Universiteit Utrecht

Master Thesis Human-Computer Interaction

Evaluating the Effectiveness of Online Learning Platforms Using Embedded Experiments in Real-World Settings

Author: Merel Das (6600034)

First Examiner:
Dr. Sergey Sosnovsky

Second Examiner:
Dr. Matthieu Brinkhuis

Daily Supervisor:
Jasper Naberman

Host Organisation:
Futurewhiz

Department of Information and Computing Sciences
Faculty of Science

October 2024

HCI-6600034

Abstract

As online learning platforms become increasingly integrated in education, there is a growing need for scalable methods to evaluate their effectiveness in fostering learning. Traditional evaluation methods, such as pre-post tests in classroom settings, are time consuming and difficult to scale. This research explores the use of platform-embedded experiments as a cost-effective, continuous method for evaluating learning platform effectiveness, using Squla and StudyGo as case studies. In Squla, an experiment was conducted to assess the impact of practicing relevant topics through the platform on quiz performance. The results showed that for math and language, answers on a new quiz were 3% more likely to be correct after practicing the relevant topic on the platform, compared to practicing off-topic content. However, no significant improvement was observed for spelling and grammar or reading comprehension. The StudyGo experiment focused on learning within a single attempt at a set of practice questions. The order of questions was manipulated and the results showed that questions were 3% more likely to be answered correctly when placed at the end of a set compared to the beginning, suggesting that students learned from previous questions. These results demonstrate the potential of platform-embedded methods for scalable and efficient measurement of learning outcomes. Future research should address the limitations of these methods, such as their limited generalizability, and explore their applicability across a broader range of content and different learning platforms.

Contents

1 Introduction

Digital technologies play an increasingly prominent role in education, both as a supplement to and an alternative to traditional teaching methods. The COVID-19 pandemic highlighted the need for these technologies, prompting educational institutions to rapidly adopt online learning solutions (Zeng, Sun, Looi, & Fan, 2024). This global shift also helped normalize digital learning as an essential component of modern education. In response, European Member states have received loans and grants to support the digital transformation in education and to address gaps in learning and technology access (Kralj, 2022). The popularity of online learning platforms and resources among families with schoolchildren was also intensified by the pandemic, offering an array of interactive learning tools and instructional methods to supplement school-based learning.

Measuring the effectiveness of educational technology is essential to understanding how well these tools support or supplement traditional classroom learning. The effectiveness of learning platforms can be evaluated in various ways, including but not limited to learning speed, knowledge retention, and student engagement. In this study, we define effectiveness in terms of knowledge gain, based on the understanding that improvement in students' understanding is the primary goal of education (Mashaw, 2012).

Another aspect of measuring the effectiveness of online learning platforms is the method used to collect data. A common approach is to use comparative pre-post assessments with a control group receiving classroom instruction. However, traditional methods often present challenges when applied to younger audiences or prove impractical for platform developers due to resource constraints such as time. In-person research methods can also significantly limit the scope of evaluations, as recruiting a sufficiently large and representative sample from the target population is challenging. This limitation can further impact the statistical power of the conducted analyses. Moreover, digital learning platforms are frequently updated and therefore require continuous efficacy measurement. To address these challenges, this research explored how the effectiveness of learning platforms can be measured in real-world settings using experiments embedded within the platform. This approach avoids the need for user feedback, does not disrupt the user experience, and offers a cost-effective, scalable method for continuous efficacy measurement.

In this research, two online learning platforms designed for schoolchildren across different age groups, Squla and StudyGo, were used to conduct case studies. An embedded experiment was devised for each, which resulted

in two unique approaches that were used to assess platform effectiveness. The analyses were performed using time-stamped log data capturing users' interactions during the experiments. A key challenge in designing these experiments was accounting for the platforms' self-directed approach to learning, where users had full control over the activities they chose to engage in.

The experiment for Squla involved creating a new quiz in which users could practice questions they had previously answered incorrectly. Performance on this quiz was compared between two groups: the on-topic practice group, who practiced the same topic again before retrying their mistake, and the off-topic practice group, who practiced different topics between making the initial mistake and answering the question again on the new quiz. Performance was measured as the correctness of answers to questions in the quiz. With this design, we aimed to determine whether practicing quizzes on Squla effectively improves learning on specific topics. Therefore, the research question we aimed to answer is:

RQ1 Can the effectiveness of an online learning platform be measured by comparing student performance on topics after on-topic practice versus off-topic practice within the platform?

We hypothesized that children's performance on a new quiz would improve after practicing quizzes on related topics within Squla, thereby demonstrating that the feature effectively fosters learning. If such a result could be measured, it would indicate that evaluating platform effectiveness through performance comparisons after on-topic versus off-topic practice is a valid approach. We anticipated that the learning effect would be modest, as practicing 10 questions on a target topic was the minimum for being considered relevant practice, which may not be enough for substantial improvement. Additionally, the time between practice sessions and the new quiz was expected to influence the results, as participants who take the quiz a day after practice are likely to retain more information than those who take it weeks later. As a result, the amount of learning depends on participants' behavior. We also expected the impact on performance to vary across the subjects for which it was implemented: math, language, spelling and grammar, and reading comprehension. Some topics may be easier to learn or are more effectively conveyed through quizzes. We expected the strongest learning effect in math, as basic math facts are effectively learned through repetitive practice, and achieving automaticity in math facts is also associated with success in more advanced mathematics (Baker & Cuevas, 2018). In contrast, subjects like reading comprehension may need more instruction than repetitive prac-

tice with quizzes to achieve substantial gains. Supporting this expectation, a meta-analysis by Higgins et al. (2012) on the impact of digital technology on academic achievement found gains tended to be greater in math than literacy. Similarly, a meta-analysis in the field of mobile-computer-supported collaborative learning showed the largest improvements in math compared to subjects like language arts, social studies, and science (Sung, Yang, & Lee, 2017).

The experiment for StudyGo was designed to determine whether users learn from the practice question sets provided for each topic, one of the platform's main features. The experiment rests on the assumption that users learn from each question they practice, and therefore questions posed towards the end of a set have a higher probability of being answered correctly. By manipulating the positions of questions in the set, this assumption was tested. The research question we aimed to answer with this experiment is:

RQ2 Can the effectiveness of an online learning platform be measured by manipulating the order of practice questions on the platform to observe changes in student performance?

We anticipated that students would learn from practicing questions on StudyGo, with this learning reflected in improved performance over the course of a set. If performance on questions was significantly better when placed at the end of a set compared to the beginning, it would indicate that students were learning from each question, thereby demonstrating the effectiveness of the practice questions. Such a result would also suggest that using an embedded experiment that manipulates question order could be a valid approach for evaluating features within an online learning platform. The expectation that users' performance would improve with repeated exposure to similar questions is supported by Thorndike's Law of Exercise, which states that repetition strengthens the connection between stimulus and response, making repeated practice essential for learning (Thorndike, 2017). However, we expected the learning effect to be small, as most practice sets on StudyGo are relatively short. The sets practiced during the experiment had an average length of seven questions, providing limited opportunities for substantial learning within a single session.

The findings from this study could provide valuable insights for designers and developers of digital education tools, demonstrating how the effectiveness of such tools can be measured through embedded experiments that integrate seamlessly into users' regular activities and allow for efficient analysis.

2 Literature review

The number and variety of online learning tools available is increasing rapidly, from applications teaching toddlers the ABCs[1] to Massive Online Open Courses[2]. These tools are not only increasing in number but are also becoming more integrated into everyday educational settings, reaching a wide audience of real users. The developers of these platforms are also continuously innovating, aiming to increase the success and impact of their products. As a result, measuring the effectiveness of these online learning tools is not only an important research challenge but also a critical key performance indicator (KPI) for developers.

Evaluating the effectiveness of educational technology involves considering various factors, such as its ability to facilitate learning, enhance learning outcomes, and promote positive user experiences and engagement. In this context, the concept of "effectiveness" can be defined in multiple ways, including knowledge acquisition, learning speed, student satisfaction, academic performance (e.g. grades), engagement, confidence, cognitive skills (such as working memory), and learning behavior, among others. This review focuses mainly on effectiveness in terms of knowledge gain, defined as the measurable improvement in understanding or skills as a result of the intervention, as this is the primary goal for most online educational systems (Mashaw, 2012) and the focus of our study.

The effectiveness of various modern educational technologies has been investigated across different contexts and application domains. Several meta-analyses have demonstrated positive impacts on learning outcomes, though results vary depending on the specific technology and its implementation. For example, Cheung and Slavin (2013) examined 74 studies on the effectiveness of educational technology in K-12 mathematics and reported an overall weighted effect size of +0.16. They found that computer-assisted instruction (CAI) had the largest effect size of +0.18, while computer-management learning (CML) and comprehensive models showed smaller effect sizes of +0.08 and +0.07, respectively. Another meta-analysis by Chauhan (2017) found an average effect size of +0.57 across 155 samples from studies evaluating the effectiveness of learning-oriented applications for elementary students. Higgins et al. (2012) reviewed 45 meta-analyses on the impact of technology on academic achievement in learners aged 5 to 18. Their findings also indicated consistent but small positive associations with learning outcomes, with a typical overall effect size between +0.3 and +0.4. However, this review in-

[1] An example is Duolingo ABC. Retrieved from `https://abc.duolingo.com/`

[2] An example is FutureLearn. Retrieved from `https://www.futurelearn.com/`

cluded studies on the impact of both educational technologies, such as CAI, and general technologies, like mobile handheld devices.

Apparent in these meta-analyses is that researchers have explored a variety of methods and metrics to estimate and assess the effectiveness of educational tools. In the next sections of this chapter, a brief overview of the approaches commonly used in the evaluation of online learning tools is presented.

2.1 Traditional methods to evaluate educational technology

A variety of methods have been developed to measure the effectiveness of educational technology. One common approach is the use of pre-post assessments, where learning outcomes are measured before and after an intervention, often with a control group receiving traditional classroom instruction for comparison.

For example, (Pilli & Aksu, 2013) investigated the impact of the educational software Frizbi Mathematics 4 on 4th-grade math achievement. The study compared 26 students receiving traditional classroom instruction with 29 students who used the software for two hours weekly, as part of their regular classes. Performance was assessed through custom pre-tests, post-tests, and retention tests. While the experimental group showed greater improvement on the post-test, only some learning gains persisted over time.

Similarly, Jansen et al. (2013) evaluated the web-based computer-adaptive application Math Garden with 58 adolescents with mild to borderline intellectual disability using pre-post tests. The control group did not use Math Garden, and the independent instrument TempoTest Automatiseren measured the memorization of math facts, focusing on addition, subtraction, multiplication, and division. While overall improvement was similar between groups, students from the experimental group who solved over 1,200 problems on the application showed significant training effects in addition and subtraction.

Papastergiou (2009) used a pre-post design to evaluate an online game's effectiveness in teaching computer memory concepts to 88 high school students. The study compared scores on a custom knowledge test between students using the game and those using a non-game version. The gaming approach was more effective at promoting both learning outcomes and motivation.

The effectiveness of the language learning app Duolingo in improving the reading and listening proficiency of Spanish-speaking English learners

was assessed in a study by Jiang and Pajak (2022). Participants's skills were measured using the standardized test STAMP 4S English, showing significant improvement after completing the initial sections of the Duolingo English course. Participants, selected through self-reports, confirmed the app was their only learning tool.

Aside from custom or standardized tests, other measures to assess pre- and post-knowledge include school grades and perceived learning through self-reports (Mashaw, 2012). This approach was employed in a study on the web-based tutor system ASSISTments for teaching mathematics (Koedinger, McLaughlin, & Heffernan, 2010). The sample consisted of 1,240 seventh graders across three treatment schools and one comparison school, for which the 6th grade year-end test served as a pre-test and the 7th grade year-end test as a post-test. The results showed significant improvements for the treatment group, particularly among special education students.

A field experiment conducted by Chirikov et al. (2020) compared traditional and online instruction for two STEM courses at three higher education institutions in Russia. A total of 325 students were randomly assigned to one of three conditions: traditional in-person classes, a blended format with online lectures and in-person discussion groups, or a fully online course. Exam scores were similar across all groups, but the online group had slightly higher assignment scores, likely due to a more lenient submission policy for online learners.

A study on the language learning app Babbel assessed its effectiveness in developing the Spanish abilities among 54 English speakers (Loewen, Isbell, & Sporn, 2020). The researchers also used a pretest-posttest design, combined with a qualitative analysis of participant comments and interviews on learning gain perception and experience. The tests included a standardized oral proficiency measure (OPIc) and grammar and vocabulary tests. Participants, who were not engaged in any formal Spanish studies, used the app daily for 12 weeks. Results showed improvements across all tests, and participants had well-formed perceptions of what they learned, namely more receptive knowledge than communicative skills.

2.2 Platform-embedded methods to evaluate educational technology

Despite the reliability of traditional methods for measuring effectiveness in terms of learning outcomes, their applicability is context-dependent, and they may have limitations. For instance, traditional methods often involve recruiting participants and conducting measurements in person, which

can result in small sample sizes due to the high costs and logistical challenges associated with this approach (De Witte, Haelermans, & Rogge, 2015). This limitation may reduce the statistical power of these studies, potentially affecting the significance of the findings. Studies using pre-post measurements are also often done in a controlled setting, which can provide accurate measurements but may not reflect natural engagement levels, potentially resulting in significantly reduced effects in authentic settings (Chen & Guthrie, 2019). Effect sizes in these studies also tend to be smaller, as they are typically measured within single sessions lasting an hour or two, which may not provide a sufficient learning period (Martin, Mitrovic, Koedinger, & Mathan, 2011).

Methods that rely on self-reports of learning outcomes can have variable reliability; for example, studies with university students have shown to be reliable in some instances (Mashaw, 2012), but not in others (Martin et al., 2011). Self-reports by children are often found to be less reliable (Broekman, Smeets, Bouwers, & Piotrowski, 2021).

The use of school grades as a measurement of learning outcomes can also pose challenges due to privacy concerns, requiring schools and parents to grant access to this information. School grades may also not be continuously available, as students are assessed at varying and often long intervals.

A few studies have addressed these limitations by using the educational tools themselves to measure learning outcomes. These embedded assessment methods offer an alternative to traditional approaches, providing a more seamless and integrated way to evaluate effectiveness directly through the software. For example, De Witte et al. (2015) examined an online and adaptive educational tool for teaching mathematics in Dutch secondary school classrooms, which provided individual training packages with explanatory movies, theory, and exercises. To start using the tool, students had to complete a pre-test as a part of the program. The post-test score consisted of a student's average score across the subjects they took. This research thus computed pre-post test scores from the data set logged by the program. They also computed engagement levels and found that doing more exercises in the program led to higher test results.

Another study using an embedded approach was performed by Chen and Guthrie (2019), which implemented mastery learning in an online tool. Students needed to master concepts of mechanical energy with 10 online modules. The modules were assigned as homework for a college physics course which counted towards the course grade, and were therefore assessed in an authentic learning setting. Students were required to attempt the assessment problems of a module before being able to access the instructional materials. The attempts at the assessment before and after accessing the instructional

materials served as pre- and post-tests. Time-stamped log data from the system was analyzed to compute the test scores and engagement levels. The analysis results were presented in sunburst charts, allowing instructors to evaluate the effectiveness of the resources used in each module.

These two studies had the advantage that the learning tools were used in a course, being assigned as homework and mandatory for course credits. Researchers could therefore instruct students to do the pre-post assessments, without having to worry about dropout or enjoyment. This is different for commercial platforms, especially when younger children are involved, as their users expect full control of how they spend their time on the platform and might not want to engage in long assessments. The previous studies also benefited from having exercises aligned with the instructional material, which could serve as pre-tests of student knowledge, an option not always available (Chen & Guthrie, 2019). Another advantage is that these studies could assume their online courses were the main source of students learning about a topic. For platforms that supplement school-based learning or cover more general topics, interaction with the system is just one part of students' education. This confounds the results and adds the difficulty of isolating the source of the measured learning outcomes, especially for studies taking place over longer learning periods, which is often required for a rigorous statistical analysis (Martin et al., 2011).

An interesting study that did not benefit from these advantages was conducted by Portnoff et al. (2021) on the language learning app Duolingo. As a self-directed learning platform, Duolingo has limited control over how users engage with its features and thus requires accurate and well-controlled assessments to measure learner achievement. The researchers devised two methods to integrate test items into the platform and analyzed the assessment data using Educational Data Mining (EDM) methodologies. The first method involved checkpoint quizzes, which contained seven pre-test items assessing proficiency for the next part of a course and seven post-test items assessing the part just completed. They observed that learners were more likely to answer post-test items correctly if they leveled up lessons, though they noted that this might reflect self-selection bias, as motivated learners are more likely to level up. To further investigate, they implemented "review exercises", where questions from previously learned material were inserted into later lessons. Accuracy on these exercises was compared between learners with the same studying behavior, except for the completion of an additional level in the source lesson. This approach reduced self-selection bias and found that learners who completed an extra level were more likely to answer correctly, providing stronger evidence of a causal relationship between leveling

up and improved performance.

However, bias may have remained as learners chose to level up, and external learning sources were not ruled out. In contrast, Jiang and Pajak (2022) excluded external sources through self-reports, confirming that Duolingo was the only learning tool used by participants.

While the studies discussed in this section demonstrated the potential of embedded efficacy measurements, they represent only a small fraction compared to the numerous studies relying on traditional in-person methods. Moreover, these studies were tailored to specific platforms or benefited from unique advantages that may not apply to other platforms. This suggests room for further research into developing in-program measurement methods that can be applied across different platforms, especially those that do not benefit from the same inherent advantages.

3 Online learning platform Squla

The online educational platform that supported the first case study for this research was Squla. The following subsection describes the system, including its context of use, and key features. Afterward, the methods, results, and discussion of the efficacy experiment in Squla are presented.

3.1 System description

Squla is an educational platform designed for both toddlers and primary school children, covering ages 3 to 12, and is commercially available[3]. It is primarily intended for at-home use, but can also be used in classroom settings. Accessible via web and mobile applications, Squla provides learning activities for a wide range of standard school subjects, as well as additional topics such as 21^{st}-century skills. The platform's content is organized according to Dutch school grades from preschool to elementary school, but all content is accessible regardless of the user's grade level. The platform has a landing page per grade, see Figure 1, on which available subjects to practice are shown as well as other features like recommendations for quizzes to play.

Each subject is divided into categories, in which related quizzes, learning games, and explanation videos are presented. A quiz may consist of one or more levels, each typically comprising 10 questions. The questions have a set order and after each attempt, the user is shown whether their answer was correct. Most questions offer an explanation of the right answer after a

[3] `https://www.squla.nl/`

Figure 1. *The mobile landing page of Squla for a user in grade 4, presented in sections to capture the full page. On top are the recommended quizzes, followed by a banner for a new quiz, and the subject tiles. At the bottom are banners for new quizzes and learning games.*

mistake, an example can be seen in Figure 2. A distinctive feature of Squla is its use of gamification techniques to engage its young audience. This includes fun question formats like puzzles and bubble poppers, as shown in an example in Figure 3, alongside rewards like games and coins to purchase goodies. Other gamified elements like leaderboards and storylines are also featured, to further enhance user engagement and experience.

The content of the school subjects on Squla adheres to educational standards, as it is based on the core targets specified by SLO, the Netherlands institute for curriculum development[4].

3.2 Methods

This research aimed to assess the effectiveness of online learning platforms in terms of knowledge gain using time-stamped log data and embedded experiments. Squla collects data on user engagement, such as login times, the activities users interact with (e.g. which quizzes they play), and the duration

[4] `https://www.slo.nl/`

Figure 2. *View of an explanation given after a mistake was made on a grade 4 math question. The bottom textbox shows the steps of how to get to the right answer of 40, which is highlighted in green.*

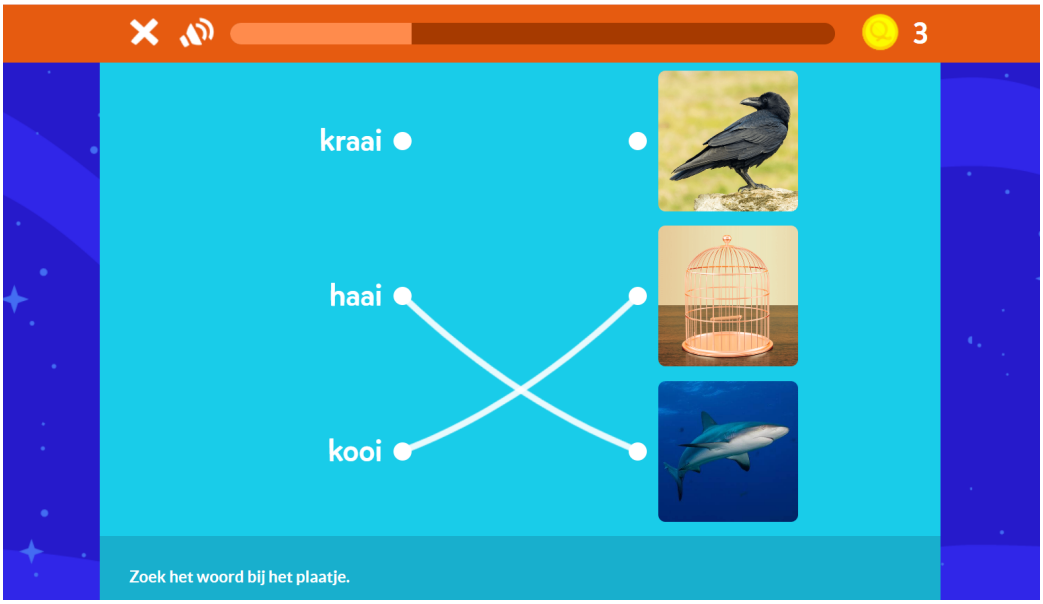

Figure 3. *A fun question format on Squla where users have to connect the words with the corresponding pictures. The question is from a quiz on spelling and grammar in grade 4.*

of these interactions. For performance, the platform tracks response times and accuracy, the specific response given, and the number of attempts per question. Since the platform supports self-directed learning, students have the freedom to choose how many and which exercises to attempt, which can introduce bias when measuring effectiveness. For example, if students who complete more exercises show higher performance, it may be tempting to conclude that the platform is effective because higher performance appears linked to engagement with the platform. However, this might not be the case, as it could also result from self-selection bias, where more motivated students or those with higher prior knowledge engage more frequently, naturally leading to better performance. Therefore, the experiment must be carefully designed to ensure participants are compared fairly, particularly with others who have similar engagement patterns.

3.2.1 Experiment design

Considering the platform's unique characteristics, available log data, and potential biases, various experimental designs were explored. One option was an online variation of traditional methods with pre-post assessments before and after a platform quiz. However, Squla's engagement data showed that children may abandon longer quizzes, leading to the self-selection of only the more motivated users who complete all assessments. Other designs requiring user feedback, such as self-reports, were also avoided due to concerns about reliability with younger users.

Instead, the experiment tested whether practicing specific topics on Squla improved performance on those topics through a new quiz, called the review quiz. This quiz contained questions that users had answered incorrectly over the past four weeks, providing an opportunity for students to revisit and practice their mistakes. The experiment aimed to compare performance on the review quiz between users who had practiced the relevant topics and those who had not.

The quiz was available across four subjects and was presented in the corresponding subject section for grades 1-8. The included subjects are the core school subjects on Squla: mathematics, (Dutch) language, spelling and grammar, and learning comprehension. Figure 4 shows the access point for the new quiz within the mathematics content section. Upon starting the quiz, users were shown a brief explanation of the quiz, see Figure 5, after which questions were presented as usual. Upon completing the quiz they were shown the same end screen as other quizzes in which the amount of coins and XP they receive is shown, as seen in Figure 6. The quiz was available from June to mid-August 2024, during which the data was collected.

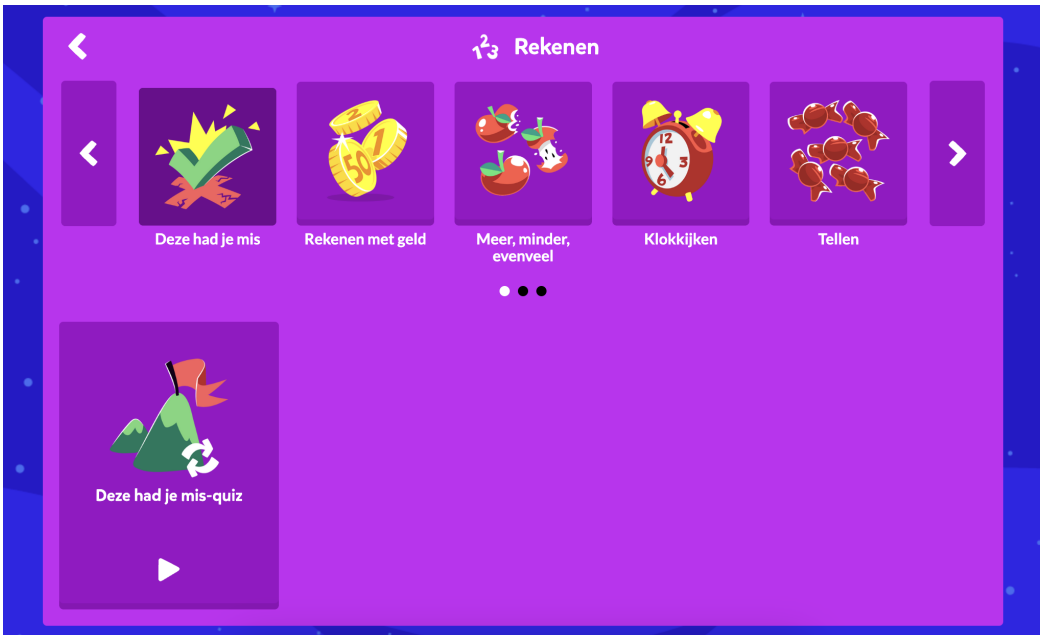

Figure 4. *The new quiz called 'Deze had je mis-quiz' shown within the mathematics section on Squla. The categories are on top, the quiz's own category 'Deze had je mis' is the first in the list.*

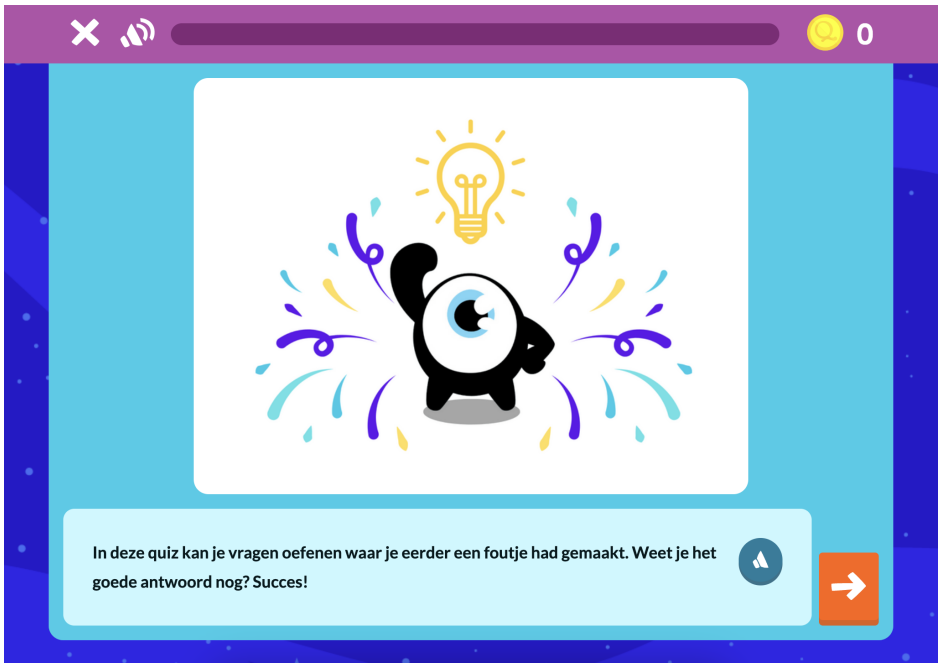

Figure 5. *The introduction screen of the review quiz, shortly explaining to users what the quiz is about.*

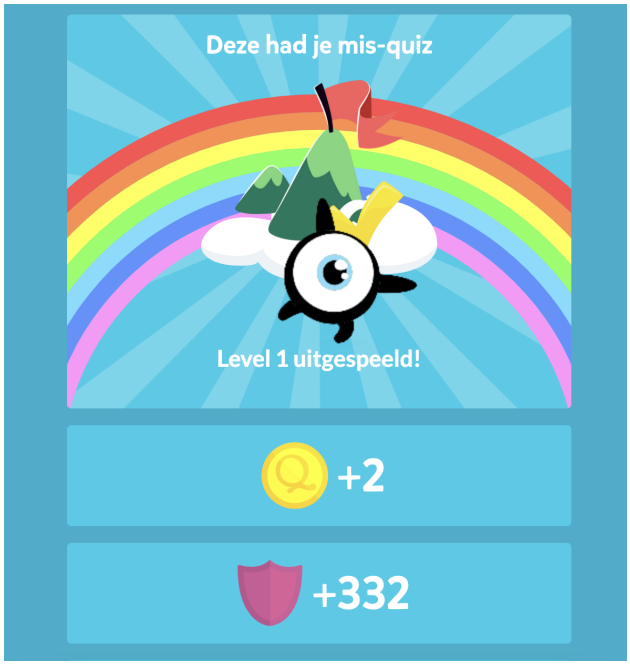

Figure 6. *The end screen shown after the review quiz is completed, displaying the amount of coins and XP earned.*

The quiz centered around revisiting mistakes as we assumed these questions would be more difficult for users and thus offer the greatest opportunity for learning, making them suitable to detect knowledge increase. To minimize frustration from overly difficult questions, we sampled mistakes only from quizzes that users had completed, based on the assumption that they would not finish a quiz if all questions were too challenging. Additionally, the quiz was kept shorter than typical quizzes on Squla, containing 5 questions instead of the usual 10, to prevent fatigue from answering too many difficult questions.

Performance on the review quiz was analyzed to determine whether practicing specific topics on Squla improves performance. We defined topics as the categories on the platform in which content was divided, as these contained quizzes covering similar concepts or skills. The accuracy of answers on the review quiz was compared between two groups: (1) answers by users who practiced the same category in which they had made a mistake (on-topic practice) and (2) answers by users who practiced other categories within the same subject (off-topic practice). An example of two cases for which performance would be compared is shown in Figure 7. Both user 1 and user 2 made a mistake in the percentages category of math, represented by the red error icon. User 1 then practiced four other math quizzes, none of which

were from the percentages category. User 2 also practiced four math quizzes after the mistake, but one of these was from the percentages category. Both users then took the review quiz, represented by the mountain icon, which included the question on percentages where they had made a mistake. Their accuracy on this question would be compared to assess whether practicing the relevant topic, percentages, leads to better performance than practicing unrelated topics. If so, this would suggest that practicing specific topics on Squla improves performance in those areas.

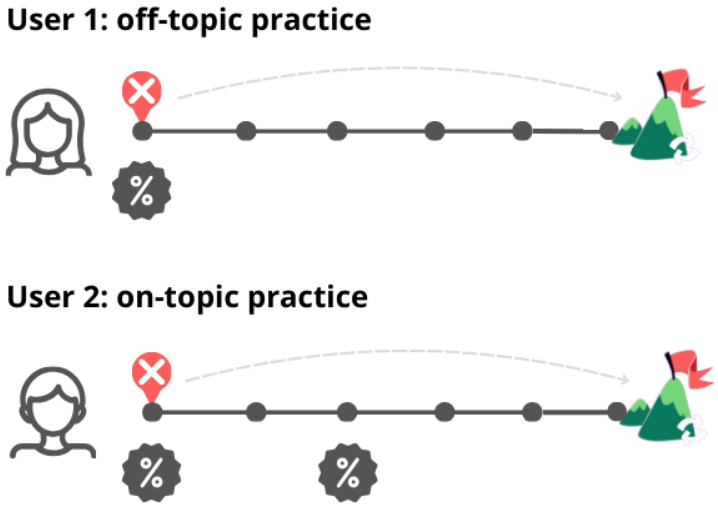

Figure 7. *Example timelines of users in the off-topic and on-topic practice groups. Each circle represents a math quiz taken by the user, with the final quiz being the review quiz. The percentage icons indicate when the taken quiz belongs to the percentages category.*

3.2.2 Data analysis

The performed experiment and analysis involved several steps, as outlined in the diagram in Figure 8. This section briefly covers steps 1-4, followed by two subsections that offer a more detailed explanation of steps 5 and 6, focusing on the matching process and the creation of the mixed model. Step 7, which covers the analysis of the results, is presented in the results section thereafter. All analyses were performed using R Statistical Software (v4.4.0; R Core Team 2021).

Users were able to take the review quiz in a subject if they had made at least five mistakes in that subject over the previous four weeks. The quiz questions were randomly sampled from the user's pool of mistakes within the subject during this period. A user could access the review quiz for multiple subjects, provided they had enough mistakes in each subject, but each quiz

was specific to one subject and only contained questions from that subject. The review quiz could include questions from multiple categories within a subject, depending on whether the user had made mistakes in one or more categories. As a result, users could belong to both the on-topic and off-topic practice groups, depending on the question. For instance, if a user's review quiz contained three questions from topic A and two from topic B, their answers were classified into the respective groups based on whether they had practiced those topics again before taking the review quiz. Therefore, a single user's answers could be classified into both the off-topic and on-topic practice groups. A user was considered to have practiced on-topic, if they had answered at least 10 questions on the topic in between making the mistake and answering it again on the review quiz. The grade in which a topic was practiced did not affect the classification, practicing the relevant topic in either grade 4 or grade 5, for example, would both be considered on-topic practice.

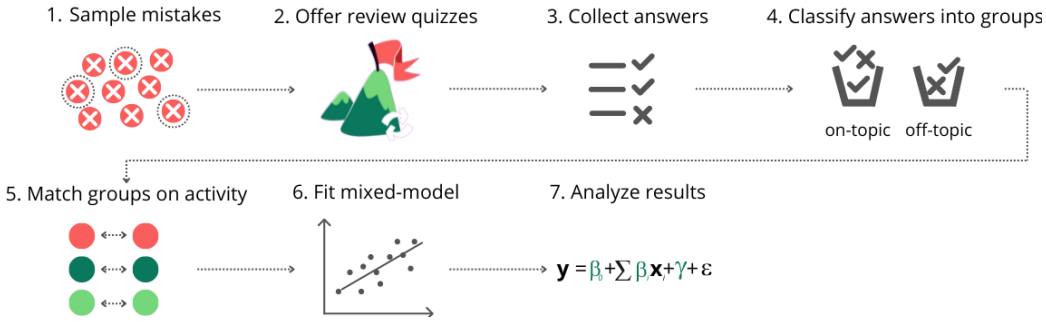

Figure 8. *Overview of the data analysis process, illustrating the key steps from sampling mistakes for the review quiz to analyzing the results from the mixed model.*

3.2.2.1 Matching

Score differences between the groups cannot immediately be compared, as users are not randomly assigned to groups and therefore this study is observational in nature. Users who practiced on-topic are likely more active than those who did not, as the latter group may have practiced less or not at all, potentially due to factors such as motivation. To mitigate the bias that high engagement may have on performance, answers from the on-topic practice group will be compared only to those from the off-topic practice group from users with similar activity patterns, specifically in terms of the number of questions answered. This approach ensures a fair comparison even when randomization is not possible.

Matching sampling was used to create comparable groups and replicate the conditions of a randomized experiment as closely as possible. This method pairs units from the control and intervention groups based on similar values of observed covariates, ensuring that the groups are only randomly different from one another on those covariates (Stuart, 2010). By doing so, matching reduces bias and helps isolate the true effect of the intervention, in this case, practicing specific topics on Squla. In this study, three key covariates were defined to capture user engagement:

1. **practiceBefore**: The total number of questions practiced across the entire platform in the month before the experiment. This variable is intended to reflect motivation and practice habits.
2. **practiceDuring**: The number of questions practiced from the relevant subject, between the mistake and the review quiz. This captures subject-specific engagement during the study.
3. **practiceDays**: The time in days between the mistake and the review quiz, to fairly compare the amount of practice over different periods.

To improve the accuracy and reliability of the matching process, outliers in the covariates were identified and removed before matching. A 95th percentile threshold was applied for outlier removal, which minimized the presence of extreme values, resulting in a more balanced distribution. By reducing the influence of outliers, the covariates were more evenly distributed across groups, facilitating better matching.

We matched individual observations, specifically users' answers to the review quiz, rather than participants themselves, as this is the level of data used in the rest of the analysis. This approach was necessary as the review quiz could contain questions from multiple categories, meaning a user could be in the on-topic practice group for one category but not for another. Users will have also made mistakes at different times, causing the time between making a mistake and taking the review quiz to vary for each question. As a result, observations were matched based on the covariates related to each answer of a user. For multiple answers from the same user, the amount of practice before the experiment would be the same, but the time and amount of practice between the mistake and the review quiz, as well as whether the topic was practiced again, could differ.

The matching process was performed using the MatchIt package in R (Ho, Imai, King, & Stuart, 2011). Several methods were explored in order to find the best balance across covariates. Nearest Neighbour matching was tried first, which selects for each treated unit the closest control unit based on a distance measure. Another method is exact matching, which only matches units with identical covariate values, offering the best balance and making it

the most powerful method. Although it is the ideal method, the downside is that many observations will be left unmatched (Stuart, 2010), especially with continuous covariates. To address this, Coarsened Exact Matching (CEM) was ultimately chosen as it groups similar values of each variable into bins and then applies exact matching to these coarsened covariates (Iacus, King, & Porro, 2012). This method provided good balance, ensuring a similar distribution of covariates across groups while maintaining an adequate sample size. The bin selection method Sturges's rule was used to select the number of bins for covariates *practiceBefore* and *practiceDays*. For the amount of practice during the experiment, the distribution was divided into 50 bins, as this produced the best balance for that covariate. Sturge's rule was not optimal for *practiceDuring* due to its highly skewed distribution before matching.

The balance of covariates was assessed using balance summary statistics for each covariate, namely: the standardized mean difference (SMD), being the difference in the means between groups divided by a standardization factor so that it is on the same scale for all covariates, the variance ratio, which is the ratio of the variance of a covariate in one group to that in the other, and eCDF statistics, the mean and largest difference in the cumulative distributions of the covariates between the groups. Standard mean differences and eCDF values close to zero indicate good balance, as well as variance ratios close to 1 (Ho et al., 2011).

Matching was done per subject, as this research aimed to explore differences in the learning effect across individual subjects. This approach ensures that the effect is estimated fairly for each subject, avoiding bias that could result from imbalanced sample distributions within subject groups. Matching was also performed per age group, as this is an important factor for fair comparison. Answers given by, for example, children in first and eighth grade are not suitable for comparison, as the type of material and its educational level vary greatly, which could otherwise lead to biased results.

The balance of the sample for language in grades 1-3 before matching is shown in Table 1, and the balance after matching is presented in Table 2. The tables indicate that the engagement covariates were severely unbalanced before matching, especially for the *practiceDuring* covariate. However, the matching process was successful, as a good balance between the on-topic and off-topic practice groups was achieved afterward. The standardized mean differences for all covariates were close to zero, indicating minimal differences between the groups after matching. Additionally, the variance ratios were close to one, suggesting similar variability across groups, while the low eCDF mean and eCDF max values indicate similar distributions between the two groups. The mean practice period in days after matching was 11 days for

both groups. The mean number of questions answered during this time was 48. There was a slight difference of 2 questions for the amount of practice the month before the experiment, but the summary statistics showed this covariate is still well-balanced.

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	12.41	10.38	0.22	1.27	0.06	0.10
practiceDuring	93.27	35.55	0.70	3.36	0.15	0.41
practiceBefore	247.03	218.62	0.09	1.30	0.03	0.06

Table 1. *Summary of balance of the sample of language answers to the review quiz in grade 1-3 before matching. Showing the means for the on-topic and off-topic practice groups, standardized mean differences, variance ratios, and eCDF measures.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	11.05	11.22	-0.02	0.99	0.01	0.02
practiceDuring	48.40	47.95	0.01	0.99	0.00	0.05
practiceBefore	175.24	172.67	0.01	0.98	0.01	0.05

Table 2. *Summary of balance of the sample of language answers to the review quiz in grade 1-3 after matching. Showing the means for the on-topic and off-topic practice groups, standardized mean differences, variance ratios, and eCDF measures.*

A visual inspection of covariate balance was also conducted using empirical Quantile-Quantile (eQQ) plots, which compare the empirical distributions of each variable across the groups (Stuart, 2010). The eQQ plot for the language sample in grades 1-3 is shown in Figure 9. The plot shows on the y-axis the value of each covariate for the on-topic practice group, and on the x-axis the value at the corresponding quantile for the off-topic practice group. Ideally, after matching, the points should fall on the 45-degree line, indicating similar distributions of covariates between the groups. Figure 9 shows the points for this sample fall closely along the 45-degree line after matching.

Thus, both the summary statistics and the eQQ plot show the bias in the sample was effectively reduced, ensuring that any observed differences in outcomes are less likely to be influenced by these engagement covariates. The summary statistics for other subjects and age groups were similar to those presented here and are provided in Appendix A.

It is important to note that not all observations in the original sample could be matched, resulting in a reduced sample size. The total sample size

across all subjects and age groups, before and after matching, is shown in Table 3.

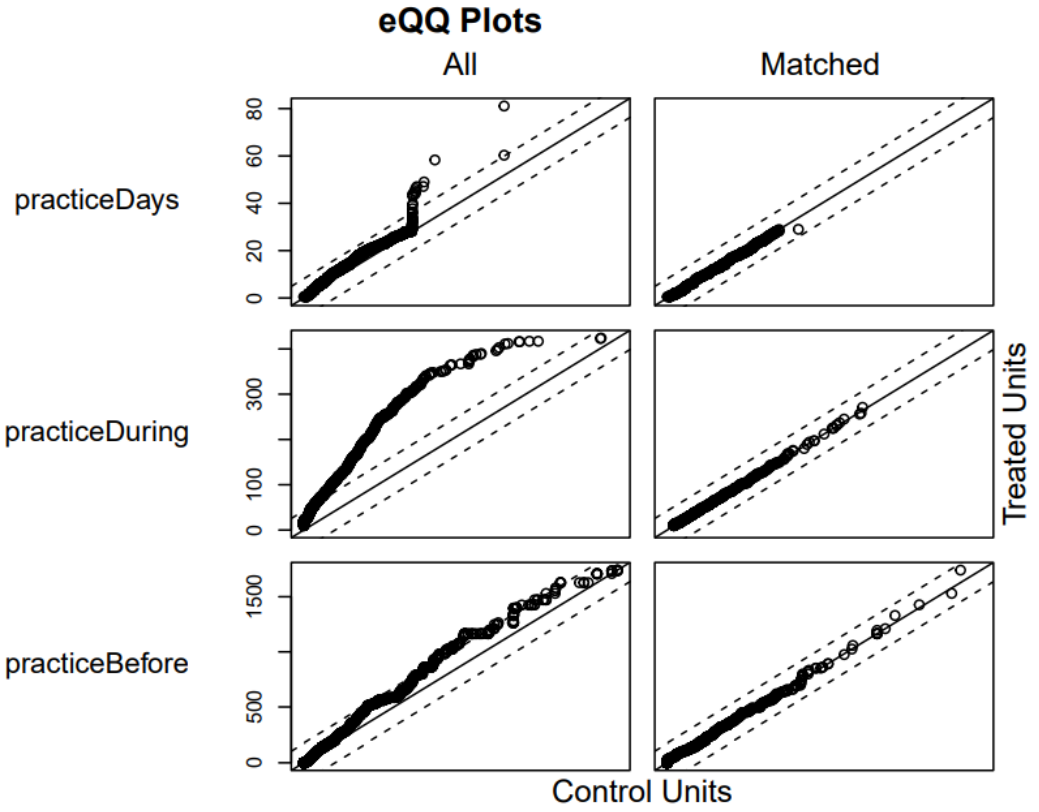

Figure 9. *EQQ plot of the sample for language in grades 1-3, showing the balance of the three covariates before and after matching.*

	Off-topic practice group	On-topic practice group
Unmatched	31,282	50,028
Matched	19,319	19,319

Table 3. *Number of answers to the review quiz in off-topic and on-topic practice groups before and after matching.*

3.2.2.2 Mixed model analysis

The analysis of the difference in performance on the review quiz between the off-topic and on-topic practice groups after matching was conducted using a generalized linear mixed model (GLMM). The model estimated the probability that a review quiz answer would be correct, based

on whether the topic of the question was practiced, and which subject the question belonged to. A GLMM with a logit link function was chosen because the outcome variable was binary: 0 for an incorrect answer and 1 for a correct answer. This function models the probability of a correct answer by transforming it into log odds, allowing for a linear relationship between the predictors and outcome (Brown, 2021). A mixed model was needed to account for the possibility of repeated measures, as users could answer between 1 and 5 questions on the review quiz, or even take the quiz multiple times. Therefore, user ID was included as a random effect, capturing the variability in performance across individual users.

The formula for the model is as follows:

$$correct \sim practicedCategory * subject + (1|userID)$$

Here, the fixed effect *practicedCategory* specifies whether or not the topic of the question was practiced, and *subject* represents the school subject the question belonged to. The formula includes an interaction as we aimed to explore whether the effect of practicing a topic differs across subjects. The interaction allowed the model to account for the possibility that practicing might improve performance more in some subjects than others, rather than assuming the effect is the same for all subjects.

The model was fitted using the *glmer*() function from the lme4 package (v1.1-35.5; Bates, Mächler, Bolker, and Walker 2015). The model structure was selected by comparing models using likelihood ratio tests performed with the *anova*() function in R. These tests showed that including *userid* as a random effect significantly improved the model's fit compared to simpler models without random effects, as indicated by lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. Model diagnostics were performed to evaluate the validity of the final logistic mixed model. The key assumptions checked were the normality of the random effect distribution, homogeneity of variances, and the presence of overdispersion (Bolker et al., 2009). The homogeneity of variances across groups and overdispersion were evaluated using the DHARMa R package by Hartig (2022), with no issues found. The normality of the random effect distribution was evaluated using a Q-Q plot (Figure 10), which showed that the random effects followed an approximately normal distribution, with slight deviations at the tails. Overall, all assumptions were met, confirming that the model is reliable for representing the data.

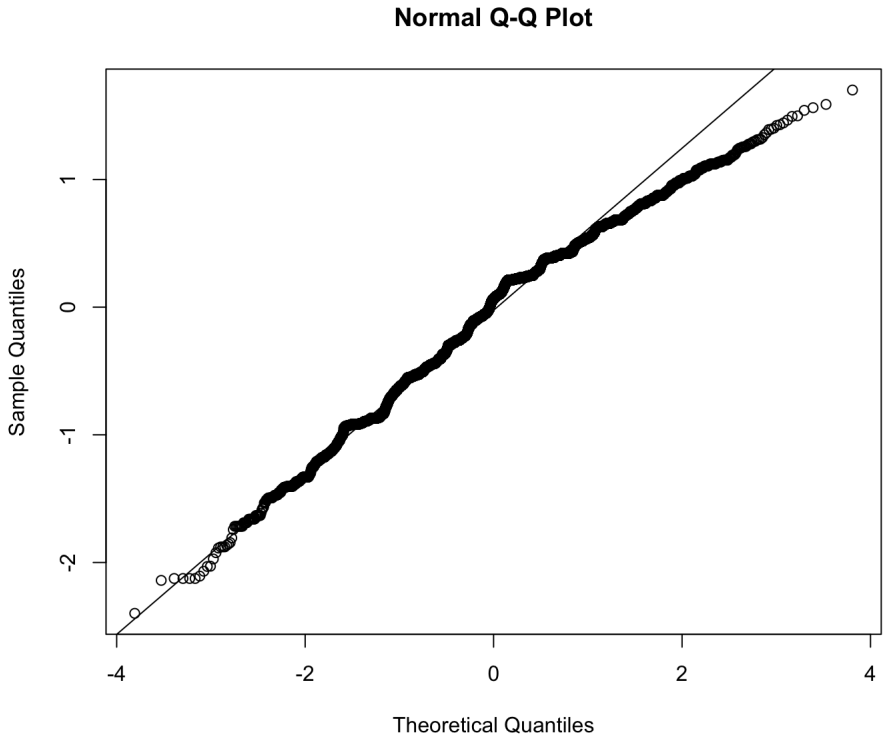

Figure 10. *Q-Q Plot for random effects (userid) showing approximate normality.*

3.3 Results

This analysis aimed to determine whether prior practice on a topic influenced performance on this topic on the review quiz, which consisted of previously missed questions. A total of 81,310 answers to the review quiz were recorded, with 31,282 answers from the off-topic practice group and 50,028 from the on-topic practice group. After matching, the number of responses in both groups was reduced to 19,319 each, resulting in 38,638 quiz responses from 7,154 participants being included in this analysis. The participants were primary school-aged children from the Netherlands. Table 4 shows the distribution of answers per grade on Squla, and Table 5 shows the number of answers given per subject across all grades.

A generalized linear mixed model was used to explore the effect of on-topic practice on the performance on the new quiz. The model was fitted using maximum likelihood estimation with the Laplace approximation. It predicted the probability of a correct answer on the new quiz based on whether the user had practiced the relevant topic and the subject to which

Grade	Amount of answers
1	4351
2	6253
3	9534
4	6108
5	4943
6	4841
7	1770
8	838

Table 4. *The number of answers per grade after matching.*

Subject	Amount of answers
Math	24098
Language	7098
Spelling and grammar	3042
Reading comprehension	4400

Table 5. *The number of answers per subject after matching.*

the question belonged. Random intercepts for individual users were included to account for repeated measures and to reflect variability in individual performance. The model had an AIC of 47,319 and a BIC of 47,396, indicating a good fit compared to alternative tested models.

The fixed effects are summarized in Table 6. The default for *practicedCategory* was 'No', indicating the off-topic practice group, and the default subject was reading comprehension. The results indicated a significant interaction between *practicedCategory* and *subject*, as the difference in the estimate of *correct* between on-topic and off-topic practice varied across subjects. The random intercept for user ID had a standard deviation of 0.95, indicating

Fixed effect	Estimate	SE	z value	p
(Intercept)	0.53	0.06	9.06	< 0.001
practicedCategoryYes	-0.09	0.07	-1.25	0.21
subjectMath	0.25	0.06	4.02	< 0.001
subjectSpelling	0.21	0.09	2.40	0.02
subjectLanguage	0.32	0.07	4.46	< 0.001
practicedCategoryYes:subjectMath	0.25	0.08	3.10	0.002
practicedCategoryYes:subjectSpelling	0.26	0.12	2.27	0.02
practicedCategoryYes:subjectLanguage	0.22	0.09	2.37	0.02

Table 6. *Fixed effects from a logistic mixed-effects model of quiz response accuracy on Squla. The reference subject is reading comprehension and the reference level for practicedCategory is 'No'.*

significant variability in the baseline performance across participants.

Post-hoc comparisons using estimated marginal means (EMMs), calculated using the emmeans package (v1.10.4; Lenth 2024), were performed to further explore the effect of practicing on- or off-topic across different sub-

jects, as shown in Table 7. These estimates have been converted to probabilities for easier interpretation, while those on the log-odds scale are included in Table 37 in Appendix B. For math, language, and spelling and grammar, the probability of a correct response increased when users practiced the relevant category. For math, the probability of answering correctly increased from 69% for off-topic practice to 72% for on-topic practice. For language, the probability of answering correctly increased from 70% for off-topic practice to 73% for on-topic practice. The estimated probabilities for spelling and grammar are 68% for off-topic practice and 71% for off-topic practice. In contrast, for reading comprehension, the probability of answering correctly decreased slightly from 63% for off-topic practice to 61% for on-topic practice.

Subject	**Practice kind**	**Probability**	**SE**	**Lower 95% CI**	**Upper 95% CI**
Math	off-topic	0.69	0.01	0.67	0.70
Math	on-topic	0.72	0.01	0.71	0.73
Language	off-topic	0.70	0.01	0.68	0.72
Language	on-topic	0.73	0.01	0.71	0.74
Spelling and grammar	off-topic	0.68	0.01	0.65	0.70
Spelling and grammar	on-topic	0.71	0.01	0.68	0.74
Reading comprehension	off-topic	0.63	0.01	0.60	0.65
Reading comprehension	on-topic	0.61	0.01	0.58	0.63

Table 7. *Estimated marginal means (EMMs) on the probability scale, showing the estimated probability of a correct response for on-topic and off-topic practice for each subject.*

To test whether the differences in accuracy between the on-topic and off-topic practice groups were significant, contrasts of the EMMs were calculated. These compare the estimated effects between the groups for each subject and are shown in Table 8. The contrasts highlight that the benefits of practicing topics on Squla depend on the subject. The 3% difference for math was statistically significant ($p < 0.001$), as well as the 3% difference for language ($p = 0.03$). The difference of 3% for spelling and grammar, however, was not significant ($p = 0.06$). The difference in accuracy for reading comprehension also was not statistically significant ($p = 0.21$).

Contrast	**Subject**	**Estimate**	**SE**	**z-ratio**	**p-value**
off-topic - on-topic	Math	-0.16	0.03	-4.62	< 0.001
off-topic - on-topic	Language	-0.13	0.06	-2.21	0.03
off-topic - on-topic	Spelling and grammar	-0.17	0.09	-1.90	0.06
off-topic - on-topic	Reading comprehension	0.09	0.07	1.25	0.21

Table 8. *Contrasts of EMMs on the log-odds ratio scale.*

To assess the size of the effects found for math and language, odds ratios were computed, which are a commonly used measure of effect size in logistic regression models (Field, Miles, & Field, 2012). Table 9 presents the odds ratios for the contrasts between the off-topic and on-topic practice groups. The odds of a question being answered correctly on a math review quiz were 14% lower after off-topic practice than after on-topic practice (OR = 0.86). Similarly, for language questions, the odds of a correct answer were 12% lower for the off-topic practice group than for the on-topic practice group (OR = 0.88).

Contrast	**Subject**	**Odds Ratio**	**Lower 95% CI**	**Upper 95% CI**	**p-value**
off-topic - on-topic	Math	0.86	0.80	0.91	< 0.001
off-topic - on-topic	Language	0.88	0.78	0.99	0.03
off-topic - on-topic	Spelling and grammar	0.84	0.70	1.01	0.06
off-topic - on-topic	Reading comprehension	1.09	0.95	1.26	0.21

Table 9. *Odds ratios for contrasts of EMMs.*

3.4 Discussion

This experiment served as a trial for a new embedded approach to evaluating the effectiveness of an educational platform that features quizzes. The goal of the experiment was to determine whether practicing specific topics on Squla helps children improve their performance on those topics. To measure this, a new quiz was introduced, containing questions that users had previously answered incorrectly. The accuracy of answers to the quiz was compared between two groups: those given by children who practiced the relevant topic of the question (on-topic practice) and those given by children who practiced other topics within the same subject (off-topic practice).

The analysis used a mixed model to account for both fixed and random effects. The variability in individual performance was substantial, with a standard deviation of 0.95. This is likely due to factors such as prior knowledge, effort, or focus, which can significantly impact performance.

A significant interaction effect was observed between the predictors *practicedCategory* and *subject*, indicating that the impact of practice type varied depending on the subject. Specifically, for math and language, on-topic practice resulted in a 3% higher likelihood of correct answers compared to off-topic practice. However, no significant difference in performance was found for spelling and grammar, and reading comprehension based on practice type. The odds ratios for math and language showed that children who

practiced off-topic material had a 14% and 12% lower odds of answering correctly, respectively, compared to those who practiced on-topic material. While these effects are moderate, they demonstrate that targeted, on-topic practice on Squla significantly improves performance in math and language.

The lack of significant effects for spelling and grammar, and for reading comprehension was further reflected in the 95% confidence intervals of their odds ratios, which included 1, suggesting no meaningful difference in odds between practice groups. Interestingly, despite the insignificance, the estimated odds ratio for spelling and grammar was the most extreme among all subjects. Additionally, the estimated marginal means suggested that the difference in performance between off-topic and on-topic practice for spelling and grammar was as large as that for math and language. One possible explanation for the insignificance of the effect in spelling and grammar, despite the comparable estimates, is the smaller sample size for this subject, which was the lowest across all subjects (as seen in Table 5). A smaller sample size can lead to greater uncertainty around estimates, reducing the power to detect effects that may be present. Another possible factor is the greater variability in performance for spelling and grammar, as suggested by the larger standard error (SE) of the contrast between practice groups. This larger SE indicates greater variability in the difference between practice and no-practice groups, which may have contributed to the non-significant result.

The different EMMs observed for reading comprehension compared to other subjects suggest that on-topic practice may not improve performance in the same way. This can be explained by the unique characteristics of reading comprehension content on Squla. Unlike other subjects where categories group quizzes on related concepts or skills, the categories in reading comprehension are based on the topics of the texts, such as stories about animals or hobbies. Therefore, practicing quizzes from the same category (on-topic practice) does not necessarily reinforce specific skills in the same way it does in other subjects. This subject was thus less suitable for the experiment as on-topic practice should be no more effective than off-topic practice, explaining the lack of improvement seen in reading comprehension performance.

The EMMs for reading comprehension not only suggested an opposite effect to what was expected, but were also generally lower than those for other subjects. This difference may be due to fatigue, as the review quiz consists of randomly sampled mistakes, leading to potentially different texts being presented in a single quiz. Questions requiring text analysis are likely more time-consuming than, for example, simple arithmetic math problems. Having to answer multiple long questions could lead to frustration or lower effort

from students. Another possible explanation is that reading comprehension questions on Squla are inherently more difficult than those in other subjects, or that students generally struggled more with this subject.

The research question we aimed to address was: "Can the effectiveness of an online learning platform be measured by comparing student performance on topics after on-topic practice versus off-topic practice within the platform?" The results obtained from the Squla experiment provide evidence that this approach can indeed be used to measure platform effectiveness in terms of knowledge gain. By comparing performance on a review quiz between students who practiced on-topic content and those who practiced off-topic content, the impact of targeted practice within the platform was assessed.

The findings showed that on-topic practice led to improved accuracy in math and language, indicating that quizzes in these subjects were effective in enhancing knowledge on specific topics. No significant effects were found for spelling and grammar or reading comprehension, suggesting that practicing specific topics on Squla may not consistently improve performance across all subjects. These findings did not fully align with our initial expectations. While we anticipated that on-topic practice would lead to better accuracy across all four subjects, significant improvements were observed only in math and language. We also expected the strongest effect in math, given that basic math facts can be effectively learned through repetitive practice. However, the significant effects for math and language were similar, with only a slightly higher odds ratio for math. This suggests that both subjects are equally effective and well-suited to the quiz-based learning format on Squla. The absence of a significant effect for spelling and grammar may indicate that this subject requires a different type of instruction or more extended practice to achieve measurable improvements.

The results suggest that the proposed method can measure the effectiveness of practicing quizzes on the platform in terms of learning outcomes. This has practical implications for other digital learning platforms, as it demonstrates how embedded experiments can be used to evaluate feature effectiveness without disrupting the user experience or requiring user feedback. By passively collecting data through embedded experiments, platforms can more efficiently assess their efficacy, especially compared to traditional in-person methods. This approach also allows for a broader reach of the target audience, as users are automatically included in the experiment without the need for active recruitment. However, the lack of random assignment in this study required the use of matching, which reduced the available data for analysis. Despite this limitation, embedded experiments remain an attractive,

cost-effective alternative for continuously evaluating educational technology, which is essential as these platforms evolve.

3.4.1 Limitations

This experiment was limited to four subjects within Squla, although the platform offers content for many other subjects. The review quiz was introduced selectively to a few subjects in order to gauge its impact on general engagement. As a result, we could not draw conclusions about the overall effectiveness of the quiz feature on Squla. To fully evaluate this feature, a broader experiment that includes all other subjects would be necessary. Similarly, no conclusions can be made about the platform as a whole, as it offers other features besides quizzes.

Additionally, while this design could theoretically be adapted to other platforms that feature quizzes, there may be practical challenges. Implementing the experiment on another platform would require identifying quiz topics, sampling users' mistakes, and integrating a new quiz into the platform. The feasibility of adapting this method therefore depends on the type and availability of data, as well as the similarity of the platform's quiz feature to the one on Squla. However, the design used in this study could serve as an example of an embedded experiment and offer inspiration for developing similar, yet tailored approaches.

Although the experiment reached a large number of users, it is unclear whether these participants are representative of Squla's entire target audience. Participants were automatically included if they played the review quiz, which may have attracted users with different characteristics, such as curiosity about new features. Additionally, the inclusion into on-topic and off-topic practice groups could be influenced by participant motivation. For example, users practicing on-topic may have been more motivated to learn about specific topics. Alternatively, users less proficient in certain topics may have been more likely to practice those topics. These underlying factors could not be fully controlled. However, through matching, we did control for user activity levels, ensuring both groups were similarly active on the platform. All participants also initially practiced the relevant topic when they made the mistake, regardless of subsequent practice behavior. This suggests that all participants had some initial motivation to engage with the topic, which helps alleviate concerns that motivation alone could explain the observed differences between practice groups.

The possibility of uncontrolled user characteristics is an important consideration, as unmeasured background covariates related to group assignment could introduce bias into the results (Stuart, 2010). We took care in defining

the variables used for matching, ensuring the inclusion of engagement metrics that reflect both general motivation and practice habits, and subject-specific engagement. While balance on these covariates was generally good, as indicated by summary statistics, there were minor imbalances. The largest difference in means between the on-topic and off-topic practice group was 13 questions, for the covariate *practiceBefore* in grades 1-3 of spelling and grammar. However, this covariate reflects engagement prior to the experiment, which is less likely to be associated with group assignment than engagement during the experiment.

Another limitation concerns the definition of 'on-topic' practice. The expectation was that after on-topic practice, performance on that topic would improve as children deepened their understanding. Quizzes on the same topics were defined as those within the same category. However, not all categories on Squla are equally coherent. For example, the math category 'shapes and figures' in grades 2 and 3 contains quizzes that focus on distinct skills, such as recognizing shapes and learning prepositions related to spatial orientation. Practice in such categories was still considered on-topic practice, which may have diluted the observed effect, as practice in these categories may not improve understanding on the topic from the previous mistake. Additionally, practice in the same category across different grades also counted as on-topic practice. While this is reasonable for grades that are close together, the relevance of practice reduces when comparing grades that are farther apart, such as grade 1 and grade 8. It is difficult to set a strict boundary for how far apart grades can be while still counting as relevant practice, as this likely varies by topic. However, since there are few categories that overlap between grades far apart on Squla, this issue likely had minimal impact on the results.

Finally, our interpretation of the non-significant effect for spelling and grammar raised concerns about whether the analysis had sufficient statistical power to detect an effect in this subject. We did not perform a formal power analysis before the experiment, which could have informed us if the sample size was adequate to detect the expected effect. A simulation-based power estimation for mixed models, such as the approach described by Kumle et al. (2021), would have been a suitable method. However, this requires reliable estimates of effect sizes and variability in the random effects. Due to the novel design of this experiment, there was insufficient empirical data to provide such estimates beforehand. As a result, we ran the experiment for as long as possible to maximize the sample size and improve the power of the analysis.

3.4.2 Future work

Although the current experiment provided valuable insights for future embedded efficacy experiments, further research is needed to address the identified limitations and to refine the methods used, especially for subjects like spelling and grammar, and reading comprehension. For reading comprehension, the current experimental design should be adjusted, as it was not fully suitable for evaluating this subject. A more refined approach would involve identifying quizzes that focus on similar underlying skills, instead of relying on category-based grouping. On-topic practice could then be redefined as practicing quizzes on the same specific topic, rather than in the same category. With this new definition, it may be possible to reanalyze the existing data to determine whether genuine on-topic practice leads to better accuracy in related questions on a reading comprehension review quiz.

For spelling and grammar, conducting a formal power analysis is necessary to determine the minimum sample size required to detect an effect in a new experiment. The estimates from this study could be used to estimate effect size and variability. The current experimental design could be replicated but extended over a longer period to collect more data if needed, or a new design could be devised. One possible approach would be to include short pre- and post-tests around spelling and grammar quizzes to better capture improvements. However, such a new experiment must be carefully designed to account for the potential biases discussed earlier.

While this study explored whether the impact of practice on Squla varies across subjects, there are other interesting groupings to consider for further investigation. For example, the observed effect of on-topic practice may differ by grade, as the material and topics covered vary significantly between grades. An attempt was made in this study to measure the effect of on-topic versus off-topic practice per subject and age group. Three age groups were defined, combining grades 1-3, 4-6, and 7-8. A mixed model similar to the one used in this study was fitted, incorporating age group as a predictor. However, the sample sizes per age group and subject were too small to draw conclusions. Although the interaction effect was significant and the differences in estimated marginal means were substantial, they were not statistically significant for many of the groups. This suggests that future research in this area is worthwhile, as the preliminary results indicate that the effect of on-topic practice on performance may vary by age group.

Another promising direction for future research with this experiment design in Squla is to explore whether certain topics of the math and language quizzes are more effective than others. If significant differences in the probability of answering correctly after on-topic practice are found across topics,

the characteristics of those topics should be investigated. This knowledge could benefit Squla and other educational platforms by helping to identify how certain topic attributes, such as complexity, influence the impact of practice and how the content can be improved.

4 Online learning platform StudyGo

The second platform that supported this research as a case study is StudyGo. For this platform, a separate embedded experiment was conducted to evaluate the effectiveness of the practice questions. The following sections describe the platform, methods, and results of the experiment.

4.1 System description

StudyGo is an educational platform aimed at secondary school students aged 12 to 18 and is commercially available[5]. It is directed to at-home use and content is available for most school subjects. The platform offers a broad array of features, including vocabulary learning, sets of practice questions, practice tests, explanation videos, and summaries. Additionally, StudyGo provides interactive support through online tutoring via a chatting function and videocall lessons, and a Q&A forum.

A key aspect of StudyGo is its collaboration with recognized publishers to offer content aligned with schoolbooks used in secondary schools in the Netherlands. Students can select the books used in their classes and the platform content will directly match this material, allowing them to easily practice and test the relevant topics. When navigating to a chapter of a book, the user can find practice exercises, explanation videos, summaries, and a test for each topic, as shown in Figure 11. While attempting a practice set, users can see information about their progress, as shown in Figure 12. The topic summary and explanation video can also be accessed from this screen, and the correct answer will be given if a question is answered incorrectly. At the end of a quiz, users can view explanations of the correct answer to each question.

StudyGo also ensures content matches educational standards as it is aligned with the core targets as specified by SLO[4]. On top of the offered content, students can also create their own quizzes and word lists to practice and share with others.

4.2 Methods

4.2.1 Experiment design

Similar to Squla, the StudyGo platform records time-stamped engagement and performance data, including login times, interactions with platform

[5]https://studygo.com/

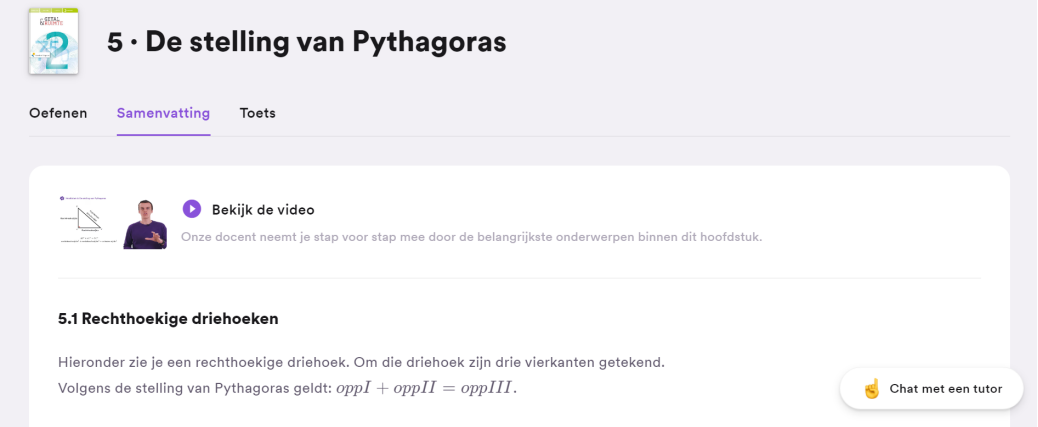

Figure 11. *This is the summary page of a math chapter on StudyGo. Users can watch an explanation video or read the chapter summary. Tabs above the video allow navigation to practice questions or the chapter test. A 'Chat with a tutor' button is located at the bottom right.*

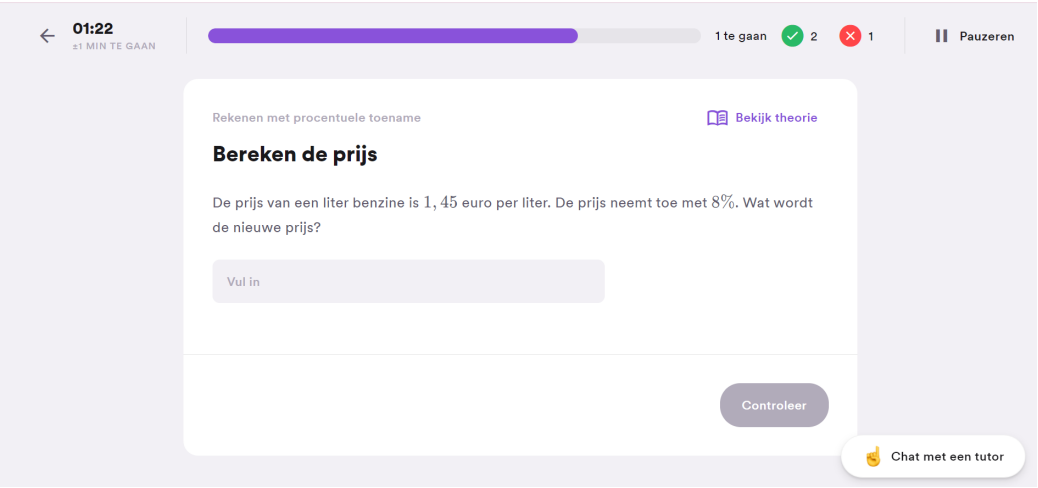

Figure 12. *View while attempting a practice set on StudyGo. The top left shows elapsed time and estimated remaining time. A progress bar is shown in the center next to the number of remaining questions and the counts of correct and incorrect answers. The 'Bekijk theorie' button allows users to access the explanation video and summary.*

features and their duration, response times, and response accuracy. This data was used to assess the platform's effectiveness. Several experimental designs were explored for this evaluation, including pre-post assessments using the platform's summative tests to measure the impact of completing practice questions on performance. However, requiring students to complete the tests and practice questions in a fixed order was not feasible, as student autonomy in managing their study activities is a key feature of StudyGo. Some

students naturally followed this order when using the platform, but using this participant group could have led to self-selection, attracting a type of student with characteristics that differ from others.

The chosen method for measuring the effectiveness of StudyGo was based on the assumption that practice questions facilitate learning. This would be reflected by fewer mistakes on questions toward the end of a set, as students learn from earlier questions. To test this hypothesis, a simple experiment was devised where the first and last questions in each set were swapped for half of the users. This allowed for a comparison of the error rates, defined as the proportion of incorrect answers, between the first and last positions for the same questions. Students were randomly assigned to one of two groups: one group received the normal order of question sets, while the other group received the swapped order. Figure 13 illustrates the setup, showing both the normal order of questions (from 1 to n) and the swapped order, where questions 1 and n are swapped. In this example, the base error rates in the first position are different, reflecting potential differences in the inherent difficulty of the questions. However, for both question 1 and question n, error rates are lower when the question is placed at the end of the set. If the error rates are consistently lower at the end, this would suggest that students are learning from the practice questions, confirming the effectiveness of StudyGo's practice questions. The experiment was live on the platform for nearly four months, from the beginning of April to the end of July, during which data was collected.

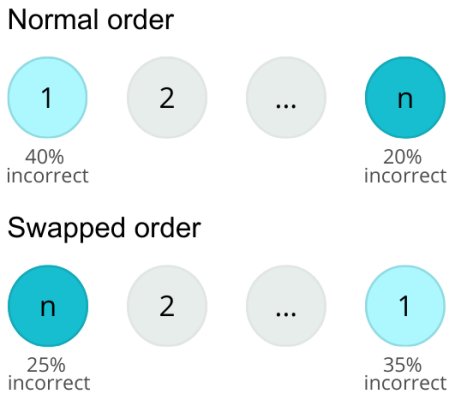

Figure 13. *Illustration of the expected effect in the StudyGo experiment. Each circle represents a question. The error rates of both question 1 and question n are lower when placed in the last position compared to the first.*

4.2.2 Data analysis

All analyses were performed using R Statistical Software (v4.4.0; R Core Team 2021). Questions were answered by participants from both the normal and swapped order groups. This ensured that each question was answered in both positions within a set, allowing for a fair comparison of error rates without introducing bias related to question characteristics, such as difficulty. Such bias might have occurred if we had only compared the error rate of the first and last questions of sets in their normal order, since the questions at the end of the set could be significantly different from those at the beginning. For example, the last questions could be easier, leading to lower error rates that might be mistakenly attributed to a learning effect. Therefore, the error rates from different positions within the set were only compared for the same questions.

4.2.2.1 Filtering

One condition for our hypothesis that students learn from earlier questions, is that questions within each practice set address the same topic or skill. Initial inspection of StudyGo practice sets revealed that many of them contained questions on related but distinct topics. For example, a set on French grammar might include questions on different parts of speech, such as verbs, nouns, and conjunctions, where learning one concept might not directly help with another. Given the large number of practice sets on StudyGo, manual inspection was not feasible, so an alternative approach was developed to identify cohesive practice sets.

We filtered the practice sets through learning curve analysis. Learning curves plot performance on a task relative to the number of opportunities to practice, showing how performance evolves per trial (Martin et al., 2011). These curves were plotted for each practice set, displaying error rates per question, using data from up to two years before the experiment. An example of such a learning curve is shown in Figure 14. For question sets on related concepts or skills, we expected the graphs to show similar error rates across questions, with performance improving as users progress through the set. Learning curves ideally follow a power law relationship, where learning improves as practice improves, with rapid improvement in the beginning that slows down as practice increases (Martin et al., 2011). The formula for a power law is:

$$P = BN^{-\alpha}$$

In this equation, P stands for the measure of performance, and N refers to the number of trials. The constant B represents the performance at the

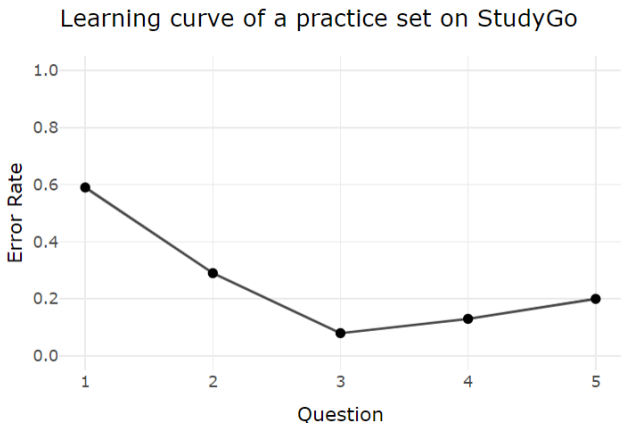

Figure 14. *A learning curve from a practice set on StudyGo, showing the error rate for each question.*

first attempt. The exponent α determines the steepness of the curve, with a steeper negative slope corresponding to faster improvement in performance.

To identify practice sets that focus on a single topic, we filtered learning curves based on how well they fit a power law model. A poor fit suggests that the questions are too varied in difficulty or topic, as indicated by inconsistent error rates, while a good fit indicates a more cohesive set of questions. We also considered the slope of the fitted curve to confirm that performance was improving, as indicated by a negative slope.

For the measure of fit, we chose not to use the commonly employed coefficient of determination R^2, as it has been found less reliable for non-linear models like the power law (Spiess & Neumeyer, 2010). Instead, we used the Residual Standard Error (RSE), which measures the average deviation of observed data points from the regression line and is appropriate for non-linear data (Jarantow, Pisors, & Chiu, 2023). A cutoff RSE value of 0.15 was set and practice sets with an RSE greater than this threshold were excluded. The cutoff point was determined by examining the distribution of RSE values and confirmed through manual inspection of the curves.

Another criterion for inclusion was that the slope must be negative, meaning that error rates decreased over time, providing evidence of learning. Figure 15 shows examples of learning curves with their fitted regression lines, featuring both included and excluded sets.

A potential issue with this filtering approach is that sets with limited data can have poor fits to the power law due to the influence of individual data points (Martin et al., 2011). To address this, for sets with fewer than 15 data points, we added the data from the normal order group during the experiment to improve the reliability of the learning curves. This also allowed

us to include newer topics that had little or no data before the experiment. However, practice sets that still did not fit well to a power law after this adjustment were excluded from further analysis. Topics were excluded regardless of their learning curve if the first or last question required users to mentally answer and then report whether their answer was correct. For these self-reported answers, we could not confirm whether an answer was actually correct, even when marked as such. Therefore, these questions were deemed unreliable for measuring learning. In total, 718 practice sets were selected for inclusion in the rest of this analysis, while 2,963 practice sets played during the experiment were excluded.

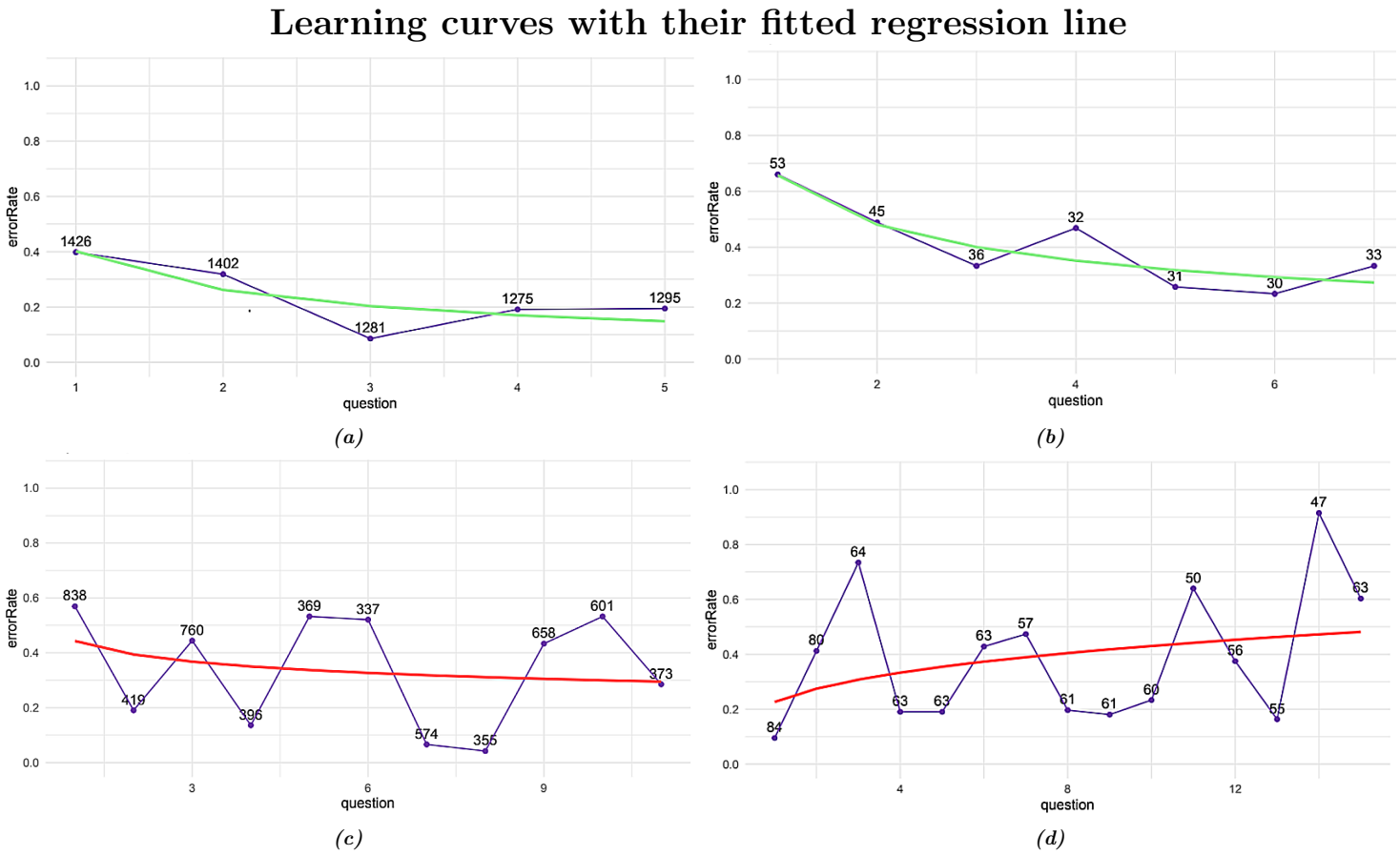

Figure 15. *Four examples of learning curves from practice sets on StudyGo. Plots a) and b) are included based on their fit to a power law model and its slope, while plots c) and d) are excluded, as indicated by the line color. The numbers next to each point represent the number of answers given to each respective question.*

Additional filters were applied to the data from the selected 718 sets,

focusing on single answer sessions (i.e., individual attempts at a practice set). We retained answers only from sessions where a set was completed until the end, to mitigate potential bias from only the more motivated students finishing the sets. We also included only answers from sessions completed in one sitting, excluding sessions where students stopped and returned later, in order to eliminate the influence of external learning sources between attempts. Any answers given in sessions where questions were skipped were also excluded, since skipping reduces the opportunity for learning from the set. Skipping behavior also presented a potential bias, as the last question in a set cannot be skipped, unlike earlier questions. This could have increased the likelihood of students guessing or giving deliberately incorrect answers to the last question if they wanted to skip but were not able to. To reduce the impact of this issue, only sessions where no questions were skipped were retained, as these sessions are less likely to involve students who wanted to skip the final question. Additionally, we excluded any answers to the first and last questions that were submitted within one second, as these were likely guesses rather than genuine responses (Chen & Guthrie, 2019).

4.2.2.2 Randomization check

Although users were randomly assigned to the experimental conditions, it was still possible that one of the groups included users with better overall performance. To ensure this was not the case, we performed an analysis comparing the performance of each group on the middle questions of the sets (i.e., all questions excluding the first and last). These middle questions remained in the same position across both conditions and were not part of the main analysis. We tested whether the difference in average performance per middle question between the conditions was significantly different from zero. Since the data was not normally distributed, a Wilcoxon signed-rank test was applied, which is appropriate for comparing paired samples with non-normal distributions. The results indicated that the median of the differences was not significantly different from zero ($p = 0.16$), confirming that the two groups of participants were comparable in terms of ability and that randomization was effective.

4.2.2.3 Mixed model analysis

The refined dataset can be divided into four groups based on experimental conditions and the position of the questions within the practice sets:

1. **First Normal**: Answers to questions in the first position by students in the normal order condition

2. **Last Normal**: Answers to questions in the last position by students in the normal order condition
3. **First Swapped**: Answers to questions in the first position by students in the swapped order condition
4. **Last Swapped**: Answers to questions in the last position by students in the swapped order condition

Note that the questions in groups 1 and 4 were the same, as were those in groups 2 and 3, they were only in a different position due to the experimental conditions.

This analysis aimed to determine whether the probability of answering a question correctly was influenced by its position within a practice set. As students may have completed multiple practice sets during the experiment, a generalized linear mixed model (GLMM) was used to account for individual differences in performance. The model was fitted using the *glmer*() function from the lme4 package (v1.1-35.5; Bates et al. 2015).

Including user ID as a random effect in the model allowed us to control for this variation in individual performance, which was justified by likelihood ratio tests showing a better model fit, as indicated by lower AIC and BIC values. In addition to user ID, other random effects were included based on the model selection process. These random effects included question type, stream ID, topic ID, and question ID. Question type refers to the different formats used, such as multiple choice or open questions, which can affect the difficulty of a question. Stream ID represents the educational track in which the question was placed, covering the Dutch system's tracks: VMBO, HAVO, and VWO. Topic ID refers to the specific practice set a question was part of, while question ID identifies the particular question answered. Question ID is nested within topic ID to reflect how questions were organized within practice sets.

The outcome variable, *incorrect*, is binary (0 for correct and 1 for incorrect) and therefore modeled with logistic regression using a logit link function (Brown, 2021). The predictors in this model included *position*, indicating whether the question appeared in the first or last position, and *variation*, specifying whether the question was answered by a user in the normal or swapped order condition. It was necessary to include variation as a fixed effect to differentiate between the two groups of questions, those normally positioned first and those normally positioned last. This ensured that any potential differences in the effect between these groups could be captured.

The formula of the final logistic mixed-effects model is as follows:

$$incorrect \sim position * variation + (1|questiontype)$$
$$+(1|topicID/questionID) + (1|streamID) + (1|userID)$$

The interaction between the fixed effects allowed us to examine whether the probability of an incorrect answer differed across the four groups introduced earlier. Therefore, we could determine if the effect of question position varied between the two sets of questions.

The final model was evaluated by checking key assumptions, including the normality of the random effect distributions, homogeneity of variances, and the presence of overdispersion (Bolker et al., 2009). Q-Q plots were used to inspect the random effect distributions, which were found to be appropriately normal. No signs of overdispersion were observed, and the transformed variances were homogeneous across groups, as confirmed by tests from the DHARMa R package by Hartig (2022).

4.3 Results

The purpose of this analysis was to determine whether the position of a question in a practice set affected the probability of answering incorrectly. A generalized linear mixed model (GLMM) was used to examine the effects of question position and experimental condition on performance, comparing the error rate between the first and last positions for two question sets. The model was fitted using maximum likelihood estimation with the Laplace approximation and controlled for random effects, including user ID, question type, stream ID, topic ID, and question ID. The final model had an AIC of 59,287 and a BIC of 59,367, indicating a good fit compared to alternative tested models. The model was fit based on 53,035 answers from 6,337 participants to 718 practice sets. Participants were secondary school students from the Netherlands, aged 12 to 18.

The fixed effects results are presented in Table 10. The default position is the first, and the default variation is the normal order, so the intercept represents the estimated log odds of answering incorrectly for questions in the first position in the normal order. The main effects of both position and variation were significant, as was their interaction. This indicates that the effect of question position on performance differs per variation.

The random effect estimates are shown in Table 11. The random intercept for user ID had a standard deviation of 0.54, indicating considerable variability in baseline performance across participants. The variability in performance for specific questions and topics was similar, with standard deviations of 0.54 and 0.62, respectively. Less variability was observed for

Term	Estimate	Std. Error	z value	p-value
(Intercept)	-0.64	0.17	-3.70	0.0002
PositionLast	-0.68	0.07	-9.37	< 0.001
VariationSwapped	-0.52	0.07	-6.95	< 0.001
PositionLast:variationSwapped	1.09	0.14	7.81	< 0.001

Table 10. *Fixed effects estimates on the log-odds scale for predicting error rates by position, variation, and their interaction.*

streams (SD = 0.18) and question types (SD = 0.21), suggesting more consistent performance across groups within these variables.

Groups	Variance	SD
User ID	0.29	0.54
Question ID:topic ID	0.29	0.54
Topic ID	0.38	0.62
Stream ID	0.03	0.18
Question type	0.04	0.21

Table 11. *Random Effects of the GLMM on the log-odds scale.*

To further investigate the combined effects of question position and variation, the estimated marginal means (EMMs) were calculated using the emmeans package (v1.10.4; Lenth 2024), and are shown in Table 12. These EMMs are shown on the response scale, and the EMMs on the log-odds scale can be found in Table 38 in Appendix B. These estimates show how the probability of answering incorrectly is influenced by both question position and variation, and how these factors interact. For example, in the normal order, the probability of answering incorrectly is higher for first-position questions than for last-position questions. However, this pattern is reversed in the swapped order group, where the probability of answering incorrectly is higher for questions in the last position.

Position	Variation	Probability	SE	Lower 95% CI	Upper 95% CI
First	Normal order	0.35	0.04	0.27	0.43
Last	Normal order	0.21	0.03	0.16	0.27
First	Swapped order	0.24	0.03	0.18	0.31
Last	Swapped order	0.32	0.04	0.25	0.40

Table 12. *Estimated marginal means (EMMs) on the response scale, showing the estimated probability of an incorrect response for each combination of position and variation.*

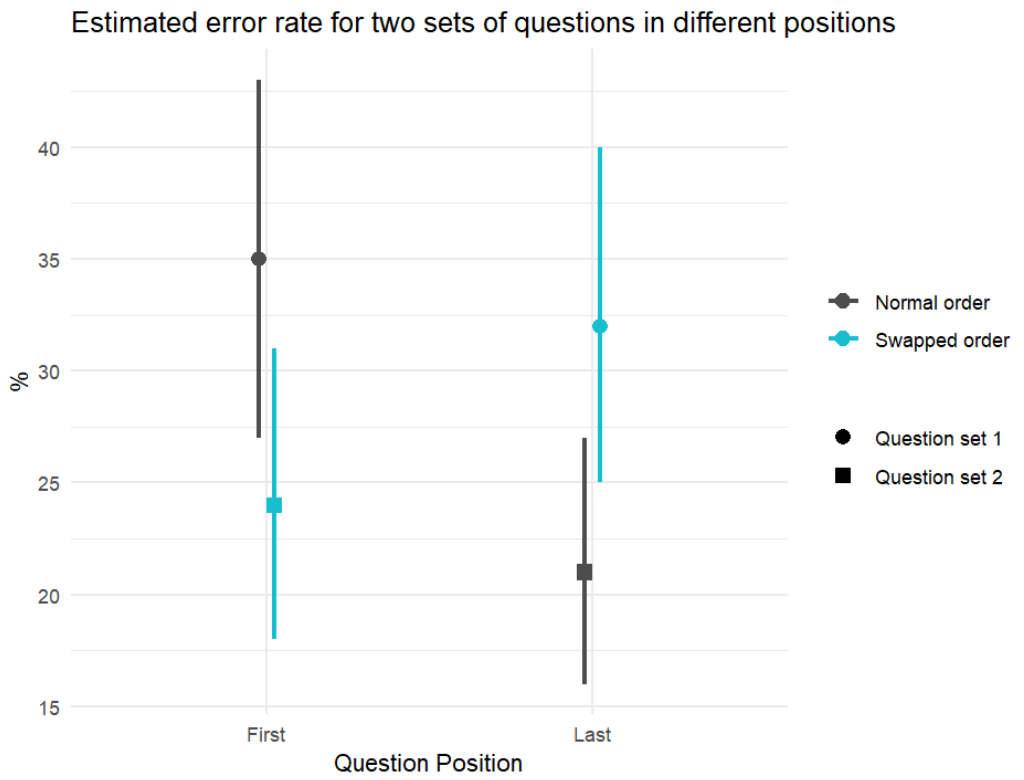

Figure 16. *Estimated marginal means on the response scale for different question positions and variations. The color indicates the variation, and points with the same shape represent the same question set and are the key comparisons of interest.*

The key contrasts of interest compare the same sets of questions in different positions: first position in the normal order vs. last position in the swapped order, and last position in the normal order vs. first position in the swapped order. Figure 16 illustrates the EMMs for the probability of answering incorrectly across the different groups, including 95% confidence intervals. While the CIs provide a measure of uncertainty around the individual group estimates, they should not be used to determine significance between groups. This is because the CIs reflect the uncertainty around each group mean independently, rather than the difference between groups. The plot shows that for questions in the first position in the normal order, the error rate is 35%, which is higher than when these questions are placed in the last position in the swapped order, where the error rate is 32%. This pattern also holds for the other set of questions (represented by square-shaped points), where the error rate is 24% in the first position in the swapped order, compared to 21% in the last position in the normal order.

Contrast	Estimate	SE	z-ratio	p-value
First Normal - Last Normal	0.68	0.07	9.37	< 0.001
First Normal - First Swapped	0.52	0.07	6.95	< 0.001
First Normal - Last Swapped	0.11	0.03	3.31	0.005
Last Normal - First Swapped	-0.16	0.03	-4.72	< 0.001
Last Normal - Last Swapped	-0.57	0.08	-7.61	< 0.001
First Swapped - Last Swapped	-0.41	0.07	-5.59	< 0.001

Table 13. *Contrasts of EMMs on the log-odds ratio scale.*

To test whether the differences in performance between question positions and experimental conditions were significant, contrasts of the EMMs were calculated, as shown in Table 13. The contrasts revealed that all comparisons of position and variation were significant. In particular, the comparisons of interest, 'First Normal vs. Last Swapped' and 'Last Normal vs. First Swapped', were both significant ($p = 0.005$ and $p < 0.001$, respectively), confirming that participants were more likely to answer last-position questions correctly.

While the contrasts revealed significant effects on performance, it is also important to assess the size of these effects. To do this, we computed odds ratios as a measure of effect size. The odds ratios for the contrasts between levels of the fixed effects are presented in Table 14. For the comparison between questions in first position in the normal order and last position in the swapped order, the odds of answering incorrectly were 11% lower for questions in the last position (reciprocal of OR = 1.12). Similarly, for the comparison between questions in the first position in the swapped order and the last position in the normal order, the odds of answering incorrectly were 15% lower for questions in the last position (OR = 0.85).

Contrast	Odds Ratio	Lower 95% CI	Upper 95% CI	p-value
First Normal / Last Normal	1.98	1.64	2.39	< 0.0001
First Normal / First Swapped	1.68	1.39	2.03	< 0.0001
First Normal / Last Swapped	1.12	1.03	1.22	0.005
Last Normal / First Swapped	0.85	0.78	0.93	< 0.0001
Last Normal / Last Swapped	0.57	0.47	0.69	< 0.0001
First Swapped / Last Swapped	0.67	0.55	0.80	< 0.0001

Table 14. *Odds ratios for contrasts of EMMs.*

4.4 Discussion

This study aimed to measure the effectiveness of StudyGo's practice questions in terms of knowledge gain, using an embedded experiment that manipulated question order. The GLMM analysis provided important insights, revealing a significant interaction between the predictors, which meant that the effect of question position on performance varied depending on the experimental condition. Additionally, the random effects played a substantial role in influencing performance. There was significant variability in performance across different users, which is expected due to differences in individual characteristics such as effort, prior knowledge, and other factors that influence learning outcomes. The random effects of topic ID and question ID also had large standard deviations, likely reflecting variability in the inherent difficulty of different topics and questions. In contrast, smaller variability was observed between different streams (school tracks), likely because the questions were tailored to their specific audiences, ensuring that they were appropriately challenging for each track. The variability in performance between different question types (e.g., multiple choice and open questions) was also relatively small but still significant, suggesting that these questions were well adjusted to the abilities of students, despite their format.

Further inspection of the fixed effects, using EMMs and contrasts, revealed that questions were 3 % less likely to be answered incorrectly when placed in the last position of a practice set compared to the first position. This pattern is illustrated in Figure 16, where error rates are lower in the last position within the same set of questions (represented by markers of the same shape). Also noticeable is the interaction effect, showing that in the normal order variation, the error rate was higher for first-position questions, whereas in the swapped order, the error rate was higher for last-position questions. This trend can be attributed to the learning curve filtering applied in the analysis (as described in section 4.2.2.1). The filtering retained only practice sets that showed a decreasing trend in error rates from the first to the last question. However, this trend likely reflects not only learning but also a progression in question difficulty, with easier questions appearing later in the set. Therefore, when the order of the questions is reversed, the difficulty increases as students progress through the set.

This effect is illustrated in Figure 17. Plot a) shows a typical learning curve, with error rates decreasing as students answer each question. If the error rates were influenced solely by question difficulty, reversing the question order would produce the trend shown in plot b) in blue. Alternatively, if error rates were decreasing only due to learning, we would expect no change in the trend when the question order is reversed, as shown in plot c). In reality, both

difficulty and learning likely influence the error rates, leading to a mixed effect when the order is swapped. In this case, easier questions appear earlier, but students make slightly more mistakes, while harder questions appear later, but are answered more accurately due to learning from previous questions, as shown in plot d). This combined effect mirrors the pattern observed in the EMMs from the experiment, although in a more exaggerated form, suggesting that the observed pattern is influenced by both question difficulty and the learning process.

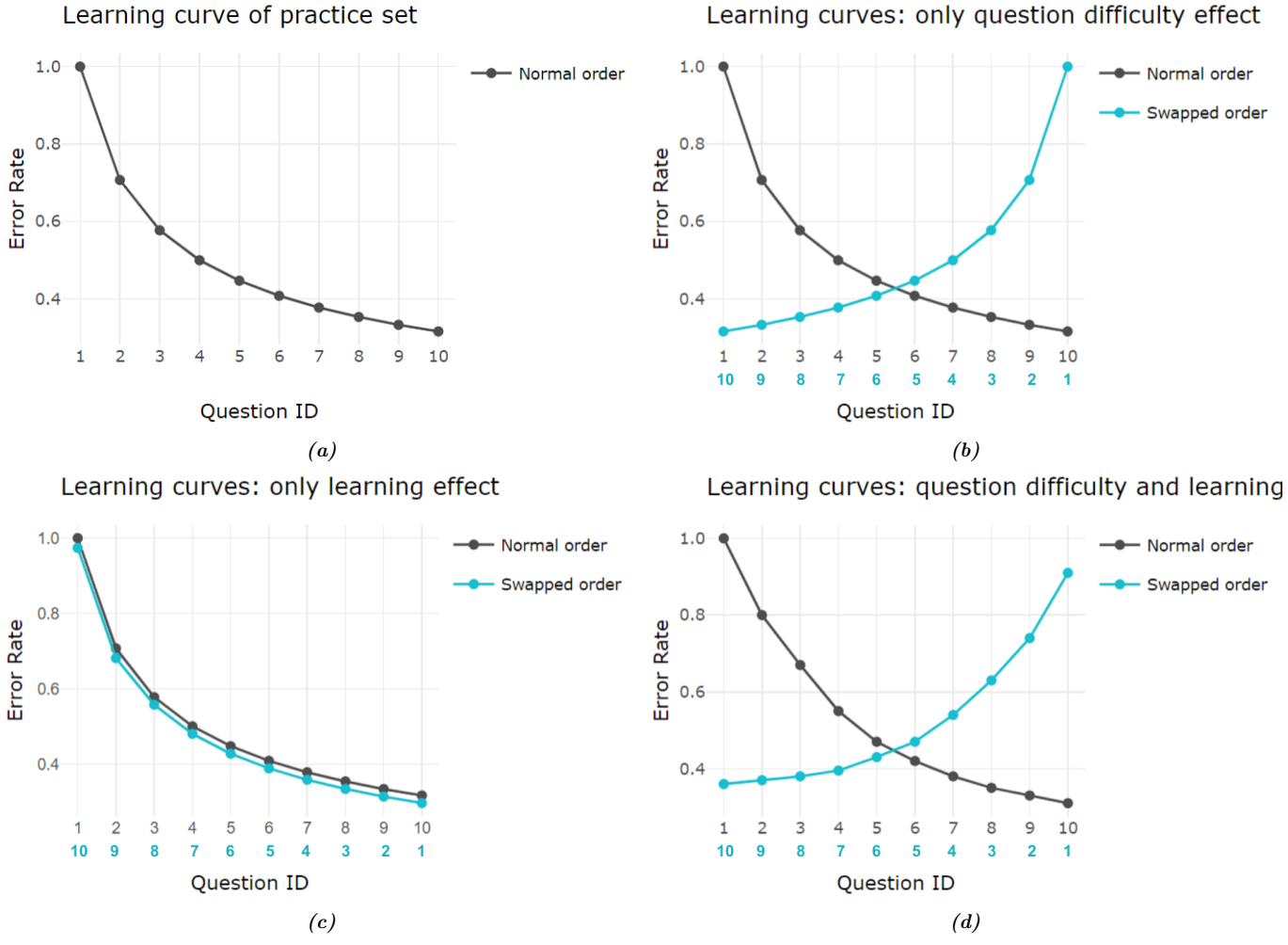

Figure 17. *Expected learning curves influenced by difficulty and learning for normal and swapped question orders.*

The small learning effect in this study was also reflected in the odds ratios. They showed a moderate difference in performance for questions within

the same set (OR = 1.12 and OR = 0.85). Other comparisons, such as the difference between the first and last positions in the normal order, had a much larger odds ratio (OR = 1.98), indicating that the odds of answering incorrectly were nearly twice as high for questions in the first position. Although the odds ratios for the contrasts of interest were smaller, they still represent significant and meaningful changes in performance between different positions within the same set of questions.

The research question addressed in this study was: "Can the effectiveness of an online learning platform be measured by manipulating the order of practice questions on the platform to observe changes in student performance?" The findings of this study suggest that this approach is indeed a viable method for assessing the effectiveness of a platform. By manipulating the order of questions on StudyGo, we were able to measure differences in student performance based on question position, which provided insights into learning patterns on the platform. Specifically, the observed changes in accuracy across different question positions allowed us to detect a learning effect and thereby assess the impact of completing practice question sets on StudyGo.

The results of the experiment showed that, for the 718 practice sets analyzed, accuracy was higher for the last question compared to the first. Specifically, the probability of answering incorrectly was 3% lower when a question was placed at the end of a set. This finding aligns with our hypothesis that the error rate for a question is lower when it appears at the end of a set compared to the beginning, due to the learning that occurs as users progress through the practice set. These results are also consistent with the theory on the Law of Exercise, which states that repetition strengthens learning (Thorndike, 2017). We also hypothesized that this effect would be small, due to the relatively short length of many practice sets on StudyGo, which provides limited opportunities to learn within a single session. Other factors may also contribute to the modest effect size. For example, some questions might be straightforward or familiar to students, leaving less room for improvement. However, the EMMs show that the mean error rates ranged from 21% to 35% percent, suggesting that there was sufficient room for improvement across questions.

Although this experiment was conducted on StudyGo, the approach has potential applications for other digital learning platforms that use question-based practice methods, such as quizzes. The results suggest that learning outcomes can be measured during single-session attempts, using only log data from users' interactions, without relying on direct feedback or traditional pre-post assessments. This offers a practical way to evaluate learning

passively, which is less time-consuming and better suited for large-scale real-world applications. The experiment also allowed users to interact naturally with the platform, without disrupting the user experience, making it an attractive option for platforms focused on self-directed learning. Additionally, the design is scalable and cost-effective, providing a continuous method for evaluating the effectiveness of learning tools. This experiment demonstrates how data-driven approaches can complement traditional evaluation methods. As digital education evolves and adapts to rapidly changing learner needs, methods like this could be valuable for providing real-time insights into the effectiveness of educational technology.

4.4.1 Limitations

Although the analysis revealed a significant decrease in the likelihood of answering incorrectly at the end of a set, these results are based on a relatively small subset of available practice sets on StudyGo. Only 718 of the 3,681 sets played during the experiment were included in the mixed model analysis. This limited sample is further constrained by the fact that many additional sets on the platform were not played at all during the experiment. One reason for this is the alignment of StudyGo topics with the Dutch school curriculum, meaning that practice sets are primarily used for exam preparation and their use varies throughout the school year. The findings can therefore not be generalized to all practice sets on the platform, and we cannot definitively conclude whether this feature is effective.

Moreover, StudyGo offers a variety of learning features beyond practice questions. The method used in this study cannot be directly applied to these other features, which would require distinct approaches to assess their effectiveness. Additionally, while this design could theoretically be applied to other platforms with similar question-based practice features, it may not be as straightforward to implement. The success of such experiments depends on the type of log data recorded and how closely a platform's features align with those in StudyGo. Nonetheless, this study can inspire future efficacy experiments adapted to other digital learning tools.

Another potential limitation is whether the observed improvement in accuracy truly reflects learning, or if it may be influenced by other factors such as familiarity with the question format, motivation to finish a set, or guessing patterns. While these alternative explanations are possible, learning remains a very plausible interpretation. The analysis included a large number of practice sets across various topics and formats, answered by a diverse group of students. This reduces the likelihood that the improvement is solely due to external factors. Although these influences cannot be ruled out, the observed

trend is consistent with a learning effect, where knowledge accumulates over the course of the set.

Several limitations are also associated with the specific methods of analysis used in this study. One key limitation is the filtering process used to identify cohesive sets focused on the same topic. Due to time constraints, the sets were not manually inspected for question similarity but were filtered based on their fit to a decreasing learning curve. This filtering method may have excluded sets that were suitable for the experiment, such as those where the difficulty increases at the same rate or more than the learning effect, resulting in a flat or rising learning curve. Many sets may have been excluded from the analysis for this reason, as ordering questions from easy to hard is a common approach in educational design, and leads to an increasing learning curve. The StudyGo content team acknowledged that they may order questions based on difficulty subconsciously.

Another limitation of this filtering method stems from the fact that the quality of a power law tends to improve with larger datasets, as the influence of individual data points decreases (Martin et al., 2011). This means that practice sets with fewer responses may have had worse fits and been excluded from the analysis, even though they might be effective at teaching the relevant concepts.

We also attempted to address the potential bias caused by the final question being unskippable by keeping only attempts in which no questions were skipped. However, this may not have fully resolved the issue, as some participants could still have rushed or guessed on the final question, leading to more incorrect answers. This may have slightly reduced the size of the observed learning effect.

Finally, this study did not conduct a formal power analysis prior to the experiment. While a simulation-based power estimation for mixed models, such as that described by Kumle et al. (2021), could have been useful, it requires reliable estimates of effect sizes and variability in random effects. Given the complexity of the mixed model in this study and the lack of empirical data to support such estimates, we chose to maximize the sample size by running the experiment for as long as possible to enhance statistical power.

4.4.2 Future work

One area for future research is investigating whether the observed improvement in performance reflects long-term retention rather than short-term gains. The limited effect size in this study is likely due to the minimal opportunities for learning, as many of the practice sets contain only a few questions. However, these practice sets are not intended as a stand-alone learning

method but are designed to complement other study materials, such as explanation videos, summaries, and practice exams available on the platform. It would be valuable to explore how learning outcomes from practice questions, or other platform features, persist over time. The cumulative effect of multiple features may also result in a greater overall impact. To assess the long-term effectiveness of the platform as a whole, traditional pre-post assessments, combined with a control group using a similar platform, could offer insights into the relative impact of StudyGo's features on long-term learning.

Another promising direction for future work would be to investigate whether certain practice sets contributed more significantly to the observed 3% improvement in performance, and thus determine if some of the sets are more effective than others. By analyzing the characteristics of these sets, such as question format, length, and the context provided, researchers could gain valuable insights into what drives their effectiveness. This information could then be used to optimize practice questions on StudyGo and other digital learning platforms.

A related question worth exploring is whether the order of questions in terms of difficulty affects the effectiveness of a practice set. In this study, only the first and last questions were swapped, and no difference in the learning effect was observed between the two sets of questions. However, future research could investigate the impact of reversing the entire order of questions, comparing sets ordered from easy to difficult versus difficult to easy. Previous research on this topic has been performed, such as the study by Anaya et al. (2022), which examined how the difficulty of earlier questions affected performance on later questions in tests on an online teaching platform. They found that arranging questions from easy to difficult resulted in the highest number of correct answers. These findings raise the question of whether the practice sets excluded from the StudyGo research, due to their increasing difficulty and corresponding learning curves, are just as effective, or maybe more so, than this study suggests. Future research should consider an alternative method for filtering cohesive practice sets that does not exclude those ordered from easy to difficult. Re-running the experiment with such sets could reveal whether the findings from this study hold or if different question sequences produce more significant learning effects. Such an experiment could contribute to the research on question order sequences and help inform best practices.

5 Conclusion

The primary aim of this research was to evaluate the effectiveness of digital learning platforms in enhancing students' knowledge through platform-embedded experiments. This study sought to develop cost-effective, continuous methods for measuring learning outcomes, addressing the limitations of traditional in-person or feedback-based approaches. Two learning platforms, Squla and StudyGo, were used as case studies, for which experiments were conducted to evaluate their key features.

The findings from Squla demonstrated that targeted practice on specific topics led to improved performance, particularly in math and language. For these subjects, we found that participants who practiced on-topic quizzes before retaking previously incorrect questions showed a 3% improvement in accuracy compared to those who practiced off-topic quizzes. However, no significant effects were observed for spelling and grammar or reading comprehension, suggesting that these subjects may require additional methods of instruction to effectively enhance understanding.

The StudyGo experiment provided evidence that students learn from practice sets on the platform, as error rates decreased when questions were placed at the end of a set compared to the beginning. This supports the idea that users gain knowledge as they progress through the questions. However, the observed learning effect was modest, likely due to the relatively short length of the practice sets, which limited opportunities for substantial learning within a single session.

These case studies allowed us to test two new embedded approaches for measuring platform effectiveness. The experiments demonstrated that targeted practice on Squla and StudyGo could enhance learning outcomes, although the extent of these benefits varied depending on the subject matter. Both studies revealed valuable insights into the effectiveness of quiz-based learning on digital learning platforms, confirming that the embedded experiments were suitable evaluation methods. However, further refinement is needed to improve the accuracy and reliability of these methods for future applications.

Several limitations were identified in both experiments. A common limitation was the generalizability of the findings. The studies focused on a subset of practice question sets, with Squla's experiment applied only to a few subjects and StudyGo's experiment constrained by extensive data filtering. Therefore, conclusions about the overall effectiveness of the platforms' features are limited. Additionally, no power analysis was conducted for either experiment, raising concerns about whether the sample sizes were adequate

to detect learning effects. This was especially a concern for the Squla experiment, where no significant effects were found for two subjects.

The adaptability of the experimental designs to other platforms is another important consideration. The approach used in StudyGo, which involved manipulating the order of questions, is likely easier to adapt to other platforms. In contrast, the Squla experiment required creating a new quiz to retry previously incorrect questions, which might require the development of new platform functionalities. This indicates that the success of similar embedded experiments depends on the type of log data available and how closely a platform's features align with those of Squla and StudyGo.

Directions for future research include expanding the experimental designs to cover longer learning periods and a broader range of subjects, which would allow for a more comprehensive evaluation of the practice features. Investigating which types of quizzes or content features are most effective could provide valuable insights for improving content design. Additionally, combining embedded experiments that assess different features could offer a more comprehensive overview of platform effectiveness. Another promising approach is to integrate platform-embedded experiments with occasional in-person methods, making use of the advantages of both to provide a more complete evaluation.

In conclusion, this research demonstrates the potential of platform-embedded experiments for evaluating the effectiveness of digital learning platforms. The experiments did not disrupt the user experience or require user feedback, as participants were able to use the platform as usual and did not need to be actively recruited. The results indicated that targeted practice on Squla and StudyGo can improve learning outcomes, though the effects were subject-dependent and modest. To better understand how digital learning tools can support education, future studies should refine these methods, address their limitations, and explore their applicability across different platforms.

Acknowledgements

I would like to sincerely thank Jasper Naberman, my supervisor at Futurewhiz, for his guidance, ideas, and assistance throughout this project. His motivation and expertise played an important role in making this work possible. I am also grateful to my project supervisor, Dr. Sergey Sosnovsky, for his valuable feedback and advice. His novel ideas helped us overcome several challenges along the way, and his critical view helped elevate the project. I would also like to thank Dr. Matthieu Brinkhuis, my second examiner, for his reviews of this work and for directing me to relevant and interesting related work. Finally, I would like to thank Jelte van den Akker for his support and encouragement, which were a great help during this process.

References

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review*, *90*, 102293.

Baker, A. T., & Cuevas, J. (2018). The importance of automaticity development in mathematics. *Georgia Educational Researcher*, *14*(2), 13–23.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, *24*(3), 127–135.

Broekman, F., Smeets, R., Bouwers, E., & Piotrowski, J. (2021). Exploring the summer slide in the Netherlands. *International Journal of Educational Research*, *107*. (Publisher: Elsevier)

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in r. *Advances in Methods and Practices in Psychological Science*, *4*(1).

Chauhan, S. (2017, February). A meta-analysis of the impact of technology on learning effectiveness of elementary students. *Computers & Education*, *105*, 14–30. doi: 10.1016/j.compedu.2016.11.005

Chen, Z., & Guthrie, M. (2019, March). *Measuring the Effectiveness of Learning Resources Via Student Interaction with Online Learning Modules.* arXiv. (arXiv:1903.08003 [physics]) doi: 10.48550/arXiv.1903.08003

Cheung, A. C., & Slavin, R. E. (2013, June). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. doi: 10.1016/j.edurev.2013.01.001

Chirikov, I., Semenova, T., Maloshonok, N., Bettinger, E., & Kizilcec, R. F. (2020, April). Online education platforms scale college STEM instruction with equivalent learning outcomes at lower cost. *Science Advances*, *6*(15), eaay5324.

De Witte, K., Haelermans, C., & Rogge, N. (2015). The effectiveness of a computer-assisted math learning program. *Journal of Computer Assisted Learning*, *31*(4), 314–329. doi: 10.1111/jcal.12090

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r.* SAGE Publications.

Hartig, F. (2022). Dharma: Residual diagnostics for hierarchical (multi-level

/ mixed) regression models [Computer software manual]. (R package version 0.4.6)

Higgins, S., Xiao, Z., & Katsipataki, M. (2012, November). *The Impact of Digital Technology on Learning: A Summary for the Education Endowment Foundation. Full Report* (Tech. Rep.). Education Endowment Foundation. (Publication Title: Education Endowment Foundation ERIC Number: ED612174)

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. doi: 10.18637/jss.v042.i08

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, *20*(1), 1–24.

Jansen, B. R. J., De Lange, E., & Van der Molen, M. J. (2013, May). Math practice and its influence on math skills and executive functions in adolescents with mild to borderline intellectual disability. *Research in Developmental Disabilities*, *34*(5), 1815–1824. doi: 10.1016/j.ridd.2013.02.022

Jarantow, S. W., Pisors, E. D., & Chiu, M. L. (2023). Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. *Current Protocols*, *3*(6), e801.

Jiang, X., & Pajak, B. (2022). Reading and Listening Outcomes of Learners in the Duolingo English Course for Spanish Speakers.

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010, December). A Quasi-Experimental Evaluation of An On-Line Formative Assessment and Tutoring System. *Journal of Educational Computing Research*, *43*(4), 489–510. (Publisher: SAGE Publications Inc) doi: 10.2190/EC.43.4.d

Kralj, L. (2022, February). Recovery and resilience plans for education: Agile collection of information.

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior research methods*, *53*(6), 2528–2543.

Lenth, R. V. (2024). emmeans: Estimated marginal means, aka least-squares means [Computer software manual]. (R package version 1.10.4)

Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, *53*(2), 209–233. doi: 10.1111/flan.12454

Martin, B., Mitrovic, A., Koedinger, K. R., & Mathan, S. (2011, August). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, *21*(3), 249–283.

doi: 10.1007/s11257-010-9084-2

Mashaw, B. (2012). A Model for Measuring Effectiveness of an Online Course. *Decision Sciences Journal of Innovative Education*, *10*(2), 189–221. doi: 10.1111/j.1540-4609.2011.00340.x

Papastergiou, M. (2009, January). Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. *Computers & Education*, *52*(1), 1–12. doi: 10.1016/j.compedu.2008.06.004

Pilli, O., & Aksu, M. (2013, March). The effects of computer-assisted instruction on the achievement, attitudes and retention of fourth grade mathematics students in North Cyprus. *Computers & Education*, *62*, 62–71. doi: 10.1016/j.compedu.2012.10.010

Portnoff, L., Gustafson, E., Rollinson, J., & Bicknell, K. (2021). Methods for language learning assessment at scale: Duolingo case study. *International Educational Data Mining Society*.

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Spiess, A.-N., & Neumeyer, N. (2010). An evaluation of r 2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a monte carlo approach. *BMC pharmacology*, *10*, 1–11.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Sung, Y.-T., Yang, J.-M., & Lee, H.-Y. (2017). The effects of mobile-computer-supported collaborative learning: Meta-analysis and critical synthesis. *Review of educational research*, *87*(4), 768–805.

Thorndike, E. (2017). *Animal intelligence: Experimental studies*. Routledge.

Zeng, J., Sun, D., Looi, C.-K., & Fan, A. C. W. (2024). Exploring the impact of gamification on students' academic performance: A comprehensive meta-analysis of studies from the year 2008 to 2023. *British Journal of Educational Technology*.

Appendix

A Matching summaries of balance

Math grades 1-3

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	11.94	11.11	0.09	1.08	0.02	0.05
practiceDuring	118.52	65.88	0.53	1.71	0.12	0.27
practiceBefore	204.81	203.78	0.00	1.06	0.01	0.06

Table 15. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.98	10.96	0.00	0.99	0.00	0.02
practiceDuring	75.13	75.04	0.00	1.00	0.00	0.02
practiceBefore	154.44	151.98	0.01	0.98	0.01	0.04

Table 16. *Summary of balance for matched data.*

Math grades 4-6

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.89	10.00	0.10	1.08	0.02	0.05
practiceDuring	119.40	68.47	0.52	1.46	0.12	0.28
practiceBefore	258.53	251.80	0.02	1.11	0.01	0.05

Table 17. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	9.71	9.82	-0.01	1.01	0.00	0.03
practiceDuring	83.29	83.27	0.00	1.00	0.00	0.02
practiceBefore	207.15	206.79	0.00	0.98	0.01	0.04

Table 18. *Summary of balance for matched data.*

Math grades 7-8

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.97	9.92	0.10	1.11	0.03	0.06
practiceDuring	105.33	51.06	0.58	1.73	0.13	0.38
practiceBefore	249.76	211.78	0.12	1.25	0.04	0.07

Table 19. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	8.56	8.67	-0.01	0.99	0.01	0.04
practiceDuring	64.32	64.06	0.00	0.99	0.00	0.04
practiceBefore	139.21	134.85	0.01	0.98	0.01	0.08

Table 20. *Summary of balance for matched data.*

Language grades 1-3

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	12.41	10.38	0.22	1.27	0.05	0.10
practiceDuring	93.27	35.55	0.70	3.36	0.15	0.41
practiceBefore	247.03	218.62	0.09	1.30	0.03	0.06

Table 21. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	11.05	11.22	-0.02	0.99	0.01	0.02
practiceDuring	48.40	47.95	0.01	0.99	0.00	0.05
practiceBefore	175.24	172.67	0.01	0.98	0.01	0.05

Table 22. *Summary of balance for matched data.*

Language grades 4-6

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	11.79	10.01	0.19	1.24	0.05	0.09
practiceDuring	92.98	30.51	0.74	3.39	0.15	0.48
practiceBefore	264.63	237.01	0.08	1.26	0.03	0.05

Table 23. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.77	10.90	-0.01	1.04	0.01	0.04
practiceDuring	50.96	50.88	0.00	0.99	0.00	0.04
practiceBefore	191.44	187.92	0.01	0.98	0.01	0.06

Table 24. *Summary of balance for matched data.*

Language grades 7-8

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	9.90	7.97	0.22	1.34	0.06	0.10
practiceDuring	84.46	21.70	0.74	4.59	0.20	0.61
practiceBefore	248.60	226.66	0.07	1.17	0.02	0.05

Table 25. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	7.46	7.54	-0.01	1.06	0.01	0.09
practiceDuring	41.77	41.42	0.00	0.98	0.00	0.09
practiceBefore	116.20	116.38	-0.00	0.88	0.02	0.09

Table 26. *summary of balance for matched data.*

Spelling and grammar grades 1-3

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.22	9.17	0.12	1.15	0.03	0.07
practiceDuring	84.09	25.62	0.78	3.66	0.16	0.55
practiceBefore	178.44	162.39	0.05	1.32	0.02	0.08

Table 27. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	8.91	9.09	-0.02	1.05	0.01	0.06
practiceDuring	44.84	44.35	0.01	0.97	0.00	0.12
practiceBefore	122.93	112.32	0.03	0.96	0.03	0.17

Table 28. *Summary of balance for matched data.*

Spelling and grammar grades 4-6

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.36	9.42	0.11	1.08	0.03	0.07
practiceDuring	107.10	50.60	0.59	1.47	0.13	0.45
practiceBefore	253.41	240.47	0.04	1.29	0.02	0.06

Table 29. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	9.57	9.82	-0.03	1.00	0.01	0.07
practiceDuring	64.26	64.16	0.00	1.00	0.00	0.05
practiceBefore	185.31	177.82	0.02	0.96	0.01	0.11

Table 30. *Summary of balance for matched data.*

Spelling and grammar grades 7-8

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.33	8.97	0.15	1.05	0.04	0.10
practiceDuring	101.20	35.24	0.73	2.21	0.19	0.52
practiceBefore	315.16	268.96	0.13	1.09	0.05	0.11

Table 31. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	6.33	6.50	-0.02	1.00	0.01	0.08
practiceDuring	48.28	47.90	0.00	0.99	0.00	0.08
practiceBefore	181.51	178.13	0.01	0.98	0.02	0.10

Table 32. *Summary of balance for matched data.*

Reading comprehension grades 4-6

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	10.61	9.50	0.11	1.29	0.03	0.06
practiceDuring	87.85	32.36	0.69	3.24	0.15	0.42
practiceBefore	258.85	268.50	-0.03	0.94	0.02	0.07

Table 33. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	8.56	8.64	-0.01	1.01	0.01	0.04
practiceDuring	46.64	46.17	0.01	0.99	0.00	0.04
practiceBefore	179.06	177.88	0.00	0.99	0.00	0.02

Table 34. *Summary of balance for matched data.*

Reading comprehension grades 7-8

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	9.34	8.46	0.10	1.11	0.03	0.05
practiceDuring	99.62	42.83	0.60	2.12	0.16	0.38
practiceBefore	218.08	208.59	0.03	1.22	0.02	0.09

Table 35. *Summary of balance for all data.*

Covariate	Means on-topic	Means off-topic	SMD	Var. Ratio	eCDF Mean	eCDF Max
practiceDays	6.63	6.77	-0.01	1.03	0.01	0.04
practiceDuring	54.13	54.08	0.00	1.00	0.00	0.03
practiceBefore	100.31	102.70	-0.01	1.01	0.01	0.05

Table 36. *Summary of balance for matched data.*

B EMMs on the log odds scale

Subject	Practice type	EMM	SE	Lower 95% CI	Lower 95% CI
Math	off-topic	0.78	0.03	0.73	0.84
Math	on-topic	0.94	0.03	0.88	0.99
Language	off-topic	0.85	0.05	0.76	0.94
Language	on-topic	0.98	0.05	0.89	1.07
Spelling and grammar	off-topic	0.74	0.07	0.61	0.87
Spelling and grammar	on-topic	0.91	0.07	0.78	1.05
Reading comprehension	off-topic	0.53	0.06	0.41	0.64
Reading comprehension	on-topic	0.44	0.06	0.32	0.55

Table 37. *Estimated Marginal Means (EMMs) on the log-odds scale from the Squla experiment model with formula:* $correct \sim practicedCategory * subject + (1|userID)$.

First	Variation	EMM	SE	Lower 95% CI	Upper 95% CI
First	Control	-0.64	0.17	-0.98	-0.30
Last	Control	-1.32	0.17	-1.66	-0.99
First	Experimental	-1.16	0.17	-1.50	-0.82
Last	Experimental	-0.75	0.17	-1.09	-0.41

Table 38. *Estimated Marginal Means (EMMs) on the log-odds scale from the StudyGo experiment model with formula:* $incorrect \sim position * variation + (1|questiontype) + (1|topicID/questionID) + (1|streamID) + (1|userID)$.

C Ethics

The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences classified this research as low-risk with no fuller ethics review or privacy assessment required.