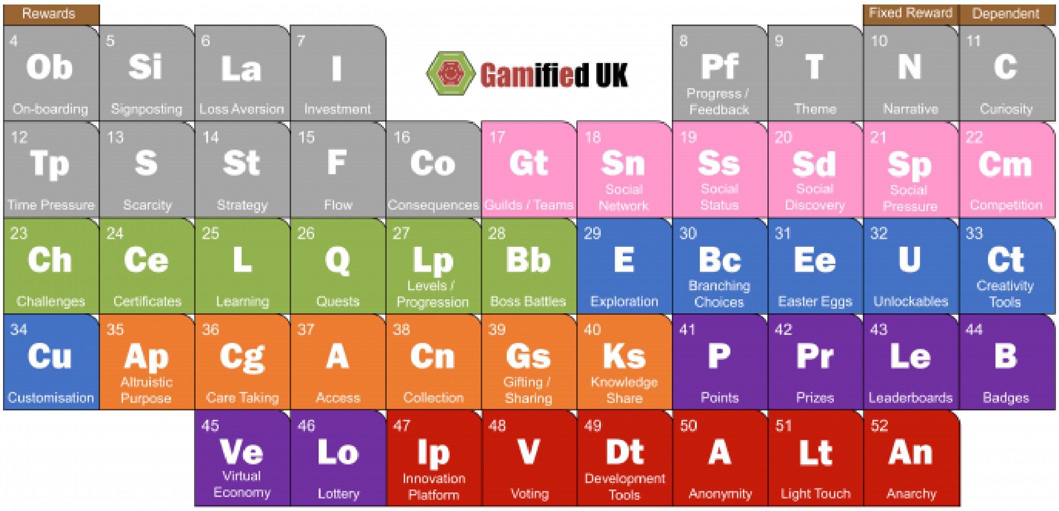


FIGURE 2.3: Gamification mechanics and elements (Gamified, 2018)

and conversions. Lastly, Chen, Liao, Cheng, Yeh, and Chan (2012) researched if it is better to use plain text or Non-Player Characters (NPC's), for example pets, to present quests in a math learning game. They concluded using NPC's to deliver quests resulted in increased enjoyment, clearer goals and stronger goal intensities. These findings could be useful in this thesis.

Chapter 3

Aims & Objectives

The ultimate objective of this thesis is to design a means by which relevant stakeholders are supported in adding governing into educational practice systems. This means can take form of a model, conceptual framework, guideline document, step-wise approach or a combination. The goal of governing is to increase student abilities, by guiding students in practicing educational content, without the need for intervention of a teacher or parent. This means is derived from the process of introducing governing in Math Garden.

We want to understand which potential governing models are suited for Math Garden. Furthermore, potential governing visualizations are hypothesized and a feasible governing visualization is selected. Moreover, we also want to understand if introducing a governing model has an effect on student engagement and ability. Therefore, the effect on student engagement and ability need to be evaluated. The overall objectives mentioned above translate into the following research question:

*"**MRQ.** How to increase student abilities, by assessing ways of governing practice in educational systems, applied to Math Garden?"*

3.1 Research Questions

To help answer the main research question, several sub-questions have been formulated. These sub-questions are as follows:

SQ1. *What are potential governing methods in adaptive educational systems?*

The first question explores potential governing methods and how these can be effectively used in educational practice systems. Therefore, governing is conceptualized and exhaustively assessed with examples. The aim of this question is to form a knowledge base of potential governing methods, which is necessary for following research questions.

SQ2. *What governing methods are present in Math Garden, and what are the effects on student behavior?*

In the second question, governing methods currently present in Math Garden are identified and evaluated. It is important to know if these methods actually do what was intended. The identification and evaluation process are described and lessons learned are added to the knowledge base from the previous research question.

SQ3. *What is an effective, feasible governing method in Math Garden?*

The third question builds further on the knowledge provided in the previous questions. It explores potential governing models and visualizations effectively usable in Math Garden. Subsequently, the solution governing method is developed and described. This process acts as a first standard and should provide the first steps and knowledge to introduce governing in other educational practice systems.

SQ4. *How can the governing method's effectiveness be measured and validated?*

The fourth question, presents a explanation how to adequately measure and validate the governing method. The solution governing method is tested in an experiment to what measures the effects on student ability and engagement.

3.2 Conceptualization

Conceptualization is the process of defining the agreed meaning of ambiguous terms used in a study. Thus, in this section the concepts used in this thesis are defined and explained.

A *Student ability* implies to what extent a student has knowledge in some topic. In context of this thesis a topic is a domain, and the degree of domain knowledge is projected in a domain-rating. Thus, *Student abilities* imply to what extent a student has knowledge in several domains.

Governing is the concept of computerized supervision within an educational practice system. It is the process of guiding a student in such a way he/she learns at least educational material necessary at the end of each grade, without intervention of a teacher or parent.

With a *governing method*, we mean every computerized mechanism that guides a student in an educational practice system. A governing method is the entire governing solution comprised of a governing model and governing visualization.

A *governing model* is the data model itself. In the context of this thesis, this is a regularly updated model that calculates the best possible domains for a student to practice. A governing model is combination of perspectives (Section 5.1.1) and strategies (Section 5.1.2).

A *governing visualization*, is the way in which the outcome of the governing model is presented to students. A visualization can be very strict, meaning a student is only able to practice the domain selected by the governing method. Yet, the implementation could also be more flexible in the form of a recommendation. Governing visualizations are explicitly explained in Section 5.1.3.

Chapter 4

Research Method

In this section the research method for this thesis is defined. This research process is structured according the design science method (Von Alan, March, Park, & Ram, 2004; Wieringa, 2014). The aim of design science is to design an artifact that solves a problem for a certain set of stakeholders in a specific domain. Wieringa's design science method is called the engineering cycle. It is a rational problem-solving process and consists of five steps (Figure 4.1). First, problem investigation identified what must be improved, while taking stakeholders and goals into account. Treatment design consists of specifying requirements that contribute to goals, identifying available treatments (i.e. artifacts) and designing new ones. In treatment validation desired effects of the designed artifacts are analyzed and if these satisfy the requirements. Next, the treatment is implemented if the desired effects are met. Lastly, the implementation is evaluated to see whether the artifact has resulted in the desired goals and effects. The latter may be the start of a new iteration in the design cycle. Since, the engineering cycle provides logical structure of tasks that does not prescribe a rigid sequence of activities (Wieringa, 2014), it allows for a iterative (agile) development method. Instead of one sequence, many sequential passes through the cycle are made. It is even allowed to have parallel iterations cycles.

The identified problem in this thesis, as stated in Section 1.1, is the strong need and desire for more independent students, while learning more effectively. The ultimate goal is to provide relevant stakeholders (developers, data-scientists and data-analysts) with means to introduce governing in educational practice systems. The idea is that governing results into a system that realizes more independent learners and allows for more effective learning.

In the Treatment design step, the eventual artifact of this thesis is a means derived from the development process of a governing method in Math Garden. This process provides the necessary steps, reasoning and knowledge to introduce governing in other

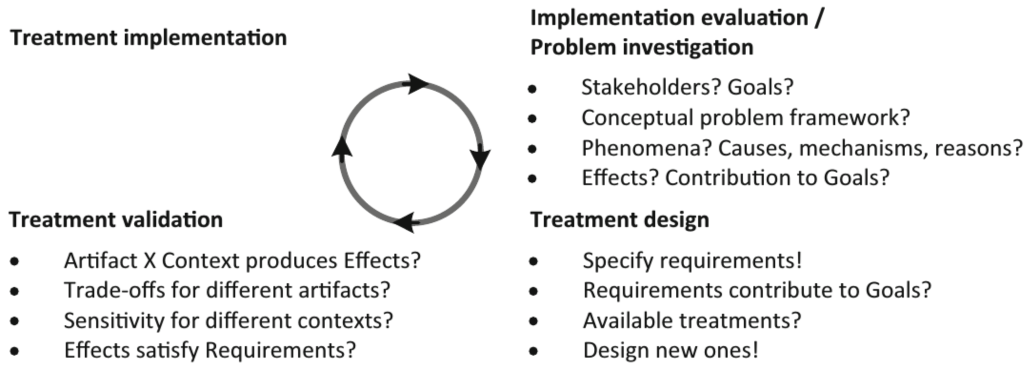

Treatment implementation
Implementation evaluation /
Problem investigation
• Stakeholders? Goals?
• Conceptual problem framework?
• Phenomena? Causes, mechanisms, reasons?
• Effects? Contribution to Goals?
Treatment validation
• Artifact X Context produces Effects?
• Trade-offs for different artifacts?
• Sensitivity for different contexts?
• Effects satisfy Requirements?
Treatment design
• Specify requirements!
• Requirements contribute to Goals?
• Available treatments?
• Design new ones!

Figure 4.1: The engineering cycle (Wieringa, 2014)

educational practice systems and serves as a baseline for future governing research. Thus, in the context of this thesis the actual design science artifact is the development and implementation of a governing method in Math Garden. The governing method was developed using the last four steps of the engineering cycle. An important requirement was to involve different stakeholders from Math Garden at the earliest possible stage. The reason being the potential system modifications that could be needed in order for the governing method to work. Therefore, the researcher chose an agile development process supported by the engineering cycle (Wieringa, 2014). This meant a functional artifact could be tested in the early stages of the research, without complete implementation. In later stages of the research, the artifact was completely implemented and released.

In order to validate if the artifact satisfies requirements in the treatment validation step, an additional method is introduced, namely A/B Testing. A/B tests, also known as online randomized controlled experiments, split test, control/treatment tests, and online field experiments (Kohavi, Longbotham, Sommerfield, & Henne, 2009), is a method that large internet companies perform regularly. It is a technique to evaluate ideas quickly in web environments using controlled experiments (Kohavi & Longbotham, 2017). Controlled experiments allow for establishing a causal relationship with high probability by forming hypothesis and evaluating these with real users. In A/B tests, users are exposed to one of two variants: Control (A), or Treatment (B) (Kohavi et al., 2009). A/B tests are particularly powerful as for instance randomization and allocation are extremely easy and interventions can be implemented homogeneously (Brinkhuis et al., 2018). Moreover, A/B tests enable iterative improvement (Williams et al., 2014), which is in line with our agile development process. Lastly, Math Garden already has A/B testing mechanisms in place. Savi, Ruijs, Maris, and van der Maas (2017) performed A/B testing in Math Garden and argued that A/B tests allow for evaluating learning intervention of large groups of students and can reveal patterns or side-effects in learning.

Thus, A/B tests are a good fit for validation of the artifact.

4.1 Metrics & Measures

In A/B tests, in order to measure the effectiveness of the treatment, quantifiable measure(s), or performance metric(s) are defined before the experiment is carried out. This is also known as the Overall Evaluation Criterion (OEC) (Roy, 2001), and is defined as the experiments objective. A single metric is highly desirable, since this forces trade offs to be made for multiple experiments and aligns the organization behind a clear objective. Metrics are statistically tested using hypotheses. In this thesis, a single performance metric to reject or accept the treatment could not be formulated, since too many factors could influence our experiment. Instead two metrics, namely *Engagement* and *Ability* have been selected. The *Engagement* metric reveals if students adhere to the solution governing method, thus willingly practice selected domains, without displaying abnormal practicing behavior compared to the default solution. The *Ability* metric reveals whether the solution governing method accomplished the intended outcome, which is increasing student abilities. The metrics are split into several explicit measures, which are described in Table 4.1.

Measure	**Definition**
	Engagement
First Click Rates	This measure checks if students click on selected domains when first logging in. This is very important, since in order for the governing method to work students should practice selected domains first, therefore click on these domains. Increased first click rates indicate the governing method is effective.
Finishing Rates	This measure analyzes whether students finish their sessions when practicing selected domains. It is important that student finish these sessions and do not stop immediately after clicking on a selected domain. Many unfinished sessions indicate the governing method is not working properly.
Bonus domain practicing frequency	This measure indicates possible negative student practicing behaviour. Math garden is divided in base and bonus domains. For a successful governing method it is important that students keep practicing base domains, since these directly relate to basic abilities. An increase in bonus domain practicing frequency indicates the governing method has undesirable side effects.
	Ability
Domain Scores	This measure analyzes domain scores of students of the default and treatment solution. An important goal of the treatment is to increase student abilities. Domain scores in Math Garden describe specific abilities and are translated into Q-scores or θ scores, therefore an increase in domain scores, is an indication of a successful solution. Moreover, this measure analyzed whether students have passed certain domain score criteria set before the experiment.
Completed Items	In order to acquire specific abilities student should practice domains. This measure monitors what students practice, thus how many items are completed in specific domains. The treatment solution requires students to practice all relevant domains, and an indication of a successful experiment is whether students practice important domains which are avoided with the default solution.

Table 4.1: Measures and definitions

4.2 Hypotheses

In this section, seven two-tailed hypotheses are formulated relating to important metrics to determine the effectiveness of the solution treatment. Hypotheses will be referred to using their abbreviated form (e.g. Hypothesis 1 refers to H1). H1, H2 and H3 relate to the *Engagement* metric and H4, H5 and H6 relate to *Ability*.

Engagement

$H1_0$: Watering can first click rates decrease significantly when governing is applied to Math Garden.

$H1_1$: Watering can first click rates increase significantly when governing is applied to Math Garden.

$H2_0$: Finishing rates of watering can practice sessions decrease when governing is applied to Math Garden.
$H2_1$: Finishing rates of watering can practice sessions increase when governing is applied to Math Garden.

$H3_0$: Bonus domain practicing frequency decreases significantly when governing is applied to Math Garden.
$H3_1$: Bonus domain practicing frequency increases significantly when governing is applied to Math Garden.

Ability
$H4_0$: The number of domains with less than 30 items completed increase significantly when governing is applied to Math Garden.
$H4_1$: The number of domains with less than 30 items completed decrease significantly when governing is applied to Math Garden.

$H5_0$: The number of domains with domain scores below a baseline increase significantly when governing is applied to Math Garden.
$H5_1$: The number of domains with domain scores below a baseline decrease significantly when governing is applied to Math Garden.

$H6_0$: Domain scores decrease significantly when governing is applied to Math Garden.
$H6_1$: Domain scores increase significantly when governing is applied to Math Garden.

Chapter 5

Governing method Design

The first phase of this project consisted of identifying and defining the problem. In the second phase, a solution governing method was designed and developed for use in a randomized experiment. As mentioned in Chapter 4, Treatment design followed the engineering cycle (Wieringa, 2014). In this section, governing is assessed and available governing methods are identified and conceptualized. Finally, the solution governing method (i.e Governer) is designed and developed.

5.1 Governing

Governing comes forth from the concept Governance (Fukuyama, 2013), which is defined as the ability to make and enforce rules and to deliver services. In this thesis governing is defined as computerized supervision within an educational practice system. It is the process utilizing data to guide students in practicing most important educational content, by enforcing certain rules, therefore excluding the need of a person with authority. We believe governing is a very comprehensive term and can be carried out in unlimited possible variations. Therefore, we divided governing into the concepts: perspectives, strategies and visualization each with distinct categories and examples.

5.1.1 Governing Perspectives

Governing can be driven from several theories (i.e. viewed from several perspectives). First, governing could be driven from education. Thus, what domain should be practiced based on grade specific governmental guidelines. For example, in the Netherlands the national institute of curricular development has clear guidelines concerning grade specific math and arithmetic development in students. The second perspective, is to look at governing from a practice system perspective. Thus, what domain(s) should a student practice to attain the highest learning gain? In context of a practice system in which domain(s) can a student gain the most rating/score? Lastly, Governing can be approached on a peer oriented perspective. Thus, how do students score in practice systems compared to other peers. This could be nationwide such as grade specific reference distributions. yet also more detailed on a school or class level. The researchers believe a governing method should not be limited to one theory only, instead a combination of theories is favorable, since this negates possible negative assumptions about the solution governing method.

5.1.2 Governing Strategies

Besides perspectives, a governing model is also comprised of a governing strategy. A strategy implies the plan of action of the data model itself. The researchers classified three data model strategies that get increasingly complex. *Naive* and *Expert* are terms coined from mental models (Gentner & Stevens, 2014). Naive refers to lack of knowledge and experience, yet an expert has existing knowledge, thus is able to solve problems more effectively. Additionally, *A/B test driven* is proposed as a third strategy.

Naive, is a strategy that does not learn and uses existing information to solve a problem. In context of this thesis, a students knowledge level is compared to a target knowledge level (i.e baseline). The target knowledge level is obtained from one or more perspective(s) as stated in the previous section. The strategy involves selecting domain(s) with the largest negative distance from the baseline. Domains are prioritized on grade specific importance. Therefore, some domains may not be included in the selection.

Expert, is a strategy that keeps learning. It predicts students domains knowledge based on knowledge in other domains. This strategy involves building a causal model such as a correlation matrix of all domains. This model, build on historical practice data, functions as an expert and identifies what domain(s) are best practiced. The model should predict domain rating variability when practicing other domains. This prediction is included in domain selection, concluding a certain domain should be practiced in order to gain most rating.

A/B test driven, is an iterative strategy that allows for evaluating learning intervention and can reveal pattern and side-effects (Savi et al., 2017). A/B tests determine what is needed for students to attain a certain knowledge level or domain rating. For example, an A/B test could be to recommend or select only one exclusive domain for each condition. This A/B test identifies how many items need to be completed in order to reach a certain domain rating, when a domain is recommended. A/B tests among all domains continuously train and update the data model.

5.1.3 Governing Visualizations

The last governing concept we formulated is governing visualization. Governing perspectives and strategies involve the data model. The model determines what domain(s) should be practiced, yet it tells nothing about how the selection of domains is presented (i.e. visualized) to students. In this thesis, how domains are presented to students is called a governing visualization. In this section, various visualizations are identified forthcoming from scientific literature. There are countless identifiable visualizations, each unique in there own way. However, each visualization applies general gamification mechanics and exploits motivational incentives to be most effective. The researchers divided governing visualizations into four categories, namely flexible, strict, rewarding and punishing. Each category includes various gamification mechanics (Gamified, 2018) with motivational incentives (Malone & Lepper, 1987). In this thesis the distinction between categories is merely for information purposes. In practice a governing visualization is not limited to one category, but can also be a mix of several categories.

Flexible Governing Visualizations
A flexible visualization offers selected domains as a recommendation. Students have the option to adhere to the recommendation, yet are not forced to do so. This mechanic increases intrinsic motivation, since students are given control over their learning. A recommendation can come in several forms, for example a simple solution would be to offer plain text assignments to students, however this will likely not increase motivation, since there is no incentive to complete the assignment. To increase intrinsic motivation, a better solution would be to exploit the curiosity of students. To invoke sensory curiosity, assignments should be visually appealing. For example, by applying distinct color pallets and attractive images to highlight selected domains (Singh, 2006). A more gamified solution would be to introduce a narrative with accompanied quests (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005). For elementary students, quests involve small assignments from non-player characters, such as pets or other childish fictional characters, since children are likely to relate to such characters. Besides sensory curiosity, Introducing non-player characters also invokes cognitive curiosity, since students can explore and discover new narrative and quests. This is supported in the findings of Chen et al. (2012), which concluded using non-player characters to deliver quests results in increased enjoyment, clearer goals and stronger goals intensities. Another approach is to invoke the challenge motivator of students. Proposing a challenging recommendation is another solution to increase intrinsic motivation of students. The addition of specific

multiple level goals could increase motivation, since a learner is more motivated when goals are clearly defined. Goals should allow for progression tracking and are more interesting if they have a meaningful outcome, such as a reward. Rewards are explicitly described in "Rewarding Governing Visualizations".

Strict Governing Visualizations
A strict visualization disabled domains, therefore limiting practice freedom of students. Control over learning is partially or completely taken away, hence decreasing intrinsic motivation. The most strict solution disables all domains not selected, yet a less strict solution disables some domains, such as domains least important to a student based on grade. The more strict the visualization the more time a students must spend in practicing selected domains. In theory this should increase student abilities. However, the drawback is less motivated students. Moreover, a decrease in task engagement is also likely (Ryan & Deci, 2000). To compensate for these drawbacks, students should be given the illusion of control, since this significantly improves motivation. For example, an appealing practicing page specific for selected domains, which is presented occasionally, could invoke curiosity and increase intrinsic motivation.

Rewarding Governing Visualizations
A rewarding visualization exploits extrinsic motivation, through rewarding students for completing certain tasks (i.e practicing selected domains). The idea is that a student is more engaged when a certain outcome is desired, such as earning a reward (Vallerand et al., 1992). For example, students could be rewarded with points for practicing selected domains or completing certain tasks/quests. When accumulating enough points, a student progresses to the next level. Furthermore, levels could be depicted in leaderboards, which shows the students with the highest level. Leaderboards invoke the direct competition motivator, since students can measure one's self against others. Leaderboards could even be on a class or school level, therefore creating competition between groups instead of individuals. This also promotes the cooperation motivator between individuals of a group, since there is a common goal of beating another group. Another approach is the introduction of gamification mechanics such as virtual economy, prizes and badges. Practicing selected domains or completing tasks is rewarded with these virtual goods. In theory, this helps raise intrinsic motivation of student by combining it with an extrinsic motivator (Hamari et al., 2014). Dividing tasks into multiple difficulty levels of goals, invokes the challenge motivator. The harder the task the better the reward a student can earn. For example, a student is presented with five tasks on a given day. A small reward is earned when an individual task is completed, yet a big reward is earned when all five tasks are completed.

Punishing Governing Visualizations

A punishing visualization is the opposite of a rewarding visualization. Instead of rewarding students for engaging in an activity (i.e. practicing selected domains). This visualization punishes students for not engaging in an activity. For example, extrinsic rewards such as virtual goods, points, levels diminish when a student does not practice selected domains for a certain amount of time. The idea is that students do not want to lose their progression, thus practice selected domains. Intrinsic motivation increases if an activity is enjoyable or interesting (Vallerand et al., 1992), yet punishment is generally accepted as not enjoyable, hence this visualization is likely to decrease intrinsic motivation.

5.1.4 Governing in Math Garden

Math Garden already has governing in place, though simplistic, it steers students in some way. As mentioned in Section 2.3.1, Math Garden is divided in a base- and bonus garden. In both gardens several domains can be practiced. Base garden domains relate to basic skills students need to learn, whereas bonus garden domains, though educational, do not directly relate to basic skills and are sometimes considered more enjoyable or more interesting by users. Base garden domains are always available, yet bonus garden domains are not. In section 2.3.1, it is explained that plants connected to domains in the base garden start to wither after a certain time. Withering of domains occurs when a domain is not practiced recently. When a student has not practiced a domain in the base garden for more then nine days, a watering can will be presented near a domain (Figure 5.1). Metaphorically, this indicates the plant needs water and translates into a domain that needs practicing by a student. When a watering can is present in base garden, the bonus garden is inaccessible, thus severely limiting the number of domains a student can practice. Watering cans are erased once a student finishes a practicing session in that specific domain. This implies that Math Garden carries out a punishing, flexible governing visualization. Students are punished for not practicing for nine days, yet are free to practice domains in base garden, without any restriction. This punishment governs students in two ways. First, by dividing all domains into ones that relate to basic skills and domains that do not, students are forced to practice the most important domains when watering cans are present. This means students who do not care for their plants at least practice a basic skill when using Math Garden. Second, students might have reasons to practice bonus domains regularly. These students are forced to practice their basic skills at least once every nine days in order to keep their bonus garden accessible. In theory, this should govern students, yet no information is present regarding the watering cans and their effectiveness in Math Garden. Therefore,

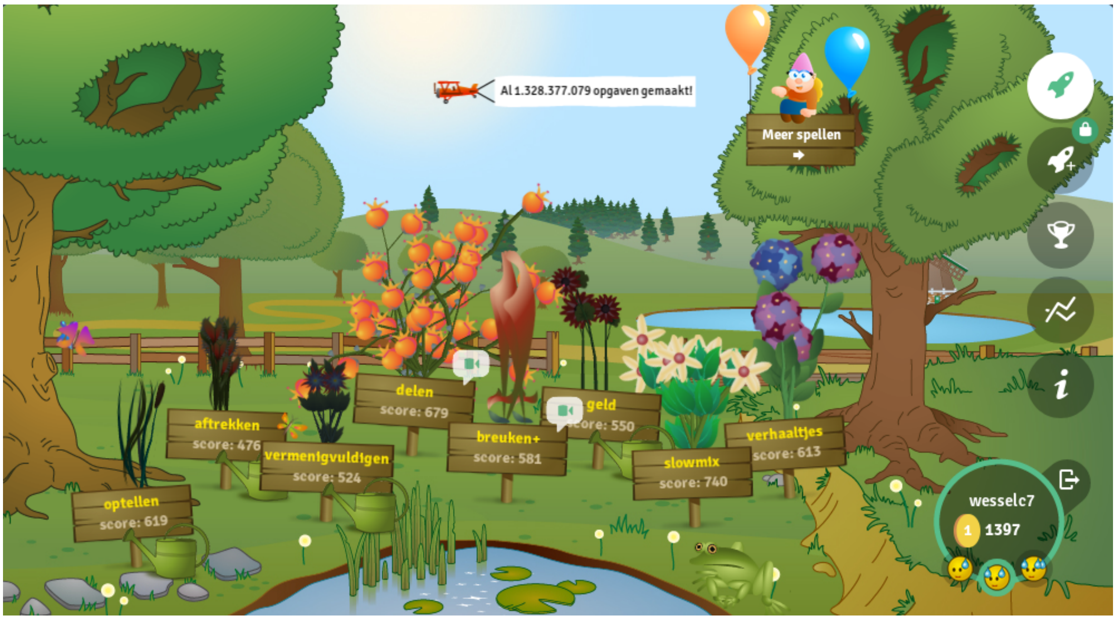


Figure 5.1: In this figure, several domains are depicted in wooden signs. Next to these signs, some watering cans are present indicating the domains needing practiced. Furthermore, domains with watering cans have their associated plants darkened (i.e withered). Therefore, the bonus garden is inaccessible, which is depicted as a lock in the top right corner.

in Chapter 7, the watering cans have been evaluated on their effectiveness and whether or not their intended goals are achieved.

5.2 Governer

in this section, the proposed solution for governing is described. This governing method is called the Governer. The Governer was developed and described utilizing the different governing concepts in Section 5.1, namely perspectives, strategies and visualizations.

5.2.1 Solution Perspectives

As mentioned in Section 5.1.1, governing can be viewed from several perspectives. For the governing solution the researchers combined all three perspectives. For the educational perspective a list composed of Math Garden domains best suited for each elementary grade was adopted. Developed by Oefenweb, this list (Table 5.1) combines curricular development core goals (SLO, 2018) with years of Math Garden knowledge into a clear distinction in which domains are relevant (dark gray) and which are not (white). The Governer only selects domains relevant to students. For example, the Governer will never select Domain 9 for a student in the 7th grade, since this domain is light gray in Table 5.1. The practice system perspective is also included in domain selection. This perspective involves selecting domains in which a student can gain most rating. For each student the Governer compares the Q-scores (Section 2.3.1) of all relevant domains and selects the domains with the lowest score. Arguably, the Governer is also including a peer oriented perspective, since Q-score comparison is utilized instead of θ-scores (Section 2.3.3) of students. Q-scores or Quantile scores are transformed θ-scores (Section 2.3.1). Based on yearly grade populations, Q-scores are calibrated on 300, 500 and 800, which translates into the average score for grade 3, 5 and 8. Thus, Q-scores disclose information of students and how they compare to other peers. For example, students in the 5th grade have an average Q-score of 500. A student in the 5th grade scoring below a Q-score 500 is under achieving compared to his/her peers in the same grade. The reason for Q-score comparison is stated in the next subsection.

ID	**Name (Dutch)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
64	Vorm en kleur	■	■						
65	Figuur en plaats	■	■						
7	Tellen	■	■	■					
41	Getallen		■	■	■	■			
67	Splitsen			■	■	■	■		
40	Mollenspel			■	■	■	■		
10	Reeksen			■	■	■	■		
1	Optellen				■	■	■	■	
2	Aftrekken				■	■	■	■	
12	Slowmix				■	■	■	■	
44	1-2-3'tje			■	■	■	■	■	■
59	Tafels				■	■	■	■	
9	Klokkijken				■	■	■		
3	Vermenigvuldigen					■	■	■	■
4	Delen					■	■	■	■
14	Geld				■	■	■	■	■
60	Verhaaltjes					■	■	■	■
61	Meten					■	■	■	■
13	Bloemencode				■	■	■	■	
5	Snelheidsmix					■	■	■	■
27	Rekenvolgorde					■	■	■	■
70	Codetaal					■	■	■	■
11	Cijfers					■	■	■	■
6	Breuken+						■	■	■

Table 5.1: Relevant domains (Dark Gray) with ID and name for grades 1-8.

5.2.2 Solution Strategy

The perspectives revealed some information of the solution governing model, yet in this section the governing strategy of the data model is described in detail. The Governer reflects a naive strategy classified by the researchers. This is a simple strategy that only needs a students ”garden” as information to select domains.

ID	**Name (Dutch)**	**Condition**
	Enabled	
1	Optellen	Tellen $\geq$ Q-Score 50
2	Aftrekken	Optellen $\geq$ Q-Score 50
3	Vermenigvuldigen	Optellen $\geq$ Q-Score 250 & Aftrekken $\geq$ Q-Score 250
4	Delen	Vermenigvuldigen $\geq$ Q-Score 250
5	Snelheidmix	Vermenigvuldigen $\geq$ Q-Score 300 & Delen $\geq$Q-Score 300
6	Breuken+	Vermenigvuldigen $\geq$ Q-Score 450 & Delen $\geq$ Q-Score 450
14	Geld	Optellen $\geq$ Q-Score 300 & Aftrekken $\geq$ Q-Score 300
9	Klokkijken	Optellen $\geq$ Q-Score 100
13	Bloemencode	Optellen $\geq$ Q-Score 100
11	Cijfers	Vermenigvuldigen $\geq$ Q-Score 600
27	Rekenvolgorde	Vermenigvuldigen $\geq$ Q-Score 600
59	Tafels	Optellen $\geq$ Q-Score 300 & Aftrekken $\geq$ Q-Score 300
60	Verhaaltjes	Optellen $\geq$ Q-Score 300 & Aftrekken $\geq$ Q-Score 300
61	Meten	Vermenigvuldigen $\geq$ Q-Score 400
67	Splitsen	Tellen $\geq$ Q-Score 50
	Disabled	
7	Tellen	Vermenigvuldigen $\geq$ Q-Score 500 & Delen $\geq$ Q-Score 500
64	Vorm en kleur	Vorm en kleur $\geq$ Q-Score 950
65	Figuur en plaats	Figuur en plaats $\geq$ Q-Score 950

Table 5.2: Domains with ID and Name. Domain either enabled or disabled (Criteria) when a certain condition is met.

Model Input

A garden represents all domains and their Q-scores enabled for students in base- and bonus garden. Not all domains are always enabled for students, since a garden is unique for every student. Unfortunately explicit garden information is not available in Math Garden data. Therefore, a garden is reconstructed applying different settings and criteria. A garden depends on so called *garden-settings*. Garden-settings are either *standard*, *alternative* or *custom*. Garden-settings determine what domains are enabled for students in base and bonus gardens. *Standard* is the default setting for students. In this setting domains are enabled/disabled once a certain condition is met (Table 5.2). For, example when a student reaches a Q-score 250 in the addition domain (ID 3), the division domain (ID 4) is enabled. In addition to the default setting, *alternative* and *custom* settings are configurable by someone with authority and do not adhere to criteria in Table 5.2. These settings are either configured on school, class or student level. *Alternative* is targeted towards preschoolers and disables all domains except preschool domains (ID 7, 64, 65). A *custom* setting is like the name suggests, where availability for all domains is adjustable. Domains can either be enabled/disabled, yet can also be placed in either base or bonus garden this is different from the *standard* setting, since base and bonus garden separation of domains is fixed for all students. Separation of domains is depicted in Table 5.3.

Garden	**ID**	**Name (Dutch)**
Base	1	Optellen
	2	Aftrekken
	3	Vermenigvuldigen
	4	Delen
	6	Breuken
	7	Tellen
	12	Slowmix
	14	Geld
	27	Rekenvolgorde
	60	Verhaaltjes
	61	Meten
	64	Vorm en kleur
	65	Figuur en plaats
	67	Splitsen
	70	Codetaal
Bonus	5	Snelheidmix
	9	Klokkijken
	10	Reeksen
	11	Cijfers
	13	Bloemencode
	40	Mollenspel
	41	Getallen
	44	1-2-3'tje
	59	Tafels

Table 5.3: Separation of domains in base or bonus garden. Only applicable for standard garden-setting

After both garden-settings and domain criteria are extracted from Math Garden data, gardens are reconstructed into complete garden data tables for each student. A typical garden data table is depicted in Table 5.4. A table consists of all domains a student has practiced combined with the current Q-score, Modified Count (i.e items completed), Last full session (i.e last time a session was completed) and domain setting (i.e where the domain is placed or if it is disabled). To clarify with an example: In Table 5.4, domain 7 is disabled. This is expected, since both domain 3 and 4 have a Q-score $\geq$ 500. This means the condition for disabling domain 7 is met, as stated in Table 5.2.

Model Output

The governing model needs garden data tables and grade information as its input. The model is build using the R programming language and selects a number of specified domains for each student per day based on clear criteria. First, only relevant domains are selected based on grade of a student (Table 5.1). Second, the researchers decided to only make base garden domains applicable for selection. This decision was taken, because base domains directly relate to basic abilities, therefore are deemed more important for students. Furthermore, the solution governing visualization required base domains to be selected. The solution governing visualization is explicitly explained in the next subsection. Next, remaining domains are selected if *modified count* < 30. The researchers believe a student need to complete at least 30 items in a domain before Q-score is a reliable representation of their abilities. Lastly, remaining domains are selected if their Q-score is below a predetermined baseline. The baseline simply is: *grade* × *100*, thus the baseline of a student in the 6th grade is Q-score 600. As stated in the previous section, Q-scores 300, 500 and 800 are calibrated for grade 3,5 and 8 respectively, however not for the remaining grades. Therefore, the researchers make the assumption that the average Q-score of 2nd and 6th grade students is 200 and 600 respectively, and so forth. All domains below the Q-score baseline are selected and ordered on *last full session*, where the domain with lowest last full session date is first in line. After these criteria some student do not have domains selected by the Governer. For these students an additional criteria is introduced, since students are likely to surpass the *grade* × *100* baseline when practicing frequently. Domains are selected if Q-score is below an increased baseline, namely *grade* × *100 + 150*. Thus, the baseline of students in the 6th grade is Q-score 750. Students without selected domains after the additional criteria are considered exceptional and governing is therefore meaningless for these students. The additional criteria was introduced to compensate for the solution governing visualization, which is explicitly explained in the next section.

5.2.3 Solution Visualization

In this section the proposed solution visualization is described, thus how the selected domains of the governing model are presented to students. At Oefenweb development resources are limited. As such, an important requirement for the Governer was to allocate little development resources in this thesis project. This implicated that introducing new functionality to Math Garden was not possible. For example, a visually appealing recommendation system described in section X could not be implemented. As described in section X, Math Garden already has a simple governing method in use, namely watering cans. The researchers decided to modify the watering can mechanic as the solution governing visualization. As mentioned, normally a watering can appears if a student did not practice a base garden domain for more than nine days. Furthermore, the bonus garden is inaccessible and the associated plants for these domains start to wither in ten phases. Phase one starts after three days until completely withered after 30 days.

In the modified version, watering cans only appear for domains selected by the governing model. The modification needs domains to be selected on a daily basis in order to operate

Student	**Domain**	**Q**	**MC**	**Last session**	**Setting**
12345678	1	707	373	18-05-22 13:19	Base
12345678	2	556	167	18-05-22 13:15	Base
12345678	3	624	144	18-05-22 13:14	Base
12345678	4	560	40	18-05-22 13:13	Base
12345678	6	321	12	18-05-22 13:11	Base
12345678	12	513	12	18-05-22 13:04	Base
12345678	14	848	181	18-05-22 13:09	Base
12345678	27	729	30	18-05-22 13:06	Base
12345678	60	897	135	18-03-29 11:27	Base
12345678	61	815	49	18-03-29 11:22	Base
12345678	67	784	83	18-02-13 11:34	Base
12345678	70	NA	21	18-05-22 13:31	Base
12345678	5	329	2	17-12-18 11:11	Bonus
12345678	9	388	42	18-03-12 9:10	Bonus
12345678	10	45	71	15-12-11 14:20	Bonus
12345678	11	88	29	15-12-11 14:11	Bonus
12345678	13	934	98	18-04-03 10:41	Bonus
12345678	40	946	306	18-03-05 11:17	Bonus
12345678	41	252	43	15-09-24 11:36	Bonus
12345678	44	541	315	17-12-18 11:11	Bonus
12345678	59	158	10	16-05-19 9:23	Bonus
12345678	7	918	382	18-03-29 11:33	Disabled
12345678	64	999	75	17-11-20 13:10	Disabled
12345678	65	969	51	17-06-12 11:23	Disabled

Q = Q-score, MC = Modified Count

Table 5.4: A typical garden table of a random student with standard garden settings. This table represents the students' garden when using Math Garden. Other information, such as garden settings are not visualized here, yet are also captured different columns.

properly. Associated plants have a fixed phase eight withered effect. A Phase eight effect was chosen to invoke the sensory curiosity motivator (Malone & Lepper, 1987), by highlighting the selected domains, subsequently increasing intrinsic motivation. A phase ten withering effect was also considered, yet was deemed too dark. Other watering can mechanics were kept. Watering cans still disappear after a session is completed and bonus garden accessibility is granted when watering cans are not present in base garden. The punishment of bonus garden inaccessibility is kept for two reasons. Most importantly, removing base and bonus garden distinction could introduce undesirable practicing behaviour. If students get unrestricted access to bonus garden, bonus domains are likely to be practiced more frequently, therefore decreasing base domain practicing frequency. Second, bonus garden inaccessibility cuts educational content and could invokes cognitive curiosity if students have the desire to practice bonus garden domains. In theory this adds an incentive to practice selected domains, which is a desirable effect.

5.2.4 Pseudo Code

In this section, everything described in the solution perspectives, strategy and visualization is combined into readable pseudo code.

Algorithm 1 Governer

```
Require: grade
Require: garden data table
Require: max domains
  applicable domains ← domains ∈ relevant domains
  applicable domains ← domains ∈ base garden
  for domains in length(applicable domains) do
    if $modifiedcount < 30$ & selected domains < max domains then
      select domain;
    end if
    if Q-score < grade $\times$ 100 & selected domains < max domains then
      select domain;
    end if
  end for
  selected domains ← sort(last full session)
```

Chapter 6

Experiment

In order to analyze the effectiveness of our solution, the Governer was applied in a real world context. A randomized controlled experiment (i.e. A/B test) was conducted to collect necessary data with which our hypotheses are tested. In this chapter the execution of the experiment is described and discussed.

6.1 Governer Configuration

The Governer was configured to select three domains per student per day. There was no scientific reasoning behind the three domain selection. Four or five domains per day were also considered, yet three domains was considered a save configuration. In the optimal situation the Governer would run daily, presenting students with the best possible domains. Daily calculation was not tested on Oefenweb servers, thus creating concerns regarding server impact. For this reason, the researchers decided to calculate selected domains for students in variant B prior to the experiment. This calculation included all selected domains needed for the entire duration of the experiment. Since, the governing model was designed to calculate domains for one day, some slight modifications were introduced. Normally, the Governer would select a number of domains up to a predetermined maximum (Algorithm 1). For this experiment, the maximum was removed, therefore all domains adhering to the criteria are selected by the governing model. These domains ran trough an additional algorithm which simply multiplied the domains until reaching 42 (3×14) total domains, and dividing them over 14 days. By doing so, students might not be presented with the best possible domain(s) each day, since a calculation prior to the experiment does not take practicing into account. For example, a student might surpass a selection criteria for a certain domain in the first week, yet still will be presented with that domain in the second week.

6.2 Data Collection

6.2.1 A/B Test

After the Governer was designed and modifications were complete, it was applied in Math Garden. For the experiment random elementary student were selected for both Control (A) and Treatment (B) variants. In the control group student use Math Garden as is, thus watering cans were presented after nine days of not practicing. In the Treatment group watering cans were presented if selected by the Governer. Variant sample included active students with standard garden settings. Students were considered active if they practiced in Math Garden in the two weeks prior to the experiment. Furthermore, only students with *standard* settings were included, since *alternative* and *custom* settings include students with very few domains. These students would have meaningless results, since few enabled domains are always selected. A variant sample includes the random active student selection together with their *standard* setting classmates. By doing so, complete classes would be treated with the Governer, therefore functionality confusion among classmates would be non existent. In total $n = 28\,255$ students were selected for the experiment, with $n = 14\,064$ in variant A and $n = 14\,191$ in variant B.

For the duration of the experiment Kohavi and Longbotham (2017) argue that the best practice to run an A/B test is for at least one week, thus capturing a full weekly cycle, and then multiple weeks beyond that. With this knowledge, our experiment ran for two full weeks starting Tuesday May 8th, 2018 up and until May 21st, 2018. The A/B test timing was not ideal, since both Ascension day and Pentecost were within its duration. However, due to time constraints delaying the experiment was not an option.

6.2.2 Pre- and Post-Experiment

To further investigate effectiveness of the the Governer the researchers decided to collect additional data. This data included data prior and post-experiment of both variants, where both variants would be using Math Garden without the Governer treatment. The idea was to analyze behaviour over time, which could support our hypothesis. To be consistent, full weekly cycles were collected. Therefore, pre-experiment data is comprised of data starting May 1st up and until May 7th 2018. Post-experiment data is comprised of data starting May 22nd up and until May 28th.

6.2.3 Garden data tables

In addition to data collected during both A/B test and pre- and post-experiment, garden tables were also gathered daily. As described in section 5.2.3, garden data of students is stored in these tables. Garden tables are used to compare a students ability just before the experiment to its ability after the experiment. Furthermore, garden settings are also captured within these tables. Garden settings could be changed during the experiment. This was important information, since students with standard settings only generate meaning full results. All garden tables were automatically reconstructed and saved at 00:01 hours each day from May 1st up and until May 28th.

Chapter 7

Results

In this section, the results of the A/B test will be discussed. First, variant demographics and responses are described. Second, the metrics: Engagement and ability are analyzed using the measures described in Section 4.1.

7.1 Responses

7.1.1 Student Responses

As described in the previous section, a total of $n = 28\,255$ students were selected in the experiment. As seen in Table 7.1, a total of $n = 13\,645$ unique students actually practiced at least one item during the two week duration of the experiment. In variant A, $n = 6\,814$ unique students are captured and $n = 6\,831$ unique students in variant B. As mentioned in section 6.2, only students with *standard* garden settings were selected and analyzed in the experiment, since students with other garden settings could produce meaningless results. Students garden settings were analyzed over all days of the experiment, resulting in $n = 6\,785$ unique students in variant A, and, $n = 6\,793$ unique students in variant B. Thus, $< 0.5\%$ of students changed garden settings. Moreover, the amount of unique students is nearly equal in both variants, which indicates randomization was successful.

Variant	(selected) n	(responded) n	(standard) n
Control: A	14 064	6 814	6 785
Treatment: B	14 191	6 831	6 793
Total	28 255	13 645	13 578

Table 7.1: Unique Students (n) participated in the experiment. (standard)n Depicts students with standard garden settings over the duration of the experiment

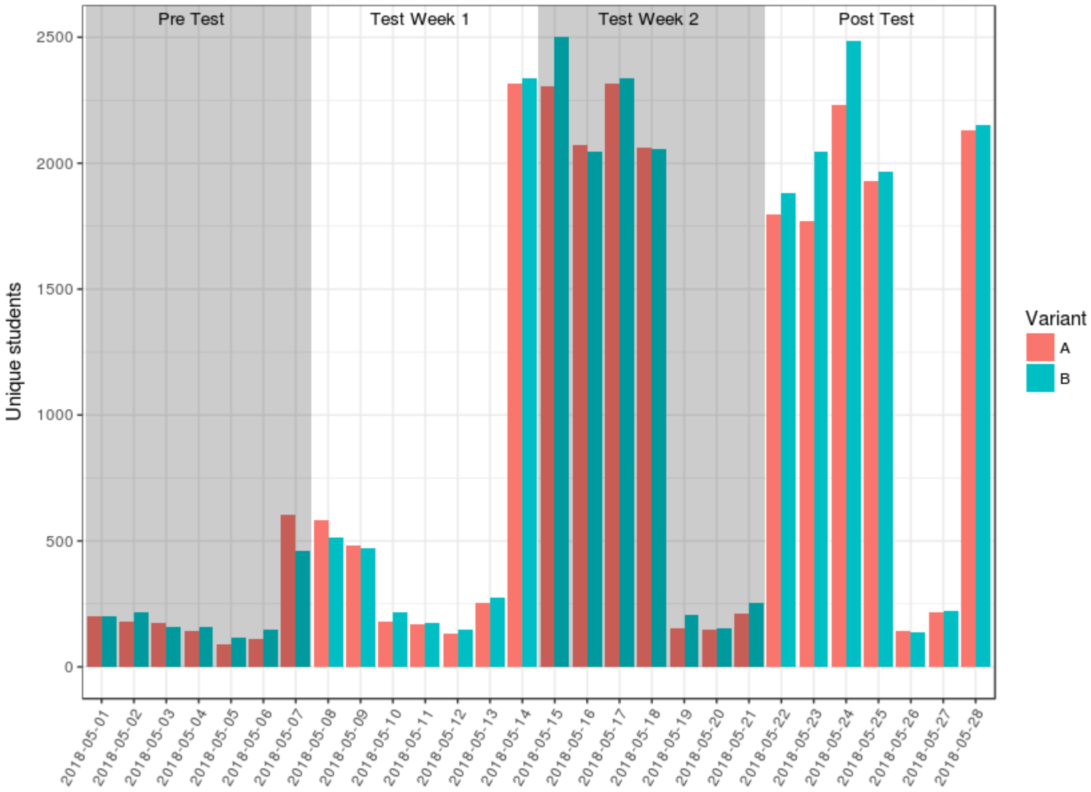


FIGURE 7.1: Unique students responses per day

Additionally, daily unique student participation was analyzed, including pre- and post-experiment. In Figure 7.1 unique students per day are visualized for both variants starting Tuesday May 1st. As depicted, the week prior to the experiment and the first A/B test week have considerably less unique students per day compared to the second A/B test week and week post experiment. The week prior to experiment had the overall least amount of unique students per day, yet n was balanced between variants. This is expected, since the Dutch May Holidays ended May 6th 2018. Less unique students in the first A/B test week were somewhat expected. May 10th and 11th were Dutch school holidays, namely Ascension day. However, the number of unique students in May 8th and 9th were not expected. A possible explanation is that students were still on May Holidays.

The second A/B test week and post experiment week have expected amount of unique students per day. Week-days have balanced unique students in both variants with $1750 < n < 2500$ unique students per week-day. The only exception is May 21st, another national holiday, namely Pentecost. Moreover, weekend-days while balanced have considerably less unique students per day, which is expected when students do not attend school. Lastly, it is noteworthy that unique students per day can not be accumulated to the amounts in Table 7.1, since students can practice daily, thus are regarded as unique in different days.

7.1.2 Item responses

Besides unique student responses, daily item responses were also checked for any unbalances. In Figure 7.2 items responses per day are visualized comparable to Figure 7.1. As depicted item responses per day resemble unique students per day. Highs and lows are proportionally the same compared to unique students per day. Furthermore, items per day are balanced between variants. Variants never exceed a ≈ 10% item difference. On school days between 70 000 and 95 000 items were completed per variant which provides high statistical power on different tests. However, holidays and weekends only produced between 5 000 and 25 000 items per variant. Therefore, holidays and weekends provide less statistical power. This implies pre-experiment and week one of the A/B test provide less statistical power than the remaining weeks. Therefore, further analyses is divided in the different weekly cycles.

7.2 Engagement

As described in Chapter 4, Engagement is one of two metrics formulated to analyze the effectiveness of the treatment solution (i.e. Governer) compared to the default solution

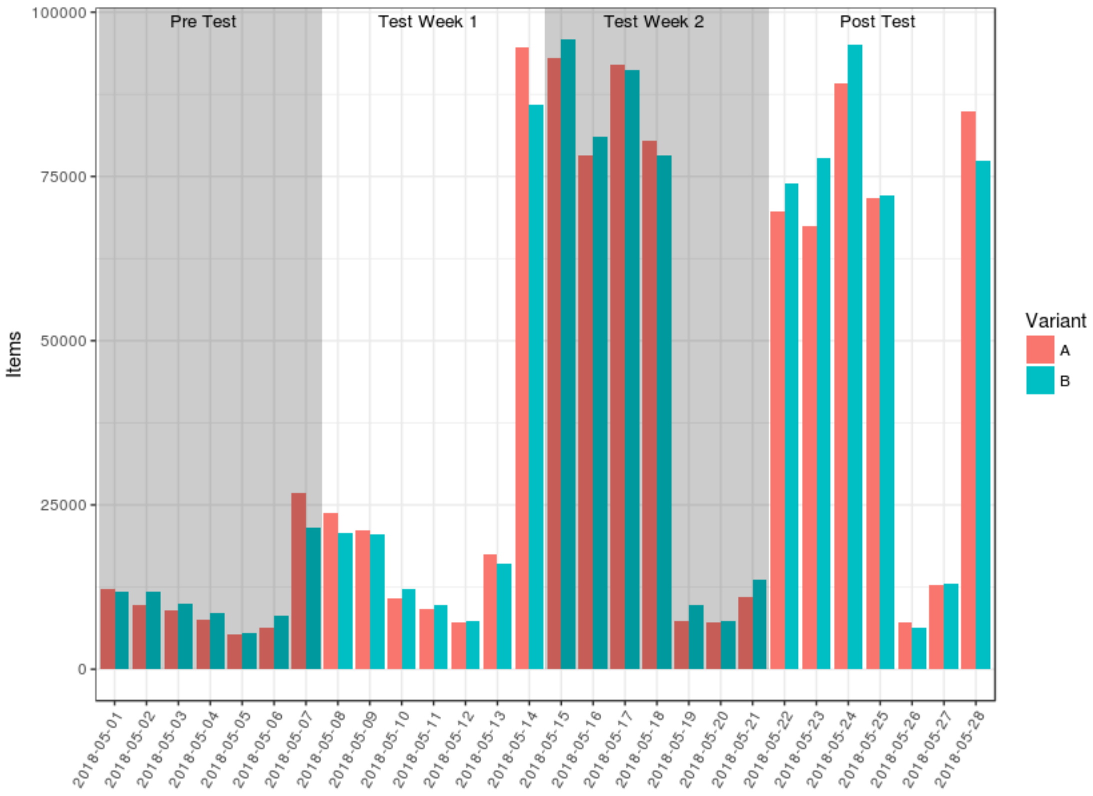


FIGURE 7.2: Item responses per day

in Math Garden. The engagement metric reveals if students adhere to the Governer, thus practice the selected domains (i.e. Watering Cans). Therefore, in this section the quantifiable measures defined in Table 4.2, namely: first click rates, finishing rates and bonus session practicing frequency are statistically tested.

7.2.1 First Click Rates

First click rates measures the the amount of students who click on a selected domain after first daily logins. To analyze this measure the mean watering can first click rates rates were calculated per variant. However, the chance a student can click on of watering can is highly depended on the amount of watering cans present in their garden. therefore, the proportion of watering cans in a garden per student was calculated and rounded by tenths. Subsequently, mean first click rates were calculated for all watering can proportions. The results are visualized in Figure 7.3. In this figure, data was aggregated per weekly cycle before calculation to visualize weekly effects. To improve figure readability, a weighted linear regression over the data was included. Moreover, the mean first click proportion if students were to click on domains at random is presented in the dotted line. Lastly, students with 0% or 100% watering cans are not included in this

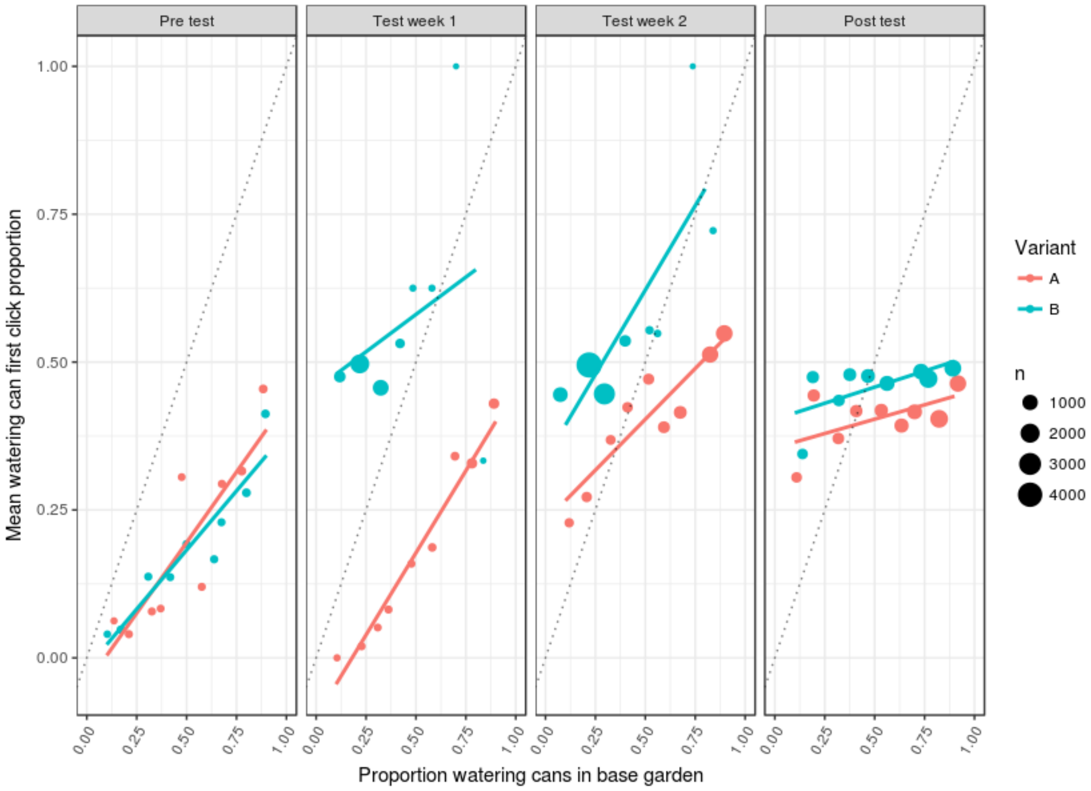


FIGURE 7.3: First click Watering can

figure, since these students either must or cannot click a watering can. In Appendix B.1, a figure including student with 0/100% watering cans is presented.

As depicted in Figure 7.3, students in variant B have higher mean first click proportions over all base garden proportions during the A/B test. Furthermore, in the pre-test mean first click proportions are similar over variants, whereas in the post week, though similar, variant B has slightly higher mean first click proportions. This indicates that on average students treated with the Governer are more likely to click on a watering can than students not treated with the Governer. Moreover, Figure 7.3 also indicates that the default solution is not functioning as intended. In all weeks, either all or a majority of data points is under the dotted line. This implied that students avoid first clicking watering cans, since clicking random domains would produce better results. In contrast, students treated with the Governer have almost all data points above the dotted line. This indicates these students do not avoid watering cans as much as students treated with the default solution.

Lastly, in Appendix B.2, mean first click proportions per day are visualized, which support the weekly findings of Figure 7.3. almost all mean watering can first click proportions of variant B are higher. A post experiment effect is even noticeable on the first two days after the experiment. This could explain the slightly higher mean first click proportions of variant B in post test.

$H1_1$*: Watering can first click rates increase significantly when governing is applied to Math Garden.*
Based on the analyses performed in this section, it can be concluded that Watering can first click rates increase when students are treated with the Governer compared to students who are not. Therefore, $H1_0$ is rejected. However, watering can first click rates are still very low in general even when governing is applied.

7.2.2 Finishing Rates

The finishing rates measure captures whether or not students finish the domain sessions selected by the Governer. To support the first click rates measure it is important that students actually finish the sessions of first clicked domains, since only then a domain would be sufficiently practiced and a watering cans would disappear. Thus, we first looked at finishing rates of first clicked watering can sessions. To analyze the finishing rates measure, all finished and not finished first clicked watering can sessions were counted and tested using Pearson's chi-square test (Table 7.2). The chi-squared test is used to determine whether there is a significant difference between number of finished

Cycle	F	NF	$F\%$	χ^2
Finished first click watering can sessions				
Pre				1.51
Variant A	316	210	60.07	
Variant B	303	170	64.05	
Week 1				21.23**
Variant A	1 043	749	58.20	
Variant B	1 146	596	65.79	
Week 2				55.84**
Variant A	3 288	2 549	56.33	
Variant B	2 572	1 455	63.87	
Post				0.25
Variant A	2 992	2 222	57.38	
Variant B	3 424	2 594	56.89	

** $p < 0.01$, * $p < 0.05$, $df = 1$,

F = Finished, NF = Not Finished

Table 7.2: Finished first session

and not finished sessions in both variants. From Table 7.2, a dependency between variant A and B can be deduced. In both A/B test week 1 ($\chi^2 = 21.23$, $df = 1$, $p = < 0.001$) and week 2 ($\chi^2 = 55.84$, $df = 1$, $p = < 0.001$) a significant difference was found. Furthermore, the percentage finished sessions of all sessions ($F\%$) indicates that student in variant B have a significantly higher chance of finishing first clicked watering can sessions than students in variant A.

Next, all practiced watering can sessions were analyzed, since first clicked watering cans only contains a portion of all selected domains. Finishing rates of all watering cans were also analyzed using Pearson's chi-square test in two approaches. In the first approach, all watering can sessions started and finished on first try were counted versus not finished on first try (Table 7.3). Finished on first try means a students finishes a watering can session on the first attempt of starting a session, therefore removing the watering can. However, a student is also able to cancel a session and attempt to finish that same watering can session in a different attempt. The second approach includes all finished selected domains versus not finished (Table 7.4).

From Table 7.3, a dependency between variant A and B can be deduced. In both A/B test week 1 ($\chi^2 = 129.86$, $df = 1$, $p = < 0.001$) and week 2 ($\chi^2 = 130.73$, $df = 1$, $p = < 0.001$) a significant difference was found. Furthermore, the percentage finished sessions of all sessions ($F\%$) indicates that student in variant B have a significantly higher chance of finishing watering can sessions on first try than students in variant A. Moreover, in the pre week a significant difference was also found ($\chi^2 = 4.03$, $df = 1$, $p = < 0.044$). ($F\%$) indicates students in variant B finish more sessions in general implicating the significant results in could be unreliable. In Figure 7.4, ($F\%$) of finished sessions first try is visualized for each day. From this figure we deduced that the portion watering can

Cycle	F	NF	$F\%$	χ^2
Finished watering can session 1^{st} try				
Pre				4.03*
Variant A	1 207	790	60.44	
Variant B	1 203	688	63.62	
Week 1				129.86**
Variant A	3 433	2 603	56.88	
Variant B	3 366	1 621	67.49	
Week 2				130.73**
Variant A	12 485	9 313	57.28	
Variant B	7 120	4 037	63.62	
Post				1.71
Variant A	11 334	8 313	57.69	
Variant B	12 688	9 065	58.33	

** $p < 0.01$, * $p < 0.05$, $df = 1$,

F = Finished, NF = Not Finished

Table 7.3: Finished session first try

sessions finished ($F\%$ 100) is higher for students in variant B for all dates in both week 1 and week 2. However, in pre and post weeks higher proportions finished vary between variants. Thus, it is assumed that students in variant B do not finish more sessions in general compared to students in variant A, implicating significant results found in week 1 and 2 are reliable.

Likewise, from Table 7.4 the same dependencies were deduced and reasoning was applied. With A/B test week 1 ($\chi^2 = 195.84$, $df = 1$, $p = < 0.001$), week 2 ($\chi^2 = 227.14$, $df = 1$, $p = < 0.001$) and pre week ($\chi^2 = 9.83$, $df = 1$, $p = 0.002$) all having significant results and higher ($F\%$) in all weeks. Thus, significant results could be unreliable. In Appendix B.3, proportion of all finished sessions is visualized per day and provides the same assumptions found in Figure 7.4. Therefore, statistical results found in Table 7.4 are assumed reliable.

Cycle	F	NF	$F\%$	χ^2
Finished all watering can sessions				
Pre				9.83**
Variant A	1 290	707	64.49	
Variant B	1 312	579	69.38	
Week 1				195.84**
Variant A	3 743	2 293	62.01	
Variant B	3 718	1 269	74.55	
Week 2				227.14**
Variant A	13 716	8 082	62.92	
Variant B	7 950	3 207	71.25	
Post				1.67
Variant A	12 376	7 280	62.94	
Variant B	13 827	7 926	63.56	

** $p < 0.01$, * $p < 0.05$, $df = 1$,

F = Finished, NF = Not Finished

Table 7.4: Finished all watering can sessions

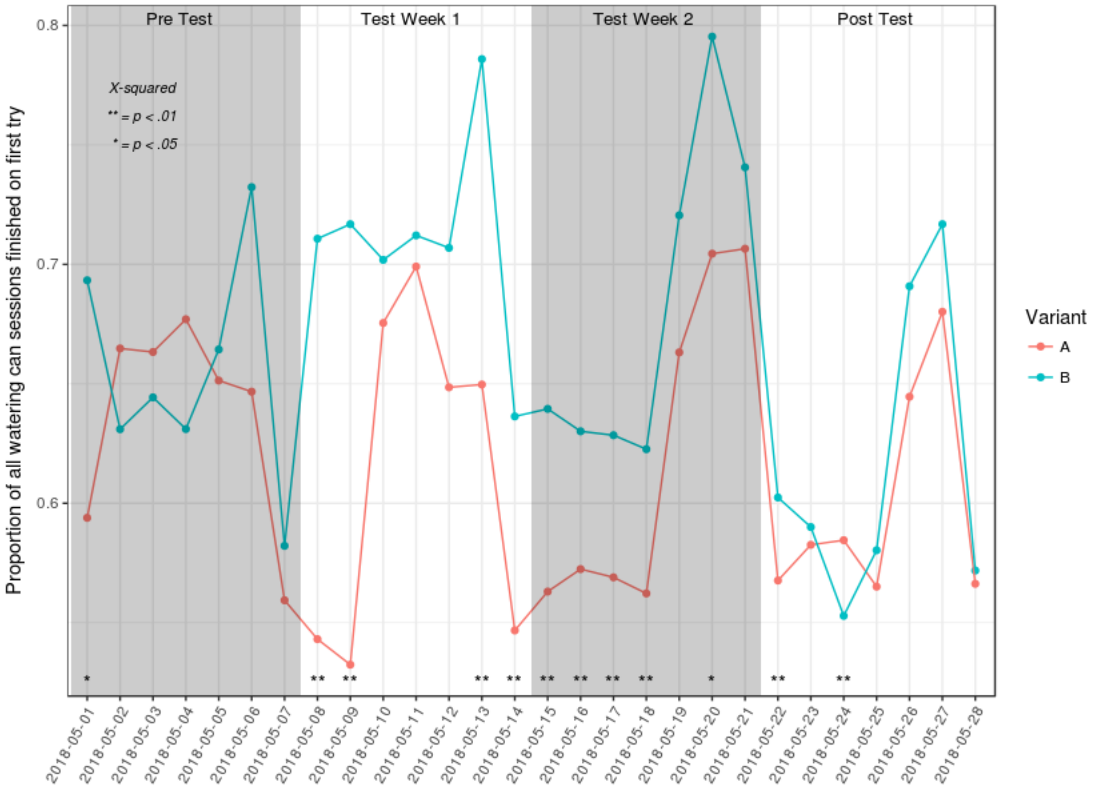


FIGURE 7.4: Proportion finished watering can first try

$H2_1$: Finishing rates of watering can practice sessions increase when governing is applied to Math Garden.

Based on the analyses performed in this section, it can be concluded that finishing rates increase significantly when students are treated with the Governer compared to students who are not. Therefore, $H2_0$ is rejected.

7.2.3 Bonus Domain Practicing Frequency

The bonus domain practicing frequency measure is meant to analyze whether students treated with the Governer practice bonus garden domains more frequently than students using the default solution. It is important that students treated with the Governer keep practicing base garden domains after selected domains are finished, since base garden domains directly relate to basic skills and bonus garden domains do not. To test this measure all finished base and bonus garden sessions were counted for both variants and statistically tested using Pearson's chi-squared test (Table 7.5). From this table significant dependencies between variants were found in both A/B test week 1 ($\chi^2 = 29.03$, $df = 1$, $p = < 0.001$) and week 2 ($\chi^2 = 2\ 833.46$, $df = 1$, $p = < 0.001$). Moreover, significant dependencies between variants in both pre-week ($\chi^2 = 783.75$, $df = 1$, $p =$

Cycle	*Ba*	*Bo*	*Bo*%	χ^2
Finished Base & Bonus sessions				
Pre				29.03**
Variant A	3 673	2 621	41.64	
Variant B	4 020	2 356	36.95	
Week 1				2 833.46**
Variant A	8 185	6 406	43.90	
Variant B	11 919	2 123	15.12	
Week 2				783.75**
Variant A	26 597	2 431	8.37	
Variant B	25 047	4 748	15.93	
Post				98.55**
Variant A	27 885	3 939	12.38	
Variant B	29 572	3 256	9.92	

** $p < 0.01$, * $p < 0.05$, $df = 1$,

Ba = Base, Bo = Bonus

Table 7.5: Base versus Bonus sessions

< 0.001) and post-week ($\chi^2 = 98.55$, $df = 1$, $p = < 0.001$) were also found. However, proportion bonus sessions of all sessions ($Bo\%$) between weeks and variants seems random. Furthermore, daily bonus sessions proportions visualized in Figure 7.5, also provides more evidence that conclusions cannot be drawn from statistical dependencies. As depicted in Figure 7.5, students in variant A have high bonus sessions proportions in both pre-test and week 1, yet drop to almost 0% as week 2 starts. After the start of week 2, proportions slowly climb to fairly consistent proportions beginning mid week 2. Furthermore students in variant B also have a drop in bonus garden proportion at the beginning of the first week, where after proportion stay fairly consistent. Since, overall bonus sessions proportions a very inconsistent significant dependencies are deemed unreliable.

$H3_1$: Bonus domain practicing frequency increases significantly when governing is applied to Math Garden.

Based on the analyses performed in this section, a significant increase or decrease of bonus garden domain practicing could not be concluded for students treated with the Governer. Therefore, $H3_0$ is neither rejected or retained. This means neither positive or negative behaviour could be found in this experiment regarding bonus garden practicing frequency.

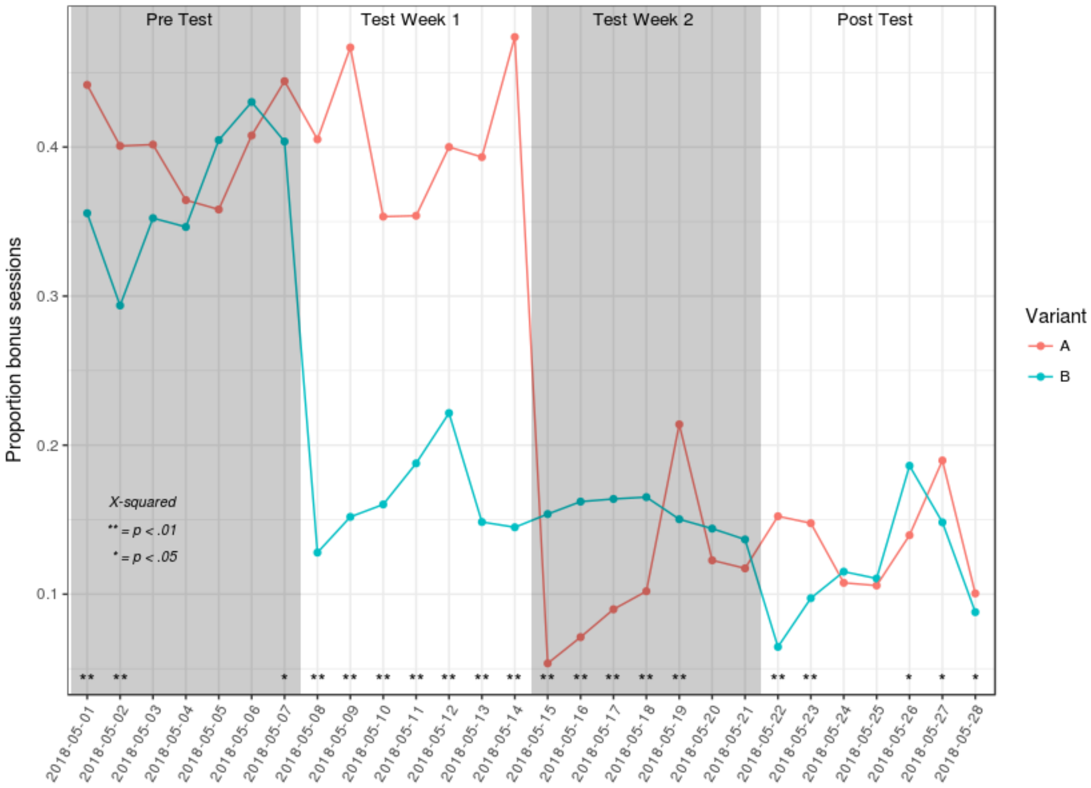


FIGURE 7.5: Proportion bonus sessions over time

7.3 Ability

As described in Chapter 4, Ability is second metric formulated to analyze the effectiveness of the treatment solution (i.e. Governer) compared to the default solution in Math Garden. The ability metric reveals whether the Governer accomplished the intended outcome, which is increasing student abilities. Therefore, in this section the quantifiable measures defined in Table 4.2, namely: Completed items and Domain Scores are statistically tested.

7.3.1 Completed items

The completed items measure monitors whether students practiced relevant domains and completed enough items in those domains. The Governer selected domains based on several criteria. One of the criteria was to select relevant base garden domains with modified count < 30. Therefore, in this experiment the completed items measures to what extend students surpassed modified count 30. To identify differences between variants, two approaches were statistically tested. The first approach only included relevant domains (Table 5.1) in base garden with modified count < 30. The second approach included all base garden domains with modified count < 30. For each student,

Variable	n	M	Mdn	Min	Max	U
			Modified Count < 30			
Relevant base domains						21 150 000**
Variant A	6 703	0.13	0.00	0.00	7.00	
Variant B	6 713	0.21	0.00	0.00	9.00	
All base domains						21 707 000**
Variant A	6 708	0.23	0.00	0.00	9.00	
Variant B	6 719	0.29	0.00	0.00	11.00	

M = Mean, Mdn = Median, U = Mann-Whitney value

** $p < 0.01$, * $p < 0.05$,

Table 7.6: Domains with Modified Count < 30

domains with modified count < 30 were counted just before and after the experiment. Differences between counts were statistically tested for both variants. As depicted in Table 7.6, the data was not normally distributed, thus the Mann-Whitney U test was performed. This test can be used to determine whether two independent samples were selected from populations having the same distribution. In this test it was assumed that a randomly selected value from variant A is less than a randomly selected value from variant B, since we expected students in variant B would practice more domains with modified count < 30. In both approaches a statistical dependency was found, with relevant base domains ($U = 21\,150\,000$, $p = < 0.001$) and all base domains ($U = 21\,707\,000$, $p = < 0.001$). Moreover, the mean difference (M) of variant B is greater than variant A. Thus, results suggest students treated with the Governer have significantly fewer domains with modified count < 30 after the experiment, than student using the default solution. This implicates students treated with the Governer practice selected domains with modified count < 30 more.

$H4_1$: The number of domains with less than 30 items completed decrease significantly when governing is applied to Math Garden.
Based on the analyses performed in this section, it can be concluded that their are significantly fewer number of domains with less than 30 items completed for students treated with the Governer. Therefore, $H4_0$ is rejected.

7.3.2 Domain Scores

The last measure compared domains scores of students between variants. An important goal of the Governer was to increase student abilities. As described, in Math Garden domain scores describe student abilities and are either θ values or Q-scores (transformed θ). First, to test this measure a method comparable to the completed items measure was taken, since an important criteria of the Governer was to select domains with Q-scores below a Q-score baseline. For each student, domains with a Q-score below the

Variable	n	M	Mdn	Min	Max	U
Q-score below baseline						
Relevant base domains						21 465 000**
Variant A	6703	0.09	0.00	-2.00	7.00	
Variant B	6713	0.16	0.00	-2.00	6.00	
All base domains						21 950 000**
Variant A	6708	0.17	0.00	-4.00	10.00	
Variant B	6719	0.21	0.00	-2.00	8.00	

M = Mean, Mdn = Median, U = Mann-Whitney value

** $p < 0.01$, * $p < 0.05$,

Table 7.7: Domains below Q-score baseline

Q-score baseline were counted just before and after the experiment using garden data tables employing the same approaches as the completed items measure. Differences between counts were statistically tested for both variants (Table 7.7). To clarify, after the experiment period a student with difference three had three domains with Q-scores below Q-score baseline fewer compared to the begin of the experiment. As depicted in Table 7.7, data was not normally distributed, hence the Mann-Whitney U test was performed. In both approaches a statistical dependency was found, with relevant base domains ($U = 21\,465\,000$, $p = < 0.001$) and all base domains ($U = 21\,950\,000$, $p = < 0.001$). Moreover, the mean difference (M) of variant B is greater than variant A. Thus, results suggest students treated with the Governer have significantly fewer domains with Q-score below Q-score baseline after the experiment compared to students not treated.

To support these findings domain scores were also tested in a different manner. The previous method only provides information whether or not students passed the Q-score baseline. However, it does not disclose to what extend students gained rating in domains. Therefore, $\Delta\theta$ scores per student per domain were also analyzed. In Figure 7.6, mean daily $\Delta\theta$ of students are visualized for every base garden domains (dotted line), furthermore mean $\Delta\theta$ over all base garden domains is also presented (solid line). In general $\Delta\theta$ is positive for both variant in all weeks, indicating student samples in both variants gained rating, which is expected. However, in both A/B test weeks $\Delta\theta$ of variant B are noticeably higher than $\Delta\theta$ of variant A. To test for significant dependence between variants A t-test was considered between mean $\Delta\theta$ of all domains, yet would have produced inaccurate results, since mean over mean data would lose substantial variance. Therefore, linear mixed models were fitted containing both fixed effects and random effects capturing more variance. Mixed models were tested with the Bayesian Information Criterion (BIC), a criterion for model selection among a finite set of models. The model with the lowest BIC is preferred. In Table 7.8, linear mixed models of $\Delta\theta$ between the beginning and end of the A/B test was fitted with and without variant interceptors. The model with variants interceptor is preferred (BIC with variants = 145 523, BIC without variants = 145 549). Furthermore, variant B intercept ($Estimate$

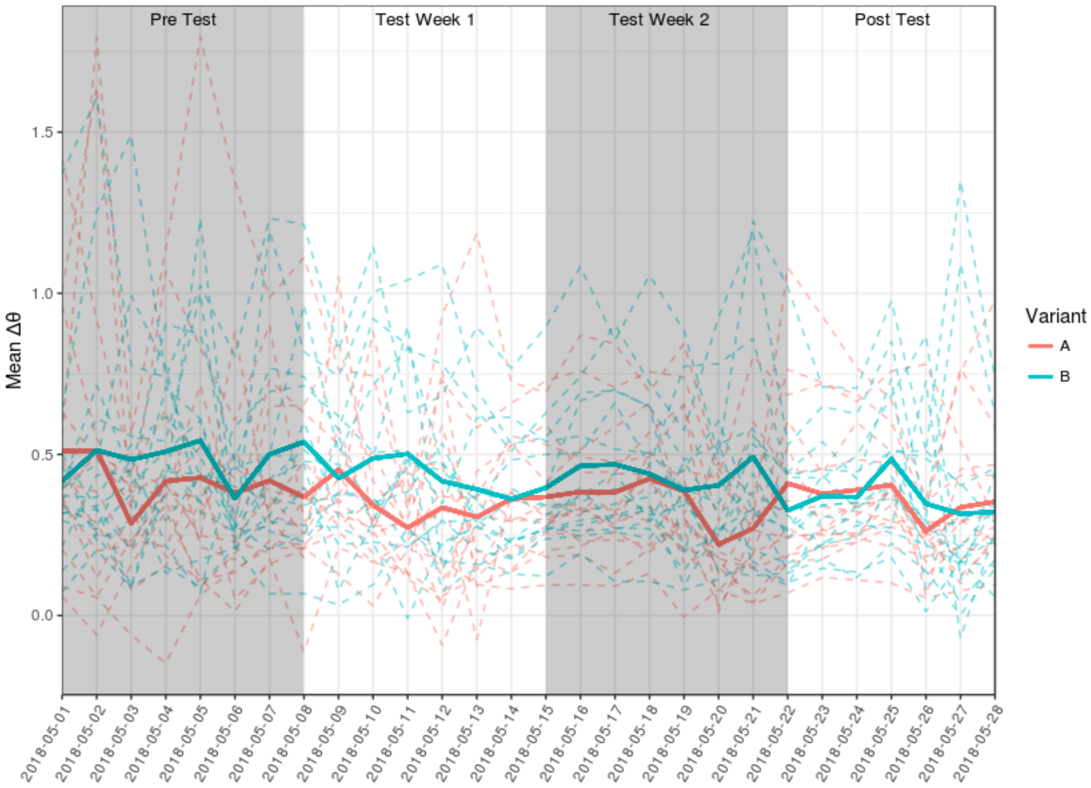


FIGURE 7.6: Mean $\Delta\theta$ over time. Mean $\Delta\theta$ base garden domains (dotted lines), Mean $\Delta\theta$ over domains (solid line)

= 0.09) suggest students in variant B gain +/- 20% more $\Delta\theta$ between the beginning and end of the A/B test compared to students in variant A. However, the inclusion of a daily random effect (Table 7.9) preferred the model without variants (with variants, $BIC = 167\,595$), (without variants, $BIC = 167\,587$). Thus, no significant difference in rating gain per day was found between variants.

$H5_1$: *The number of domains with domain scores below a baseline decrease significantly when governing is applied to Math Garden.*
Based on the analyses performed in this section, it can be concluded that their are significantly fewer number of domains below the Q-score baseline for students treated with the Governer. Therefore, $H5_0$ is rejected.

$H6_1$: *Domain scores increase significantly when governing is applied to Math Garden.*
Based on the analyses performed in this section, it can be concluded that significant increase in domains scores was not found over short periods of time. However, significant increase were found over a longer period of time. Therefore, $H6_0$ is rejected.

	Fixed Effects			
Variable	*Estimate*	*SE*	*t*	*BIC*
	A/B test weeks			
With Variants				$145\,523^{-}$
Intercept	0.47	0.06	7.38	
Variant B	0.09	0.01	6.57	
Without Variants				$145\,549^{+}$
Intercept	0.52	0.06	8.18	
	Post-week			
With Variants				$126\,998^{+}$
Intercept	0.45	0.06	7.96	
Variant B	-0.01	0.01	-0.84	
Without Variants				$126\,981^{-}$
Intercept	0.44	0.06	7.92	

SE = Standard error, *BIC* = Bayesian Information Criterion

+/- = Higher/Lower *BIC*

Table 7.8: Linear mixed model: $\Delta\theta$ over A/B test. Random effects: students, domains

	Fixed Effects			
Variable	*Estimate*	*SE*	*t*	*BIC*
	Days in A/B test			
With Variants				$167\,595^{+}$
Intercept	0.39	0.05	8.26	
Variant B	0.03	0.01	3.29	
Without Variants				$167\,587^{-}$
Intercept	0.41	0.05	8.70	
	Days in Pre- and Post-week			
With Variants				$441\,681^{+}$
Intercept	0.39	0.05	8.34	
Variant B	-0.01	0.01	-0.71	
Without Variants				$441\,661^{-}$
Intercept	0.39	0.05	8.31	

SE = Standard error, *BIC* = Bayesian Information Criterion

+/- = Higher/Lower *BIC*

Table 7.9: Linear mixed model: $\Delta\theta$ per day. Random effect: Students, domains and days

Chapter 8

Discussion & Conclusion

8.1 Discussion

In this thesis, an attempt was made to increase student abilities by assessing ways of governing practice in Math Garden. The so called Governer was developed and implemented, subsequently a randomized experiment showed that governing was successfully introduced in Math Garden.

Based on metrics and measures formulated in Section 4.1, it was found that the Governer had a positive effect on engagement compared to the default solution. First click rates of selected domains increased whilst also increasing finishing rates. Moreover, a significant increase in undesirable practicing behaviour (bonus garden practicing frequency) was not found. This means students adhere to the Governer and willingly practice selected domains. Additionally, the introduction of the Governer also had a positive effect on the ability metric. The number of domains with few items completed significantly decreased whilst domains with scores considered to low also significantly decreased. Domain score gains over a longer period of time were also significantly higher with the introduction of the Governer. This means abilities increase to a greater extend with the introduction of the Governer. The findings of the experiment per hypothesis are summarized in Table 8.1.

Despite positive result described in the previous section, they only disclose whether or not the Governer is more effective than the default solution. Figure 7.3 suggests first click rates on selected domains are considerably higher when the Governer is introduced, yet are far from desirable. The default solution is not working as intended, where students even avoid selected domains, implicating a better solution (i.e Governer) would more easily bring positive results.

	Hypothesis	Accepted?	Observations
	Engagement		
$H1_1$	Watering can first click rates increase significantly when governing is applied to Math Garden.	Accepted	Treated students had increased first click rates
$H2_1$	Finishing rates of watering can practice sessions increase when governing is applied to Math Garden.	Accepted	Treated students had increased finishing rates
$H3_1$	Bonus domain practicing frequency increases significantly when governing is applied to Math Garden.	Not Accepted	Neither positive or negative bonus domain practicing behaviour was found
	Ability		
$H4_1$	The number of domains with less than 30 items completed decrease significantly when governing is applied to Math Garden.	Accepted	Treated students had significantly fewer domains with less than 30 items completed
$H5_1$	The number of domains with domain scores below a baseline decrease significantly when governing is applied to Math Garden.	Accepted	Treated students had significantly fewer domains below the baseline
$H6_1$	Domain scores increase significantly when governing is applied to Math Garden.	Accepted	Significant increase of domains scores were found over a longer period of time

Table 8.1: Hypothesis outcomes

Moreover, the goal of governing was to increase student abilities. Whilst the experiment provides evident increase in domain scores, this cannot be generalized in overall increase of student abilities. Students abilities are not domain scores, yet domains scores explain Math Garden abilities. Thus, we only can conclude Math Garden abilities increase to some extend when the Governer is introduced.

8.2 Limitations

This research project was conducted as part of a Master's Thesis, thus was bound to time and resource constraints from both researchers and Oefenweb developers. This meant some decisions had to be made that limited the outcome of this project.

First, the developed governing visualization had some strict requirements concerning development resources and time. This meant potentially better governing visualizations were hypothesized, yet could not be developed. Furthermore, despite early involvement of all relevant stakeholder, release of the solution visualization passed the initial deadline. This meant the experiment had to be delayed. As described in Section 6.2, this resulted in non ideal experiment timing, since Dutch Holidays were in its time frame. Therefore, responses were imbalanced between days (Section 7.1) In hindsight, starting the experiment two weeks later could have provided balanced responses resulting in more accurate results.

Lastly, other restrictions meant selected domains needed to be calculated prior to the experiment instead of daily. Thus, students were not presented with the best possible domain(s) each day. Results might have been different if the experiment included daily selected domain calculation. However, we believe such an improvement would only provide additional strength to the tested governing method.

8.3 Research Questions

The ultimate objective of this thesis was to design a means by which relevant stakeholders are supported in adding governing into educational practice systems. Several research questions were formulated in Section 3.1. to assist the research and development process of a solution governing method (i.e Governer) in Math Garden. We believe these research questions acts as a baseline to provide the first steps and reasoning to assist in introducing governing in educational practice systems. Therefore, concluding this research project, the sub-questions will be briefly summarized referring to important sections. Finally, the answer to the main research question will be provided in the next section.

SQ1. *What are potential governing methods in adaptive educational systems?*

As discussed in Section 5.1, governing is a very comprehensive term and can be carried out in unlimited possible variations. Therefore, we divided governing into perspectives, strategies and visualization each with distinct categories and examples. A governing method is comprised of a combination of perspectives, strategies and visualizations. A perspective (Section 5.1.1) was defined as the theory from which governing is driven. Theories were divided in education oriented, system oriented and peer oriented. A strategies (Section 5.1.2) is the plan of action of the data model itself. strategies were classified into Naive, Expert and A/B test driven. A visualization (Section 5.1.3) is the manner in which governing is presented to students and were categorized into flexible, strict, rewarding and punishing. Each category includes various gamification mechanics (Gamified, 2017) with motivational incentives (Malone & Lepper, 1987).

SQ2. *What governing methods are present in Math Garden, and what are the effects on student behavior?*

in Section 5.1.4, an existing Math Garden governing method was identified utilizing the knowledge acquired in **SQ1** .Though simplistic this governing method is driven from a system oriented perspective using a naive strategy, since only base garden domains are selected on the condition that they are not practiced for more than nine days. Governing is visualized by placing watering cans beside selected domains and darkening associated plants. The visualization punishes students for not practicing by making bonus domains inaccessible, yet is flexible for practicing basic skill domain. The idea of this governing model is to force students to practice basic skills permanently or once every nine days, if bonus domain practice is desired. This governing method effectiveness was never evaluated and was also analyzed in the experiment. Results captured in variant A. To summarize, in section 7.3 it can be concluded that the default governing method is not working as intended. In Figure 7.3 it is depicted that students avoid selected domains (watering cans).

SQ3. *What is an effective, feasible governing method in Math Garden?*

In Section 5.2, the solution governing method (i.e Governer) was developed and described using all knowledge acquired in **SQ1** and **SQ2**. To summarize, the Governer utilized a Naive strategy which is driven from both educational, system and peer perspectives. Domains selection was based on reconstructed garden tables (Table 5.4) using a simple algorithm (Algorithm 1). Selected domains were visualized by modifying existing watering can functionality and was effectively tested in an experiment.

SQ4. *How can the governing method's effectiveness be measured and validated?*

In Chapter 4, A/B testing was introduced to validate the artifact (i.e Governer). Students are exposed to one of two variants: Control (A), or Treatment (B) (Kohavi et al., 2009). A/B tests require specific quantifiable measures, or performance metrics before the experiment is carried out. In Section 4.1, two metrics, namely engagement and ability were formulated with explicit measures (Table 4.1). Also, in A/B tests adequate random variant selection is important and experiments should capture full weekly cycles (Kohavi & Longbotham, 2017). Analyses between Variants should be performed for every defined measure and results were statistically tested, which concluded the solution governing method's effectiveness.

8.4 Conclusion

In this research project, governing was hypothesized as a solution to guide students in practicing their most important domains in educational practice systems, such as Math Garden. Since there was no prior research on this topic, Math Garden was studied to explore ways of governing practice. The main question that this thesis attempted to answer was formulated as follows:

> *"**MRQ.** How to increase student abilities, by assessing ways of governing practice in educational systems, applied to Math Garden?"*

This question was answered using several formulated sub-questions. First, the researchers explicitly described governing in the concepts: Governing perspectives, governing strategies and governing visualizations with examples to assess the potential ways of governing practice in educational systems. A governing method was subsequently developed and tested in Math Garden utilizing the knowledge contained in the concepts.

The governing method was found to have positive effects on both engagement and ability. Significant increases in measures: first click rates and finishing rates concluded students were willingly practicing selected domains. Furthermore, significant increases in the completed items measure concluded that students practiced domains they previously barely touched. Also, domains scores significantly increased, concluding students acquired more Math Garden abilities, hence we do not claim that student abilities were effectively increased. Still, we were able to effectively introduce governing in Math Garden.

Concluding, this thesis explored governing and provides the first steps, knowledge and reasoning to introduce governing in educational practice systems. Moreover, this thesis acts as the standard to introduce governing in Math Garden.

8.5 Future Research

Governing in educational practice systems is a new concept, which was first explored in this thesis. Therefore, many relevant and interesting opportunities for future research exist. In this section, opportunities will be discussed most relevant to Math Garden, since governing was already successfully introduced.

First of all, to validate the findings in this thesis, it is necessary to run a second A/B test in Math Garden. Yet, for a longer period of time, with larger sample sizes and without holiday interference during the experiment. The next step would be to release the Governer for all students if results prove to be positive. Ideally the Governer should run daily. However, this thesis proved that daily domain calculation in not a necessary requirement for governing to be effective. Thus, calculating the best domains once or twice a week is also a viable option.

Furthermore, it would also be interesting to improve the Governer itself trough several iterations. A new iteration is A/B tested against the previous iteration. The current Governer utilized a simple naive model for domain selection. In Section 5.1.2, more complex models are explained which could be the next iteration of the Governer and improve domains selection even further, subsequently selecting domains more effectively. Other iterations might even be to use different models per student category, which is not covered in this thesis.

Besides improving the model, it is also interesting to improve the governing visualization (Section 5.1.3) in various iterations. Examples are already provided in Section 5.1.3, yet the first step could be to increase motivation by introducing an extrinsic motivator (Section 2.4) to watering cans, in the form of a reward. Math Garden already has virtual goods, such as coins and cabinet prizes (Section 2.3.1) which could be utilized. Other gamification mechanics could also be introduced such as points and levels and leaderboards (Hamari et al., 2014) that can be earned by practicing selected domains. Different color pallets and new images could also be iterative testes. Different images and colors should be simultaneously A/B tested by including more than two variants, therefore concluding which has the effect on the engagement and ability metrics. Lastly, the current watering can visualization punished students, thus has a negative effect on motivation. Therefore, the researchers believe a combination of flexible and rewarding visualizations could provide the necessary increase in intrinsic motivation to drop the watering can functionality altogether and introduce functionality specifically designed for governing.

Bibliography

Apple, M. (2013). *Teachers and texts: A political economy of class and gender relations in education*. Routledge.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest atlantis, a game without guns. *Educational technology research and development*, *53*(1), 86–107.

Basispoort. (2018). Een inlog voor uw digitale lesmatriaal. Retrieved from http://info.basispoort.nl/

Bomasch, I. & Kish, C. (2015). The improvement index: Evaluating academic gains in college. Retrieved from https://www.knewton.com/results/

Brinkhuis, M., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*(3), 319–338.

Brinkhuis, M., Savi, A., Hofman, A., Coomans, F., van der Maas, H., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system.

Bullough, R., Hall-Kenyon, K., MacKay, K., & Marshall, E. (2014). Head start and the intensification of teaching in early childhood education. *Teaching and Teacher Education*, *37*, 55–63.

Chen, Z.-H., Liao, C., Cheng, H., Yeh, C., & Chan, T.-W. (2012). Influence of game quests on pupils' enjoyment and goal-pursuing in math learning. *Journal of Educational Technology & Society*, *15*(2), 317.

Choppin, B. (1968). Item bank using sample-free calibration. *Nature*, *219*(5156), 870.

Christenson, S., Reschly, A., & Wylie, C. (2012). *Handbook of research on student engagement*. Springer Science & Business Media.

Ciampa, K. (2014). Learning in a mobile age: An investigation of student motivation. *Journal of Computer Assisted Learning*, *30*(1), 82–96.

Cito. (2018). Centrale eindtoets. Retrieved from http://www.cito.nl/onderwijs/primair%20onderwijs/centrale_eindtoets

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PloS one*, *11*(5).

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. In *Proceedings of the 15th international academic mindtrek conference: Envisioning future media environments* (pp. 9–15). ACM.

Doignon, J.-P. & Falmagne, J.-C. (2012). *Knowledge spaces*. Springer Science & Business Media.

Earl, L. (2012). *Assessment as learning: Using classroom assessment to maximize student learning*. Corwin Press.

Eggen, T. & Verschoor, A. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*(5), 379–393.

Elo, A. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Embretson, S. & Reise, S. (2013). *Item response theory*. Psychology Press.

Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In *Formal concept analysis* (pp. 61–79). Springer.

Flatla, D., Gutwin, C., Nacke, L., Bateman, S., & Mandryk, R. (2011). Calibration games: Making calibration tasks enjoyable by adding motivating game elements. In *Proceedings of the 24th annual acm symposium on user interface software and technology* (pp. 403–412). ACM.

Fukuyama, F. (2013). What is governance? *Governance*, *26*(3), 347–368.

Gamified. (2018). Gamification mechanics and elements. Retrieved from https://www.gamified.uk/user-types/gamification-mechanics-elements/

Gartner. (2015a). Gartner highlights the top 10 strategic technologies impacting education in 2015. Retrieved from https://www.gartner.com/newsroom/id/2994417

Gartner. (2015b). Top five strategic technologies impacting k-12 education in 2016. Retrieved from https://www.gartner.com/doc/3170221/top-strategic-technologies-impacting-k

Gartner. (2016). Top five strategic technologies impacting k-12 education in 2017. Retrieved from https://www.gartner.com/doc/3558520/top-strategic-technologies-impacting-k

Gartner. (2018). Adaptive learning definition. Retrieved from https://www.gartner.com/it-glossary/adaptive-learning

Gentner, D. & Stevens, A. L. (2014). *Mental models*. Psychology Press.

Glickman, M. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, *3*, 59–102.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In *System sciences (hicss), 2014 47th hawaii international conference on* (pp. 3025–3034). IEEE.

Kelliher, C. & Anderson, D. (2010). Doing more with less? flexible working practices and the intensification of work. *Human relations*, *63*(1), 83–106.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824.

Kohavi, R. & Longbotham, R. (2017). Online controlled experiments and a/b testing. In *Encyclopedia of machine learning and data mining* (pp. 922–929). Springer.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. (2009). Controlled experiments on the web: Survey and practical guide. *Data mining and knowledge discovery*, *18*(1), 140–181.

Lemke, C. (2014). Intelligent adaptive learning: An essential element of 21st century teaching and learning. Retrieved from http://www.dreambox.com/white-papers/intelligent-adaptive-learning-an-essential-element-of-21st-century-teaching-and-learning

Lord, F. & Novick, M. (2008). *Statistical theories of mental test scores*. IAP.

Malmberg. (2018). Zo digitaal als jij het wilt. Retrieved from http://digitaalvoorjou.nl/

Malone, T. & Lepper, M. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction*, *3*, 223–253.

Maris, G. & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633.

Martin, A. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology*, *77*(2), 413–440.

Momento. (2018). Digitaal leren met momento. Retrieved from https://momento.nl/

Muntean, C. I. (2011). Raising engagement in e-learning through gamification. In *Proc. 6th international conference on virtual learning icvl* (Vol. 1). sn.

Newman, D. (2017). Top digital transformation trends in education. Retrieved from https://futurumresearch.com/top-digital-transformation-trends-education/

Newman, D., Bryant, G., Fleming, B., & Sarkisian, L. (2016). Learning to adapt 2.0: The evolution of adaptive learning in higher education. Retrieved from http://tytonpartners.com/tyton-wp/wp-content/uploads/2016/04/yton-Partners-Learning-to-Adapt-2.0-FINAL.pdf

Noordhoff. (2018). Digitaal. Retrieved from https://www.noordhoffuitgevers.nl/basisonderwijs/digitaal

Paramythis, A. & Loidl-Reisinger, S. (2003). Adaptive learning environments and e-learning standards. In *Second european conference on e-learning* (Vol. 1, *2003*, pp. 369–379).

ParnasSys. (2018). Welkom bij parnassys. Retrieved from https://www.parnassys.nl/

Pelánek, R. (2014). Application of time decay functions and the elo system in student modeling. In *Educational data mining 2014*. Citeseer.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Roy, R. K. (2001). *Design of experiments using the taguchi approach: 16 steps to product and process improvement*. John Wiley & Sons.

Ryan, R. & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.

Savi, A., Ruijs, N., Maris, G., & van der Maas, H. (2017). Delaying access to a problem-skipping option increases effortful practice: Application of an a/b test in large-scale online learning. *Computers & Education*.

Selwood, I. & Pilkington, R. (2005). Teacher workload: Using ict to release time to teach. *Educational Review*, *57*(2), 163–174.

Singh, S. (2006). Impact of color on marketing. *Management decision*, *44*(6), 783–789.

Sleeman, D. & Brown, J. S. (1982). *Intelligent tutoring systems*. London: Academic Press.

SLO. (2018). Tule inhouden & activiteiten. Retrieved from http://tule.slo.nl/index.html

Snappet. (2018). Over snappet. Retrieved from https://nl.snappet.org/informatie/over-snappet/

Squla. (2018). Waarom leuk leren werkt. Retrieved from https://www.squla.nl/

Straatemeier, M. et al. (2014). Math garden: A new educational and scientific instrument. *Education*, *57*, 1813–1824.

ThiemeMeulenhoff. (2018). Digitale leeromgeving. Retrieved from https://www.thiememeulenhoff.nl/primair-onderwijs

Vallerand, R., Pelletier, L., Blais, M., Briere, N., Senecal, C., & Vallieres, E. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and psychological measurement*, *52*(4), 1003–1017.

van den Bergh, M., Schmittmann, V., Hofman, A., & van der Maas, H. (2015). Tracing the development of typewriting skills in an adaptive e-learning environment. *Perceptual and motor skills*, *121*(3), 727–745.

van der Linden, W. & Hambleton, R. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

van der Lubben, M. (n.d.). The end of primary school test. Retrieved from http://www.iaea.info/documents/paper_1162d212f6.pdf

Van Droogenbroeck, F., Spruyt, B., & Vanroelen, C. (2014). Burnout among senior teachers: Investigating the role of workload and interpersonal relationships at work. *Teaching and Teacher Education*, *43*, 99–109.

van Grinsven, V. & van der Woud, L. (2016). Rapportage onderzoek passend onderwijs. Retrieved from https://www.duo-onderwijsonderzoek.nl/wp-content/uploads/2016/06/Rapportage-Passend-Onderwijs-22-juni-2016.pdf

Von Alan, H., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, *28*(1), 75–105.

Wang, H. & Woodworth, K. (2011). Evaluation of rocketship education's use of dreambox learning's online mathematics program. *Center for Education Policy.*

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562.

Wieringa, R. (2014). *Design science methodology for information systems and software engineering.* 10.1007/978-3-662-43839-8. Springer. doi:10.1007/978-3-662-43839-8

Williams, J., Li, N., Kim, J., Whitehill, J., Maldonado, S., Pechenizkiy, M., ... Heffernan, N. (2014). The mooclet framework: Improving online education through experimentation and personalization of modules.

Williamson, J. & Myhill, M. (2008). Under 'constant bombardment': Work intensification and the teachers' role. In *Teaching: Professionalization, development and leadership* (pp. 25–43). Springer.

Wilson, K. & Nichols, Z. (2015). Knewton adaptive learning. Retrieved from https://cdn.tc-library.org/Edlab/Knewton-adaptive-learning-white-paper-1.pdf

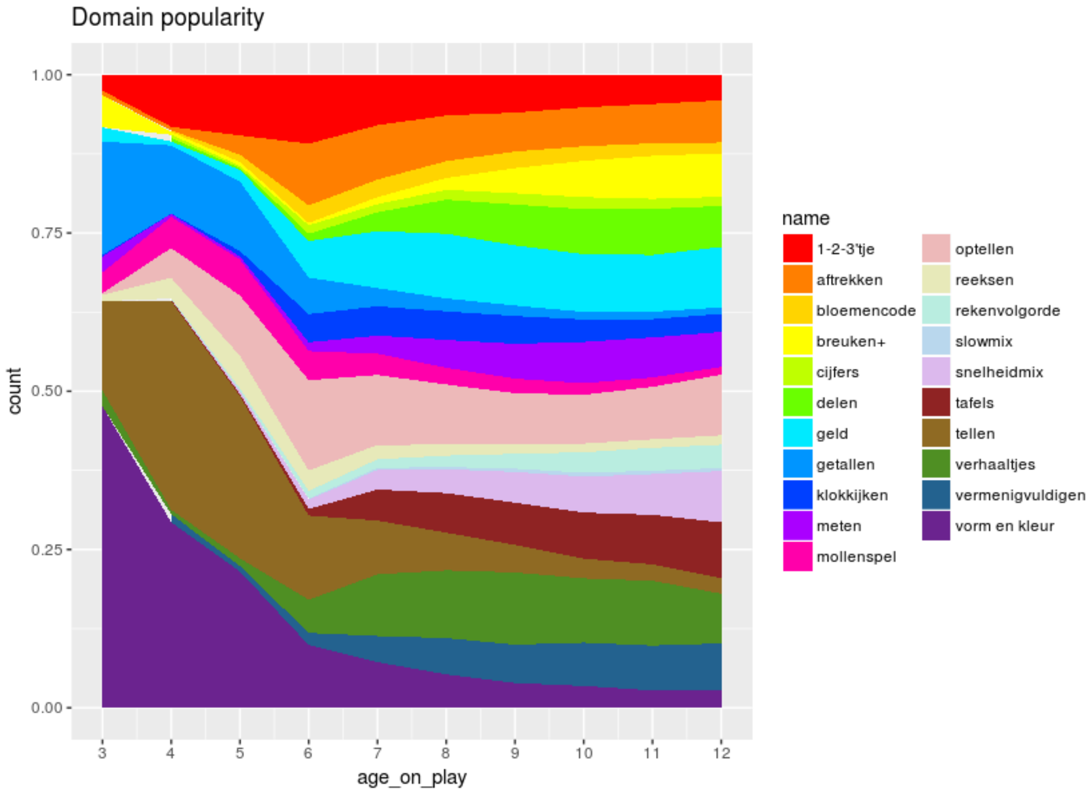
Zichermann, G. & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps.* " O'Reilly Media, Inc.".

Appendix A

Exploratory Data Analysis

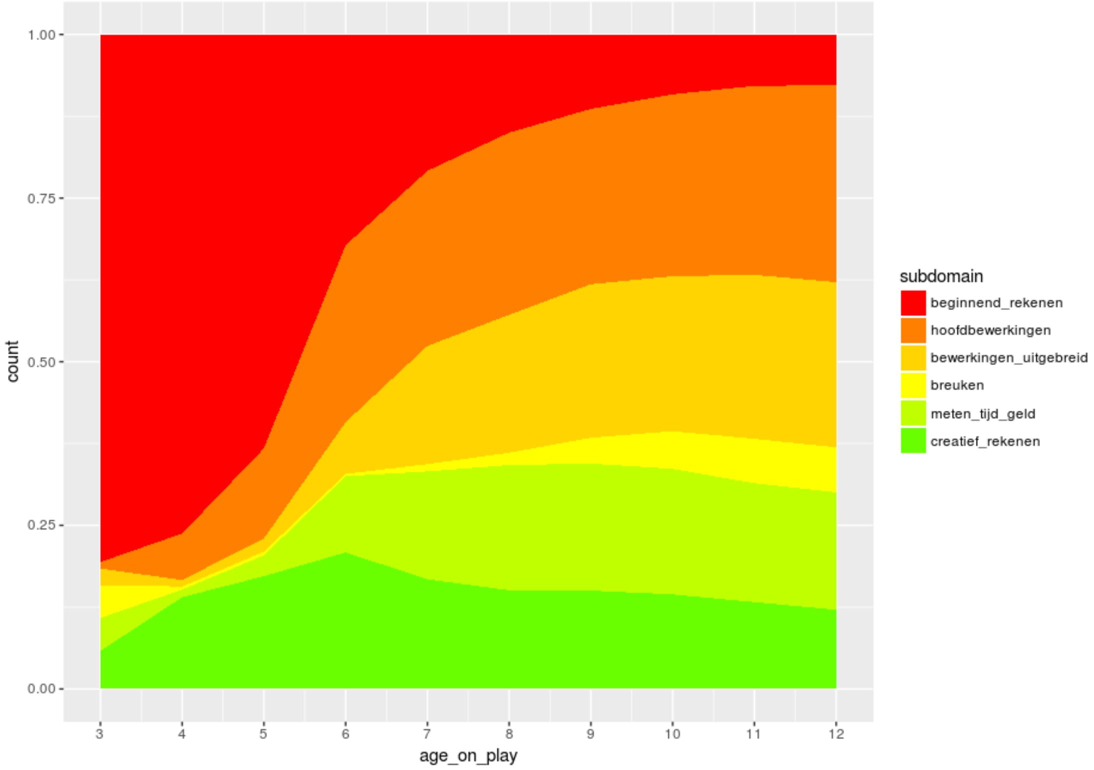
A.1 Domain popularity of all domains

Area plot of proportions per domain, for ages 3-12. Plot based on 6,3 million items completed in a week in January 2018.



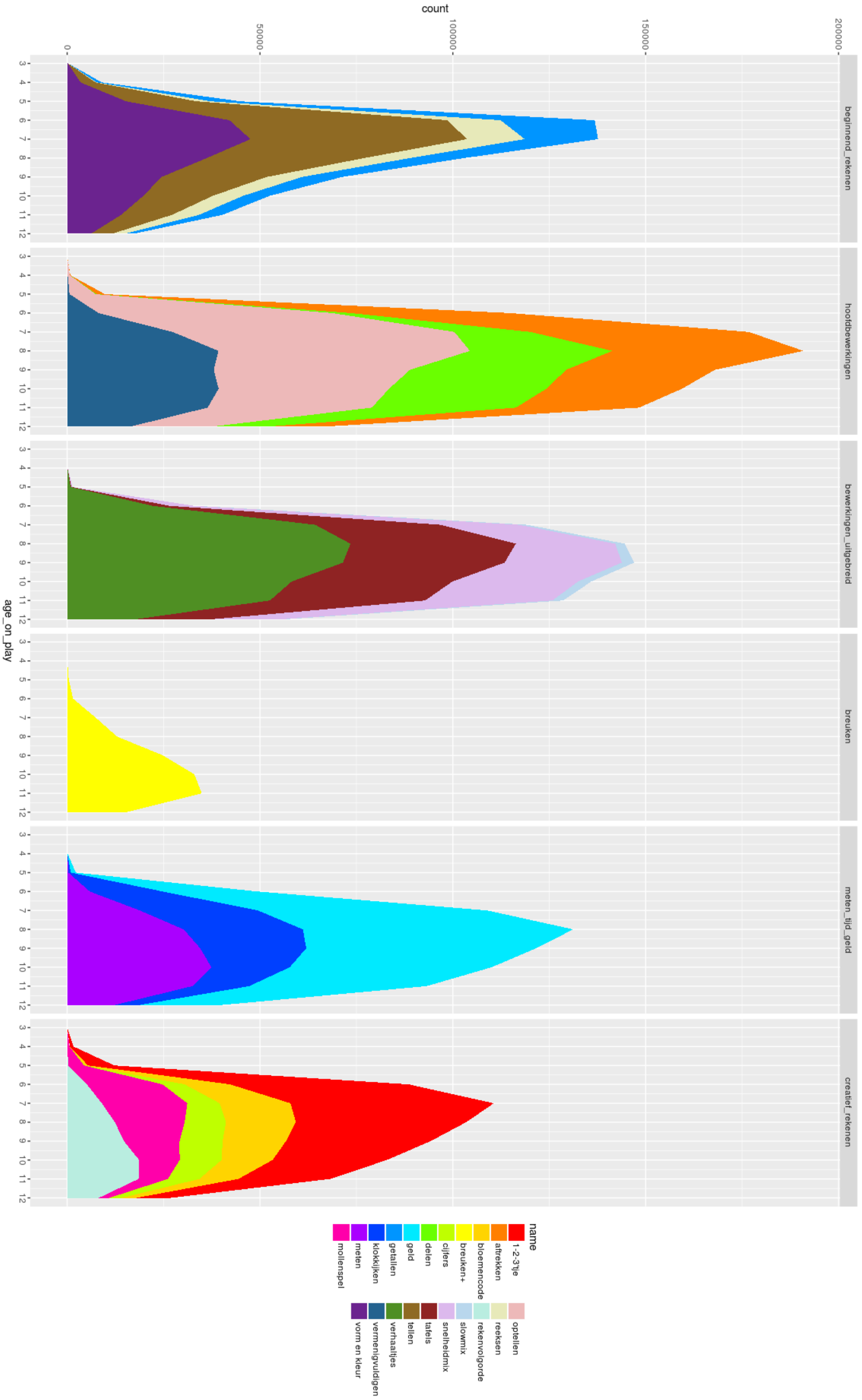
A.2 Domain popularity per sub-domain in Math Garden

Area plot of proportions per domain, for ages 3-12. Plot based on 6,3 million items completed in a week in January 2018.



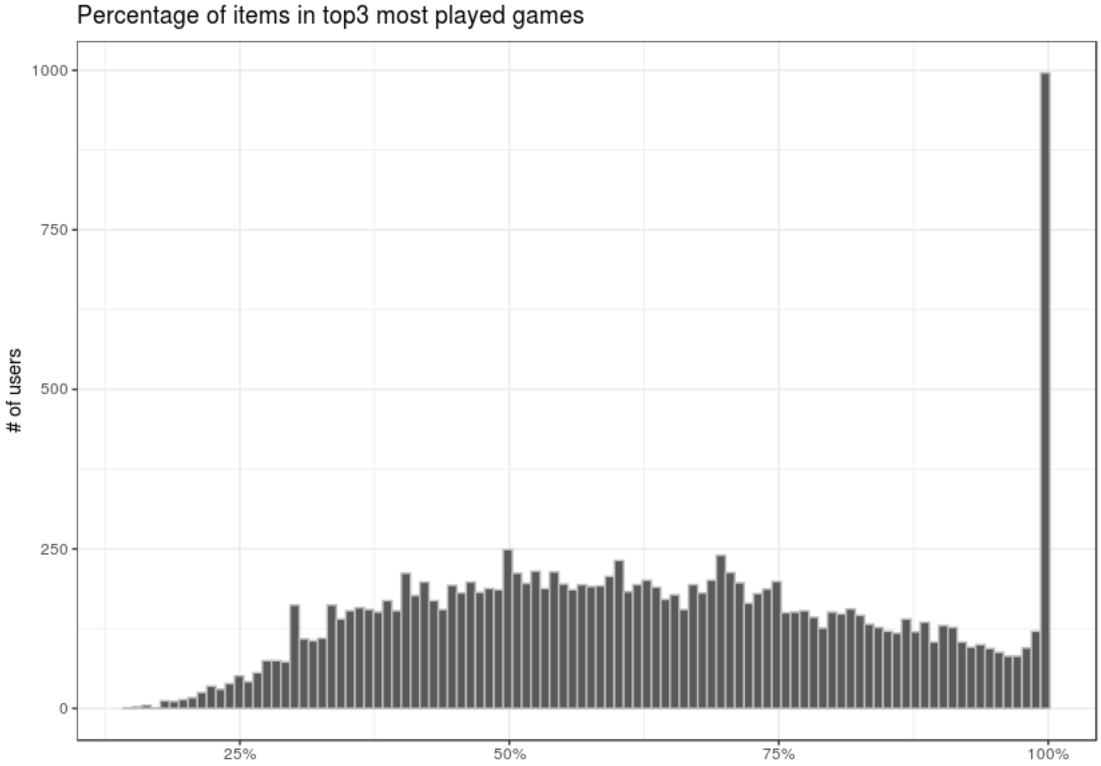
A.3 Math Garden domain popularity divided in 5 subdomains

Area plot of total items per domain per from age 3-12. Plot based on 6,3 million items completed in a week in January 2018.



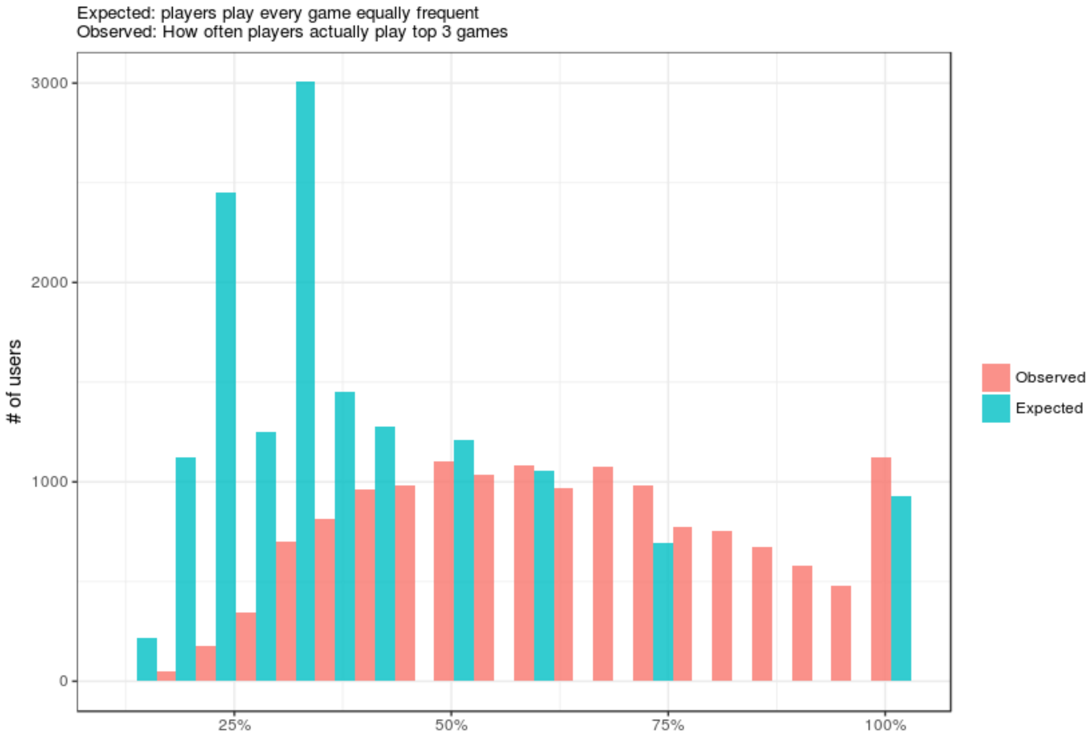
A.4 Percentage of items in Top 3 most practiced domains

Plot based on students who completed more than 100 items in a week in January 2018. 14668 students included. In this graph it is clearly depicted, that a substantial amount of student only play 3 games.



A.5 Observed/Expected percentage top 3 games

Plot based on students who completed more than 100 items in a week in January 2018. 14668 students included. In this graph it is clearly depicted, that expected items played in top 3 games are much lower than observed items played in top 3 games.



Appendix B

Results

B.1 First Click Rates

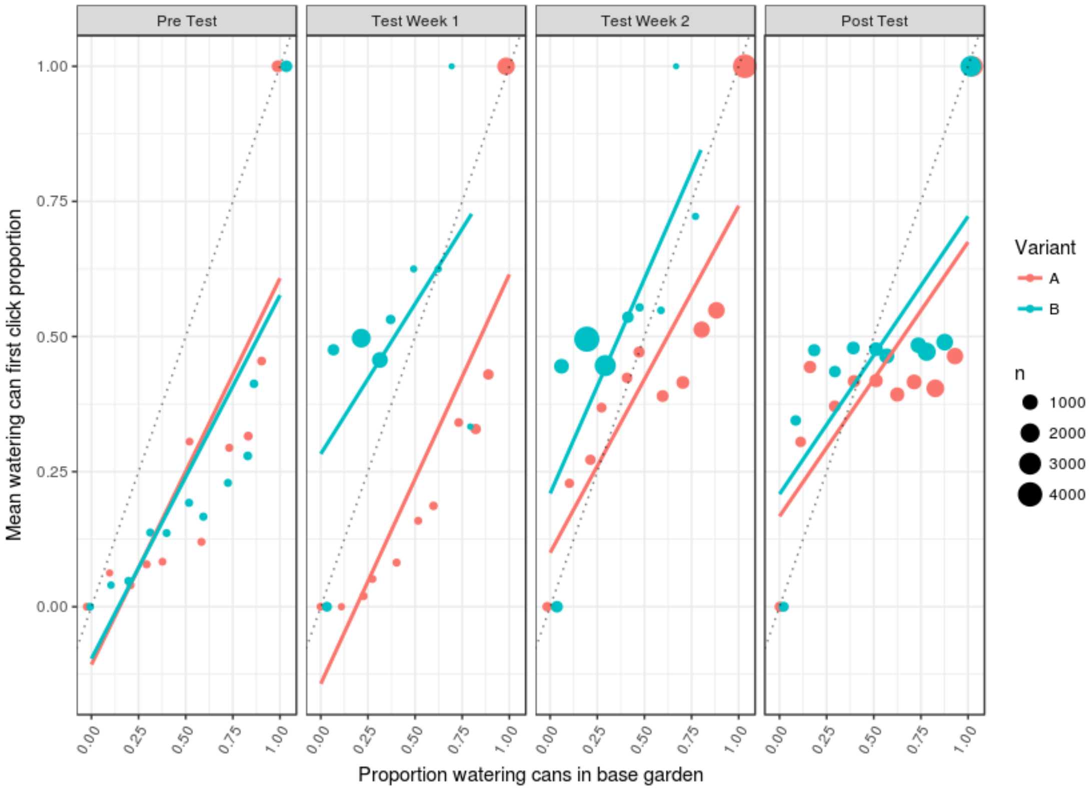


FIGURE B.1: Watering can first click rates with 0 and 100%

B.2 First Click Rates per day

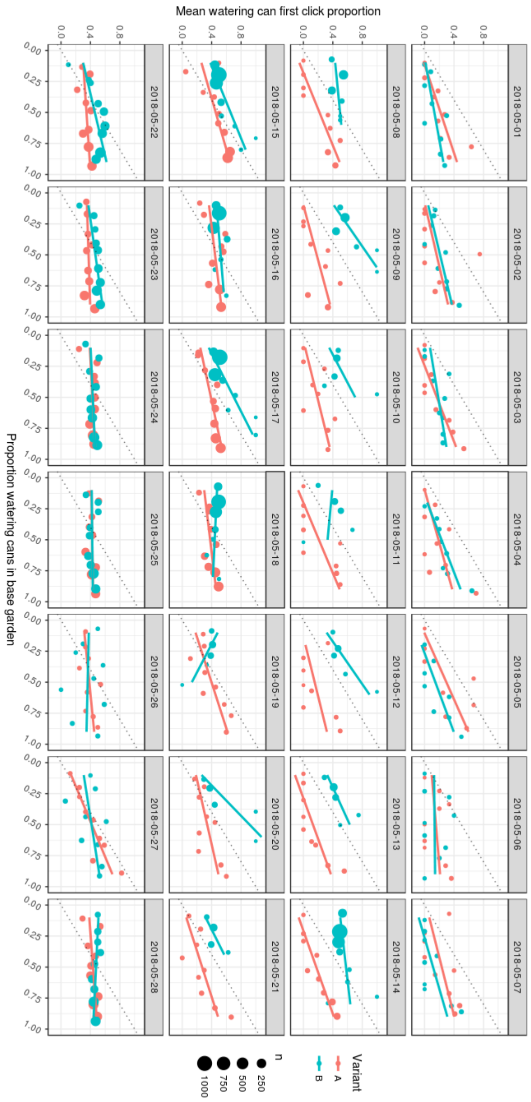


Figure B.2: Watering can first click rates per day

B.3 Finishing Rates

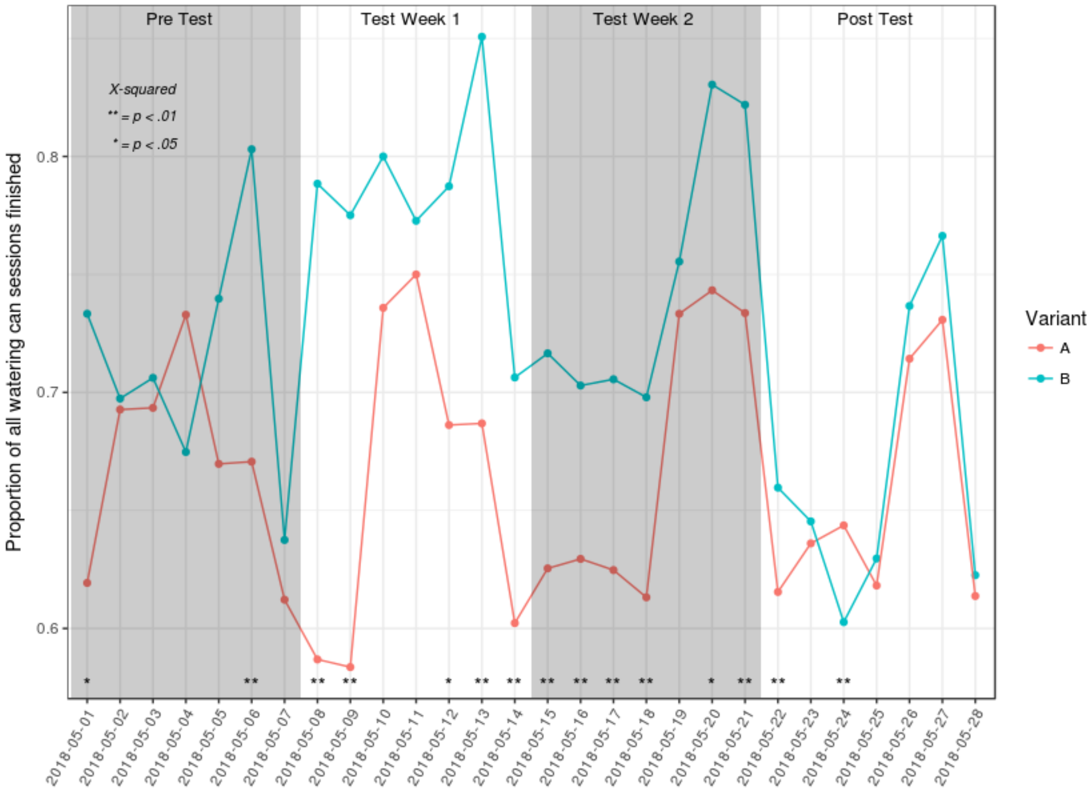


FIGURE B.3: Proportion finished watering can