MASTER THESIS

Digital, adaptive software: the new big thing?

A multilevel study using longitudinal data to investigate the effects of digital adaptive software on mathematics performance in sixth grade

G. R. Bruijn
S2350408

Faculty of Behavioural and Management Studies (BMS)
Educational Science and Technology

EXAMINATION COMMITTEE
Dr. J. W. Luyten
Dr. M. R. M. Meelissen

October, 2020

8043 words

Key words: Snappet, digital education, adaptive instruction, differentiated instruction, formative testing, mathematics.

Table of contents

Acknowledgement

Writing a thesis is always an important project. After all, it is the final step before graduating and entering the work field. When I started this project in February 2020, I was looking forward to doing research while also doing an internship at the Inspectorate of Education. Because of the pandemic that followed, this plan could not continue. As everyone was working from home, it was not possible to do the internship as elaborate as I was planning to. Instead, I focused solely on research and writing my thesis. Although this was not what I hoped for my final year as a student to be, I am proud that I am still graduating without too much of a study-delay. This would not have been possible without the guidance of my supervisor dr. Luyten from the University of Twente and Maarten Balvers from the Inspectorate of Education. I want to thank them for supporting me and helping me throughout the process, especially when things did not go to plan.

Abstract

This study aimed to measure the effects the software program Snappet has on the mathematics performance of primary school pupils. The software is known for its adaptive approach to learning, facilitating differentiated instruction and a personalised learning path. First, theoretical aspects of Snappet are analysed to show in what way the software is expected to increase performance. To measure the effects of the software, standardised test data from 2016 to 2019 was analysed through multilevel regression analysis. Two groups of 6-th grade pupils were compared. Results showed that the usage of Snappet was not a significant predictor for mathematics performance. Limitations of the study are discussed as well as practical and scientific implications of the study. Recommendations for further research include providing more detailed information about the extent to which the software is implemented and accounting for the differentiated effects previous research found between high- and lower-performing students. The final conclusion is that this software can benefit education in multiple ways, but the effects on the actual mathematics performance seem limited.

Introduction

The educational system needs to be accessible with pupils of all sorts and sizes. Over time, pupils get more and more sorted into different types of education, with different levels and content. In Dutch primary schools, this is not yet the case. Apart from age, there has hardly been any form of selection or sorting of pupils into groups. This results in classes with pupils that have different needs and thus need different types of guidance. One of the main tasks of education is to give every pupil the guidance it needs to reach its highest potential. To do so, differences among pupils are traced so that teachers can act accordingly. This is called adaptive teaching and testing and is currently a popular approach to offer a solution for overcoming the big differences among pupils that exist in primary schools (Ikwulmelu, Oyibe, & Oketa, 2015; Ismajli & Imami-Morina, 2018; Yenmez, & Özpinar, 2017). To help teachers implement adaptive teaching and testing, software programs were developed. Dutch primary schools make extensive use of these methods, the main three being Snappet (https://nl.snappet.org/), Gynzy (https://www.gynzy.com/nl/) and Exova MATH (https://exova.nl/math/). The methods consist of a software program executed on a tablet or laptop. Pupils follow their individual learning path, receive immediate feedback while teachers receive real-time information on the progress of every individual. Although suitable for multiple subjects, the methods are mostly used for mathematics.

These methods are still quite new and research on the effectiveness is therefore limited. Existing research focused on measuring the effects of these software programs and the first results appeared to be small, but positive (Drijvers, 2018; Faber, Luyten & Visscher, 2017; Molenaar & Knoop van Campen, 2017). One common limitation of these studies is that they took place over a year or shorter. There is no longitudinal research available yet. This study will focus on data gathered over three years, to see whether these first effects last over time. Effects found so far deserve a proper follow-up. These effects might have been moderated by the novelty effect, meaning new stimuli receive more cognitive attention, the treatment effect, meaning results in an experiment might increase because of the increased attention given to a subject during the treatment, and may decrease over time because of habituation, which is the effect of decreased reaction to a stimulus due to the amount of exposure to this stimulus (Siddigue, Dhakan, Rano & Merrick, 2016). The goal of this research is to indicate the effect the programs have on the results of the standardized tests in the long run. The main research question will, therefore, be: *What is the effect of the use of digital adaptive software for mathematics achievement in primary education on the results of the standardised test at the end of primary school (after being used for 3 years)?*

The study will focus on Dutch primary schools that used Snappet for their mathematics classes. This because Snappet is the most popular of the three aforementioned software programs. This paper starts with a brief explanation of the main theoretical concepts behind the software programs to give insight into why these are seen as promising. After that, the methods of this study will be discussed. To evaluate the effects of Snappet, the results of the standardized test (Toetsbeleid PO, 2014) pupils make when leaving Dutch primary

schools[1], will be used. Two samples will be compared, one containing schools using the software program and schools that do not. After the methodology is discussed, the results will be presented, followed by a discussion and concluding remarks. Directions for further research will be given.

[1] In this thesis, all grades referred to are based on the American school system where primary school consist of grade 1 (ages 4-5) to 6 (ages 11-12). This study refers to Dutch schools, but all grades are converted to the American school system to make this thesis comprehensive to international readers as well.

Research context

Research findings

This chapter will elaborate on the theoretical concepts underlying this study. First, an overview of existing research will be given. Based on these studies, a few important theoretical concepts will be explained in further detail. Finally, this chapter will explain what the software used in this study, called Snappet, entails and why it is supposed to enhance educational performance. This will lead to the hypotheses that are tested in this study.

The idea of implementing digital adaptive software is not new. Ysseldyke et al. (2003) found significant positive effects on students' performance on mathematics after using comparable software for a year. Ysseldyke & Bolt repeated the study in 2007 and found a significant difference in mathematics performance between groups that used the software for either a full year, half a year, or not at all. This research concluded that the extent to which the software is implemented affects students' performance as well since the performance increased for the group that used the software for a longer period. Although the software analysed in this study uses more advanced technology than the one studied in 2003 and 2007, the results are important to keep in mind because there is often a focus on finding quick, short-term results, while this study showed that results increase over time.

Faber, Luyten & Visscher (2017) performed an experimental study with 3rd-grade students to measure the effects of Snappet on motivation and performance. They found a significant positive effect on mathematics performance but did find that, overall, high performing students benefit more from Snappet than medium or low performing students. Next to that, their results indicated that the more a student used Snappet, the more their performance increased. They noted that this might be moderated by motivational factors or general mathematics ability. They conclude that Snappet seems to benefit mathematics performance but that moderating factors might play a role as well (Faber, Luyten & Visscher, 2017).

Molenaar & Knoop van Campen (2017) compared an experimental group that used Snappet to a control group that used a traditional method, both in grades 2 and 4. The experimental group used Snappet for mathematics, this resulted in increased performance in grade 4, again especially for high-performing students. There were no significant effects found for the respondents in grade 2. This adds to the conclusion of Faber, Luyten & Visscher (2017), the effects of Snappet can be positive but it seems to differ per ability group and maybe also per grade.

Drijvers (2018) wrote a general overview of the state of the art research concerning digital adaptive approaches to teaching and testing. This contained studies that investigated digital technologies in mathematics in general, so conclusions do not tell something about the specific adaptive, differentiated software programs used in this study per se. They do, however, paint an image that shows that results are overall positive and significant but since most effect sizes are small, their impact remains small.

Most digital adaptive software programs are based on the same core concepts. They create opportunities for differentiated instruction and an adaptive approach to suit the needs of each student. Next to that, they make use of frequent testing, instead of focusing on the final test with a more summative nature. This type of formative testing results in frequent and specific feedback, which then again leads to specific training with an adaptive and

differentiated approach for each student. Because these concepts form the core of the software Snappet that is analysed in this study, each concept will now be elaborated on.

Differentiated instruction

Differentiated instruction is an important part of the growing focus on the individual student in education and can be defined as "an instructional method complying with the constructivist approach that takes individual differences into consideration" (Yenmez & Özpinar, 2017). Taking these differences into account, each student can reach their highest potential. Differentiated instruction is crucial to meet the needs of all students and to early identify what aspects need more attention (Ismajli & Imami-Morina, 2018). Current research provides little knowledge about the way differentiated instruction affects student achievement. Faber, Glas & Visscher (2018) concluded that the amount and quality of differentiated instruction are very hard to measure reliably because of the differences between teachers and even between lessons of the same teacher. They attempted to define these effects in an observational study. The role of the observed differences was hard to define, making it unclear whether the effect was caused by differentiated instruction or other moderating factors (Faber, Glas &Visscher, 2018).

Differentiated instruction leads to properly challenging the students, thus enhancing their motivation. Deci and Ryan (2000) state that intrinsic motivation is enhanced when individuals are challenged at their intellectual level because their needs for feeling competent and autonomous are fulfilled. Bolkan (2015) agrees and states that when individuals are intellectually challenged, they feel more engaged in learning and appreciate the learning content more, enhancing their motivation.

Differentiated instruction cannot exist without data about the performance of the students. This data is used by the teacher to adjust their instruction. Part of the instruction is the feedback they give their students. Snappet is built around this principle and helps teachers define what feedback and instruction their students need. This way, they can guide their students through the learning process.

Adaptive approach

When differentiated instruction is part of an educational setting, an adaptive approach to learning is implemented. Snappet is famous for this adaptive approach to learning which can be defined as "a method of online instruction that involves providing personalized learning experiences resulting from a data-driven approach to curriculum design" (Shelle, Earnesty, Pilkenton & Powell, 2018). In other words, previous performance is analysed and used as input to personalise current and future learning content. When implemented successfully, adaptive learning avoids 'teaching to the middle', meaning that the instruction focuses on neither advanced nor lagging students alone (Educause, 2017). Adaptive learning offers possibilities for every student to maximise their learning gain because they are being challenged at their own level. Being challenged at your own level is likely to create a positive learning experience, which enhances one's confidence (Weltman, Timchenko, Sofios, Ayres & Marcus, 2019). Frustration caused by being tested at a level that is too high and boredom caused by being tested at a level too low are avoided.

The role of feedback

The adaptive approach and the differentiated instruction that comes with it are enhanced by the way Snappet collects data about the performance of students. This data enriches the insight teachers have of their students' development, which creates new possibilities for including feedback in their instruction. Snappet is built upon the results of research performed by Hattie & Timperly (2007). Their work emphasized the importance of proper feedback on the students' performance. Only when given in the right educational context, relating to concrete learning goals, feedback is positively affecting the students' progress. Next to that, Hattie & Timperly (2007) concluded that feedback directed towards students should be personalised, also taking into account environmental aspects that may influence a student's behaviour. The importance of these environmental aspects proves the importance of the teacher, since only they know exactly what is going on in the classroom. Snappet offers detailed insight for both the student and the teacher into the progress of the student. This makes it easier for the teacher to direct his feedback and instruction. Hattie & Timperly (2007) distinguish feedback on multiple levels, stating that feedback leading to self-regulated learning should be aimed for.

Because of the insight Snappet offers, this goal becomes more realistic. The teacher can use the feedback of Snappet to talk to the student, analysing their progress together, guiding them towards self-regulated learning, which is known as a trait possessed by effective students (Hattie & Timperly, 2007).

Mastery-based learning

Snappet accounts for every individual learning path and offers extra practice for those who need it and more challenging exercises for those who can handle a higher level. Every student needs to master the current learning goal before moving on to the next. This approach is based on the work of Bloom (1984), who developed the concept of mastery-based learning. Mastery-based learning entails that students first need to reach a certain level of knowledge before moving on to the next topic or level. The idea behind this is that most students are capable of reaching the highest level when given the opportunity to practice as often and as long as they need. When a student fails, this is more likely a problem of instruction than lacking capabilities of the student. This concept was used in an experimental study where three conditions were compared: a personal, one-to-one tutoring condition; conventional, classroom-broad instruction condition and the mastery-based learning condition that had the same teacher-student ratio as the conventional condition. Results showed that one-to-one tutoring is the most effective, creating the highest performance. Second highest performance was achieved by the mastery-based learning condition and the third was the conventional condition (Bloom, 1984). This shows that conventional education can be improved by focussing on mastery-based learning.

Snappet uses this principle, giving students as many exercises as they need before reaching the learning goal. Only then can they move on to the next learning goal. One-to-one tutoring would be even more effective, but this is not realistic when looking at the current ratio of students to teachers. Therefore, mastery-based learning is a second-best option. Snappet implements this, expecting to improve learning outcomes.

Formative assessment

Snappet is a software program that uses continuous testing to monitor the progress of each student. This way, summative tests are hardly necessary because students are being tested daily. This type of continuous testing where the output serves as input for further practice is called formative testing. A lot of definitions of formative assessment exist, and there has been an elaborate discussion about it among researchers. For this paper, the definition of Black and William (2009) is used, for they redesigned the definition in a way that most researchers nowadays refer to. They state that: "Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited." (Black & William, 2009, p.7). In other words, formative assessment is the process in which the teacher collects data on the performance of students to improve future learning and instruction (Hopster-den Otter, Wools, Eggen & Veldkamp, 2019). To do so, testing usually becomes more frequent, providing the teacher with process data next to outcomes based on tests that have a more summative nature. By providing teachers with detailed insight into the performance and development of their students, Snappet facilitates formative assessment.

Klute, Apthorp, Harlache & Reale (2017) conducted a review study on the effectiveness of formative assessment. They concluded that formative assessment appeared to be positively related to student academic achievement. However, this effect differed over subjects and approaches to the assessment. Bennet (2011) also stated that research implies that formative assessment has a positive effect on learning. Problematic is the variety in the types and sizes of the different implementations of formative assessment. Bennet (2018) states that this harms the validity and reliability of the findings. However, the general goal of formative assessment is to provide the teacher with evidence to improve their teaching. The general opinion is therefore that, when done correctly, formative assessment will lead to more effective teaching (Hopster-den Otter, Wools, Eggen & Veldkamp, 2019; Van der Kleij, Vermeulen, Schildkamp & Eggen, 2014).

In short, formative assessment can offer advantages to teaching and learning and it helps to guide the learning process. That makes it a good approach for both adaptive learning and differentiated instruction. But since the debated effectiveness of the approach, one should be careful implementing it. When formative assessment eliminates summative tests, it could harm the reliability and validity of students' performance, because there is a chance that too much variation in tests exists among schools and teachers. Summative testing should never be eliminated to account for this risk.

Deliberate practice

The core principles of the theory of deliberate practice support the idea of formative assessment and an adaptive approach and are hence also implemented in Snappet. The theory of deliberate practice is focused on the principle that to obtain a mastery level of something, conscious and precise training is needed, as well as many repetitions and receiving immediate feedback (Ericsson, Th. Krampe, & Tesch-Romer, 1993). It emphasizes the importance of individualised training, specifically and deliberately training only those aspects that are not

fully mastered yet. This idea is implemented into Snappet. Snappet offers many exercises, giving students the possibility to practice as long as they need to obtain the desired level. Students receive feedback right after each exercise to make sure the practice is adjusted to their needs.

Introduction into Snappet

The previous paragraphs showed that, when analysing the core principles of digital adaptive software programs, it looks very promising. Snappet implemented these principles as well. This program works in the following way.

The teacher decides what today's learning goal is and what topics should be addressed. Exercises that correspond to that learning goal and topic are then put into the devices (tablet or laptop) of the pupils. There are regular exercises, extra exercises for pupils who need more practise and exercises for pupils who want an extra challenge. Usually, the teacher will give a generic instruction after which the pupils start working individually. The pupils practise on their tablet and receive immediate on-screen feedback. There are visual cues that tell pupils whether a question was answered correctly. When a question is correct, the pupil moves on to the next one. When it is incorrect, a similar question on the same knowledge-level will appear. This way, questions are adapted to the level of the pupil. The teacher, at the same time, gets an overview of the progress of all his pupils and can see to what extent questions are being answered correctly, also related to the learning goals. With this data, the teacher can adapt their instruction to the needs of the pupils. This adaptive approach leaves room to give differentiated instruction to pupils on their own level, while also ensuring that every pupil is being challenged.

Figure 1 shows an example of the overview the teacher gets to see. On the left are all the pupils mentioned, and next to their name there are colour-coded squares to show whether the exercise was correct or not.

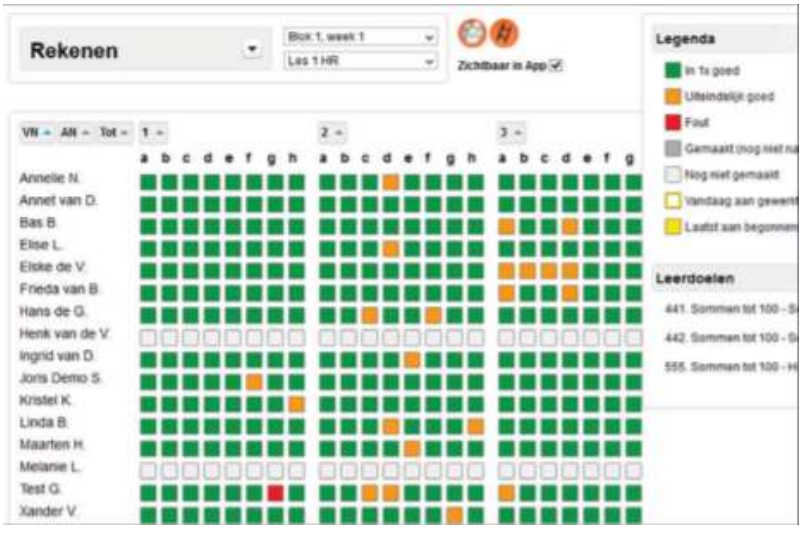

Figure 1: overview of students' progress. Retrieved from: http://static.snappet.org/downloads/pdf/quickstartguide.pdf

It is not surprising that software programs like Snappet are gaining popularity. It is built upon a broad scientific foundation, using multiple educational concepts. As research

showed, these principles are supposed to enhance learning and instruction, making it a promising software. Therefore, one would expect that working with Snappet has a positive effect on student performance as well. This study will show to what extent these expectations are met in practice. Based on existing research, the expectation is that Snappet enhances mathematics performance and that this difference might increase over time. This results in the following hypotheses:
H_0: there is no significant difference between the scores of the experimental group versus the control group.
H_1: the performance of mathematics of the experimental group increases over time.

Method

Design

The research question of this study is: "*What is the effect of the use of digital adaptive software for mathematics in primary education on the results of the standardised test at the end of primary school (after being used for 3 years)?"*
This question was answered using existing quantitative data that was gathered in the past years. Secondary data analysis was performed. This provided a longitudinal view of the change in mathematics performance. Considering existing research outcomes, expectations are that the performance in mathematics will increase over time when the software programs are used. This results in the following hypotheses:
H_0: there is no significant difference between the scores of the experimental group versus the control group.
H_1: the performance of mathematics of the experimental group increases over time.

Procedure

To identify the schools that fit the sample criteria, their instructional methods were analysed. Schools publish a yearly booklet, providing an overview of the main content and regulations per school, including the instructional methods used. The booklets from 2016-2017 were analysed via text mining. All school booklets that mentioned the word 'Snappet', were registered in a different file for further analysis. When a school booklet mentioned the software and also the year in which they started using the software, this was registered in an excel-file.

After that, the control sample was composed using the remaining school booklets that did not pass the text mining process. These school booklets were analysed and when they mentioned another method than the one mentioned before and made no comment about using tablets or laptops extensively during mathematics, they were included in the control sample. When both samples were complete, the weight score of each school that was calculated over the years of 2016 to 2019 (Inspectie van het Onderwijs, 2019) was added. Every school in Dutch primary education has a weighted score ('schoolweging') that is expected to influence the outcomes of that school. This score is calculated by Statistics Netherlands (CBS, n.d.) and is based on several variables, such as the educational background of the parents, the average educational level of all mothers to that school, the native land of the pupils' parents, the duration of the mothers' stay in the Netherlands and whether or not the parents are in a

financially stable position (Inspectie van het Onderwijs, 2020). This weighted score ranges between 20 and 40 and is normally distributed. The higher the score, the lower the expected outcomes are. Schools with absent weight scores in the dataset of Statistics Netherlands were deleted from the sample. This resulted in two samples of each 142 schools.

The final samples were registered in an SPSS-file. This file was then sent to Statistics Netherlands (CBS, 2020) where all the schools' names were anonymized and coded. With these codes, it was possible to link the school to the scores on the CITO-test in the relevant years without identifying individual students or schools. The data consisted of scores on the CITO-test at the end of sixth grade. This test is standardised and aims to objectively measure the level of the pupils when they leave primary school (College voor Toetsen en Examens, 2019). The test covers multiple subjects but this study focused on the scores for mathematics. This study used the test scores of the years 2015-2016 to 2018-2019. This provided a longitudinal view on the subject as well as the chance to analyse the data further and examine the changes over the years.

Unfortunately, the CBS-data did not contain all data of the CITO-test for all schools included in the samples. Schools that did not provide enough data were excluded from the sample. This is probably caused by the fact that some schools have the choice to test the skills of their pupils with a different test. In those cases, CITO-data was lacking. Deleting these schools resulted in a total of 72 schools in the experimental sample and 74 schools in the control sample. Because the schools were coded and anonymized, it was not possible to analyse which schools were eliminated and whether these schools had specific characteristics in common (e.g. focus on digitalisation in general). A summary of the final sample sizes can be found in Table 1.

Table 1 Frequencies in both samples

	2016			2017			2018			2019		
Group	N	Missing	Total	N	Missing	Total	N	Missing	Total	N	Missing	Total
Control	74	0	74	73	1	74	69	5	74	65	9	74
Experiment	72	0	72	71	1	72	59	13	72	66	6	72

Samples of schools

Initially, two samples of 142 schools were drawn. This size was deemed enough to account for schools that might be excluded from the samples if they would appear to be unsuitable when analysed into further detail. The schools were collected by analysing publicly available school information. Schools were included in the experimental sample if they mentioned the usage of Snappet for mathematics lessons. Next to that, the software needed to be in use in the academic year 2016-2017 or earlier in Grades 3 to 6. This to ensure that the sixth-grade pupils whose test scores were analysed, used the software for at least three years in 2018-2019. Schools were included in the control sample when they mentioned another mathematics method than the ones mentioned above and did not mention extensive usage of tablets or chrome books in every-day class. In both samples, a few demographic characteristics were identified to make sure the samples were comparable to each other. Next

to that, these characteristics were identified to provide insight into the extent to which the samples provided a representative image of the population. The characteristics will be discussed in the next paragraph.

First, it was made sure the samples did not differ significantly on the average weight score of 2016-2019. The mean of the experimental sample was 29.92 (SD=3.11). The mean of the control sample was 30.03 (SD=3.70). It is desirable to have similar samples, to eliminate factors that otherwise might have an impact on the mathematics scores. An independent samples t-test showed equal variances with a t-value of .189(df=144) and a confidence interval overlapping zero (-1.23 to 1.01) meaning that the difference between both samples can be assumed to be zero. An overview of this data is shown in Table 2.

Table 2 School weight in samples compared

Group		School weight	Independent samples t-test		
	N	Mean (SD)	t	Sig. (2-tailed)	95% confidence interval
Control group	74	30.03 (3.70)	.189	.85	-1.23; 1.01
Experimental group	72	29.92 (3.11)			

Next to the software, schools use instructional materials in mathematics lessons. Schools are free to choose from certified instructional materials. The specific materials that were applied in the schools in the samples were identified to make sure this did not influence the results. A frequency analysis showed that all instructional methods were distributed similarly in both samples. This means that both samples used comparable instructional materials. The exact distribution of these materials and the relation to the population can be found in the Appendices (in Table 1).

Reliability and validity

The samples are as comparable as possible to make sure the differences found are not caused by anything else than the usage of the software.

Reliability was ensured by using data from a standardised test (College voor Toetsen en Examens, 2019). This is an objective measure for mathematics in primary school and does not relate to a particular method. This ensures an unbiased representation of the mathematics performance of pupils. Despite the effort of the creators of the CITO-test, the difficulty is likely to be slightly different each year. The absolute scores can, therefore, not be directly compared over the years. Because of this, a comparison was made between the samples for each year individually.

Instruments

The main instrument that was used is the CITO-test (College voor Toetsen en Examens, 2019). This standardized test tracks the performances of every pupil during primary school in several subjects. At the end of sixth grade, pupils make a final test that indicates their level at the start of secondary education. The test is based on nation-wide core goals and

uses reference-levels, which are defined by law (Wet referentieniveaus Nederlandse taal en rekenen, 2010). The test attempts to objectively measure each pupil's knowledge and skills. Statistics Netherlands gathers the data of these tests and provides details about the scores for each subject. For this study, only the scores for mathematics were analysed. The total score of the CITO-test ranges between 501 and 550. Most years, the mean of the population is close to 535.

Data analysis

For this study, analyses were done using SPSS 25. To identify the differences that already might have existed in the starting year of the study, a few independent samples t-tests were conducted. This was done with the mean CITO scores in 2016 and the mean percentile-scores for mathematics in 2016. After that, a two-level multilevel regression analysis was conducted. Pupils (level 1) are nested in schools (level 2), explaining part of the difference among pupils. Multilevel regression analysis provides an opportunity to account for this difference. A multilevel analysis offers insight into the effect of Snappet on the difference in mathematics results, while also accounting for the school-level. Three multilevel analyses were conducted, for the years 2017, 2018, and 2019. This way, it was attempted to gain insight into the development of the differences between pupils in schools that did and did not use Snappet.

Results

Results per year

2015-2016: The Baseline year

When interpreting results that give information about changing data, it is important to have a clear and unbiased baseline. To clarify the starting point of this study this baseline is presented in the following section.

First, an analysis was performed to tell whether the two samples are representing the same population when it comes to CITO-results. The mean of the control sample was 535.01 (SD=3.89) and the mean of the experimental sample was 534.49 (SD=3.57). An independent samples t-test compared the means to the population. It showed equal variances and a confidence interval overlapping zero (-1.81 to .63) meaning the distribution of the mean CITO-score is assumed to be equal to the population's mean CITO-score. With this result in mind, there is no reason to assume that the choice to use Snappet is related to the performance level of these schools. This is important to keep in mind when interpreting further results. A summary of this data is shown in table 3.

Table 3 Mean CITO-scores in 2016 compared

Group		CITO-Score	Independent samples t-test		
	N	Mean (SD)	t (df)	Sig. (2-tailed)	95% confidence interval
Control group	74	535.01 (3.89)	-.952 (144)	.34	-1.81; .63
Experimental group	72	534.49 (3.57)			

When focusing on mathematics, the baseline of the samples was the following. To identify the mathematics level of the schools in the sample, a mean percentile-score was calculated. This consisted of the percentile scores per student combined in a mean score per school. To check to what extent the schools are representing the population, an independent samples t-test was performed. This showed the following details. The mean of the control sample was 51.51 (SD=10.11) and the mean of the experimental sample was 50.54 (SD=10.01). This shows that, on average, both groups scored neither high nor low compared to the population. Next to that, the t-test showed that both samples are comparable and equal variances can be assumed. This result on top of the average CITO-score in the same year is a strong indication that the samples are representing the population well. A summary of this data is presented in Table 4.

Table 4 Mean percentile-scores for mathematics in 2016 compared

Group		Percentile-score	Independent samples t-test		
	N	Mean (SD)	t (df)	Sig. (2-tailed)	95% confidence interval
Control group	74	51.51 (10.11)	.580 (144)	.56	-2.32; 4.26
Experimental group	72	50.54 (10.01)			

2016-2017: One year of using Snappet

In 2017, the pupils took the standardised test after having used the software for one year. The group-factor (0= control group), describing whether a pupil was in the experimental or control group, did not appear to be significant (b= -3.65, *p*=.277). The p-value did not reach the boundary of .05 and the confidence interval was rather large. On the basis of this result, the null-hypothesis cannot be rejected, meaning no effect was measured after using the software. The school weight did have a significant influence on the mathematics scores (b=-.73, *p*=<.001). A negative correlation was found, meaning that pupils enrolled in a school with a higher mean weight score are likely to perform less on mathematics. These results are summarised in Table 5.

Table 5 2016/2017: Fixed effects after one year of Snappet

Parameter	b	*t*	*p*	95% confidence interval
Group (0=control)	-3.65	-1.091	.277	-10.26; 2.96
Mean School weight score	-.73	-6.276	<.001	.96; -.50

2017-2018: Two years of using Snappet

In 2018, the pupils were using Snappet for two years. The group-factor was not significant (b=1.70, *p*=.592). Again, the confidence interval was quite large and the p-value did not reach .05. No significant effect could be found, indicating that there is no reason to reject the null hypothesis. The school weight score appeared to be significant (b= -.72, *p*=<.001). This negative correlation was similar to the previous year and shows that pupils

enrolled in a school with a higher mean weight score are again likely to perform slightly less on mathematics. A summary of these results can be found in Table 6.

Table 6 2017/2018: Fixed effects after two years of Snappet

Parameter	b	*t*	*p*	95% confidence interval
Group (0=control)	1.70	.538	.592	-4.57; 7.98
Mean School weight score	-.72	-6.710	<.001	.94; -.51

2018-2019: Three years of using Snappet

After the third year of using Snappet, the group-factor was still not significant (b=-4.52, *p*=.084). There is no indication that Snappet influenced the mathematics results based on these samples. This result indicates that there is no reason to reject the null hypothesis, indicating no difference based on the use of the software. The school weight score remained a significant predictor of the mathematics results (b= -.60, *p*=<.001). The strength and direction of the correlation were similar to previous years. Table 7 contains a summary of these results.

Table 7 2018/2019: Fixed effects after three years of Snappet

Parameter	b	*t*	*p*	95% confidence interval
Group(0=control)	-4.52	-1.743	.084	-9.66; .62
Mean School weight score	-.60	-7.034	<.001	.77; -.43

Summary

Overall, the usage of Snappet was not a significant predictor of mathematics performance on the standardised test. Given the results, there is no reason to reject the null-hypothesis that stated that there is no difference between the two groups when looking at mathematics performance. The school weight, however, was a significant predictor of mathematics performance. This correlation is stable over time.

Conclusion

The goal of this research was to add to the existing research concerning the effectiveness of Snappet. Since existing research focused on rather short-term results, this study aimed to gain insight into the long-term effects of the software. The main research question in this research was: *What is the effect of the use of digital adaptive software for mathematics in primary education on the results of the standardised test at the end of primary school (after being used for 3 years)?*
To answer this question, a multilevel regression analysis was performed. Two groups of schools that were comparable based on previous performance and social-economic structure served as input. Pupils' results on the standardised test were analysed. To gain insight into the development of the mathematics performance, test results of several years were analysed.

Looking at the results, it can be concluded that using Snappet did not have a significant influence on the mathematics performance of pupils. Although previous studies showed otherwise, the difference in mathematic scores cannot be explained by the grouping variable. The difference in scores could partially be explained by the school weight. It was clear that the school weight of the pupils' school was of significant influence on their mathematics performance. This is not surprising, considering the fact that schools with a low school weight usually exist of pupils with a more advantaged background (e.g. higher educated parents, parents financially stable, Dutch as native language) meaning these schools tend to score higher overall.

Considering these results, an answer can be formulated in response to the research question. Based on the data analysed in this research, there is no indication that the use of Snappet in primary education affects the results on the standardised test on mathematics that pupils take at the end of 6th Grade. All results suggest accepting the null-hypothesis.

Discussion

The findings of this study do not support the results of previous research. There are a few possible explanations for that. This chapter will begin with addressing these explanations, followed by the practical and scientific implications of this study. After that, the limitations of this study will be described, and a final conclusion will be given.

First of all, the biggest difference between previous studies and the currents study is that this study measured data that was not collected through an intervention. It was a collection of results based on the day-to-day practice in these schools rather than a snapshot of a designed environment. This, in combination with the long-term focus of this study, might be the reason why the studies resulted in different conclusions. It could be that, because teachers and pupils were aware of the experiment, they acted differently. Even unconsciously, teachers might have been more self-aware because they were working as part of an experiment and pupils might have been excited to work with something new, increasing their motivation for a while. All these scenarios are likely to have influenced the results at least a bit. This factor, caused by the novelty and the experimental setting, was eliminated in this study. Because the

data was collected without the schools knowing and without them being part of an experiment, the results paint a picture of their regular performances.

Second, in the research of Faber, Luyten & Visscher (2017), a positive, significant effect was found. In their experimental study, the post-test showed significantly improved results after a 6-month intervention. A few explanations why this result was not replicated in this study are the following. One important finding from their study was that high-performing students improved more with this software than low- or medium-performing students. Because of the set-up of this study, this factor could not be taken into account. This study used data that was already gathered and the average performance of students was not part of this data. This differentiated effect was therefore not accounted for in this study, which might be the cause of the lacking significant results. Next to that, the extent to which teachers are being trained and coached to use the software to its full potential could have played a role. In the study of Faber, Luyten & Visscher (2017), all teachers in the experimental group received an introduction and training beforehand and also had a coach available to help them with any problems that might come up. Because the current study gathered data in hindsight, it is not clear to what extent this was the case for the schools in the experimental group. Training the teachers might have been a moderating factor in the effectiveness of the software. Lastly, Faber, Luyten, and Visscher (2017) used Snappet log files to identify the intensity of the use of the software and this also appeared to correlate with the eventual post-test scores. Log files were not collected in this study and this factor was therefore not part of the study. It could be that the schools in the experimental group did not use the software with the same intensity, decreasing the final effect. Although these claims are suggestions, they are important to mention as the combination of them might be the cause of the discrepancy between previous and present results.

Molenaar & Knoop-van Campen (2017) also found a positive effect after their quasi-experimental study. Their results added to Faber, Luyten & Visscher (2017), concluding that especially high-performing pupils benefitted from using Snappet. The discrepancy between these results and the ones in this study might be caused by the small sample that Molenaar & Knoop-van Campen (2017) used. This sample size might have affected the results. Both the studies of Faber, Luyten, & Visscher (2017) and Molenaar & Knoop-van Campen (2017) found that the performance level of students at the beginning of the intervention was a predictor of the effect the software would have. This study did not account for this difference and the effects of the high- and low-performance pupils might have cancelled each other out.

Limitations

This study has a couple of limitations. First of all, the fact that secondary data analysis was performed based on data that already existed, was both a strength and a weakness. The advantages of this approach have already been discussed, but there were downsides to this approach as well. Because of this approach, details about the two groups were lacking. It was not clear to what extent the software was implemented exactly. Since Faber, Luyten & Visscher (2017) found that the intensity of the software usage influenced the effect it caused, this was important information that was now lacking. It is also unknown if, and to what extent the teachers working with the software were trained to do so. Furthermore, more details about

the control group are necessary to exclude factors that might cause the lack of significant results. It could be that the control group did not use the software, but did use another method that included the effective elements of the software (adaptive approach, differentiated instruction, and/or formative assessment). If this would have been the case, it could explain the lack of difference between the two groups. Future studies could take this into account but this would require in-depth analysis of all materials used and in-class observations, since the practice of the teacher is also influencing the amount of differentiation and adaptation. This information was not part of the school booklets that were used in this study.

Next to that, the main source used to indicate which schools were using the software was their information booklet that most schools publish online. These booklets are also used to showcase the best assets of the school, and they are a somewhat biased source because of that. When the software was mentioned as a method that was in use for mathematics this school was included into the experimental sample, but it was not always clear since what year this software was implemented. The only thing that was checked, was whether the software was in use in 2015-2016, to make sure all pupils in the test of 2019 worked with the software for three years.

The final limitation concerns the fact that the data did not account for the difference in the difficulty of the standardised test for each year. Each year, the test is slightly different and the exact difficulty inevitably differs over time. There is a reference-score that is calculated by the developers of the test that accounts for this difference and this was the score that would have been ideal for this study. However, due to COVID-19 and the holiday period in which this study took place, it was not possible to obtain this data in time.

Recommendations and future research

To improve further research, a lesson can be learned from the limitations of this study. First and foremost, more information should be gathered about both samples. The exact implementation of the software should be indicated, ideally from the log files that Snappet provides. It should also be identified to what extent the teachers are trained to work with the software. Next to that, the control group should be investigated in more detail. Important details about the instructional materials that are being used, should be noted to make clear exactly what is the difference between both groups. The long-term and large scale approach of study are strong points and should be integrated with future research as well. If future research takes place, it should focus on the reference score for mathematics to gather data that is comparable over time. A final recommendation to the scientific side of this topic is to account for the differentiated effect this software causes for high- and low-performing pupils.

Practical implications

The software that was tested in this study, Snappet, is created by a company, meaning one of the goals is to sell their product and make a profit. Especially in education, it is important to thoroughly test new interventions and instructional materials instead of implementing them without a second thought. This software serves multiple goals. First of all, it claims to improve performance. As discussed, there is evidence that both supports and rejects this claim. An advantage of Snappet is that it will save teachers time and effort. They no longer need to check all work by hand but they can look at their dashboard and

immediately have an overview of the performance of their pupils. This way, their instruction can be more focused on the problems that do arise, making it easier to have an effective, differentiated instruction. Considering the busy agenda teachers have, this is an important advantage that should not be overlooked.

Another implication one should consider is that this type of software seems to benefit one group of pupils more than the other. High-performing pupils appear to benefit most (Faber, Luyten & Visscher, 2017; Molenaar & Knoop-van Campen, 2017). Previous research also pointed out the effect that ability grouping has on both high- and lower-performing students (Van Damme, Opdenakker, & Van Landeghem, 2006; Faber, Glas & Visscher, 2018) It appeared that being in a high-performing class is beneficial for high-performing students. This effect was not found for average- and low-performing students. More recent research (Buttaro & Catsambis, 2019) added to this, concluding that grouping students to their ability is increasing the differences among them since high-performance students, again, benefit most. These conclusions should be taken into account when looking at adaptive, digital software like Snappet. It should be noted that not all students benefit equally and it is advised to implement this software with this in mind. Some pupils might benefit from it and this should be encouraged, but some pupils might not and their need for an additional approach should not be overlooked.

The reader should realise that the absence of an effect found in this study is not a judgment of the quality of the software. The software can have a positive impact on both teachers and pupils and it can offer a solution for several issues and optimize day-to-day education. One should, however, be cautious when implementing it, realising the complicated nature of education where no solution is the answer to everything.

References

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education Principles Policy and Practice, 18*(1), 5-25. Doi: 10.1080/0969594X.2010.513678

Black, P. & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability, 21*(5), 5-31. Doi: 10.1007/s11092-008-9068-5

Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher, 13*(6), 4-16.

Bolkan, S. (2015). Intellectually stimulating students' intrinsic motivation: The mediating influence of affective learning and student engagement. *Communication Reports, 28*(2), 80-91.

Buttaro, A., Catsambis, S. (2019). Ability Grouping in the Early Grades: Long-Term Consequences for Educational Equity in the United States. *Teachers College Record, 121*(2), 1-50.

CBS (n.d.). Organisation. Retrieved on 20-04-2020 from: https://www.cbs.nl/en-gb/about-us/organisation

CBS (2020). Microdata: conducting your own research. Retrieved on 29-06-2020 from: https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research

College voor Toetsen en Examens (2019). *Terugblik Centrale Eindtoets 2019*. Retrieved from https://www.centraleeindtoetspo.nl/publicaties/publicaties/2019/11/11/terugblik-centrale-eindtoets-2019

Deci, E., & Ryan, R. (2000). Intrinsic and extrinsic motivations: Classic definition and new directions. *Contemporary Educational Psychology, 25*(1), 54-67.

Drijvers P. (2018) Empirical Evidence for Benefit? Reviewing Quantitative Research on the Use of Digital Tools in Mathematics Education. In: Ball L., Drijvers P., Ladel S., Siller HS., Tabach M., Vale C. (eds) Uses of Technology in Primary and Secondary Mathematics Education. ICME-13 Monographs. Springer, Cham

Ericsson, K., Th. Krampe, R., & Tesch-Romer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review, 100*(3), 363-406.

Faber, J., Glas, C., & Visscher, A. (2018). Differentiated instruction in a data-based decision-making context. *School Effectiveness and School Improvement, 29*(1), 43-63.

Faber, J., Luyten, H., & Visscher, A. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, (106), 83-96. Doi:https://doi.org/10.1016/j.compedu.2016.12.001

Hattie, J. & Timperly, H. (2007). The power of Feedback. *Review of Educational Research 77*(1), 81-112. Doi: 10.3102/003465430298487

Hopster-den Otter, D., Wools, S., Eggen, T., and Veldkamp, B. (2019). A General Framework for the Validation of Embedded Formative Assessment. *Journal of Educational Measurement, 56*(4), 715-732. doi:10.1111/jedm.12234

Ikwulmelu, S., Oyibe, O., & Oketa, E. (2015). Adaptive teaching: an invaluable pedagogic practice in social studies education. *Journal of Education and Practice, 6*(33), 140-144.

Ismajli, H. & Imami-Morina, I. (2018). Differentiated Instruction: Understanding and Applying Interactive Strategies to Meet the Needs of all the Students. *International Journal of Instruction, 11*(3), 207-218.

Inspectie van het Onderwijs. (2019). Schoolweging 2016/2017, 2017/2018, 2018/2019. Retrieved from: https://www.onderwijsinspectie.nl/documenten/publicaties/2019/10/07/schoolweging-van-alle-scholen-2019-2020

Inspectie van het Onderwijs. (2020, January). Onderwijsresultatenmodel PO. Ministerie van Onderwijs, Cultuur en Wetenschap. Retrieved from https://www.onderwijsinspectie.nl/onderwerpen/onderwijsresultaten-primair-onderwijs/documenten/publicaties/2019/12/10/onderwijsresultatenmodel-voor-het-primair-onderwijs

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence* (REL 2017–259). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from http://ies.ed.gov/ncee/edlabs.

Molenaar, I., & Knoop van Campen, C. (2016). Learning analytics in practice: The effects of adaptive educational technology snappet on students' arithmetic skills. *Proceedings of the sixth international conference on learning analytics & knowledge (LAK '16)* (538-539). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/2883851.2883892

Siddique, N., Dhakan, P., Rano, I., & Merrick, K. (2017). A review of the relationship between novelty, intrinsic motivation and reinforcement learning. *De Gruyter Open,* (8), 58-69.

Shelle, G., Earnesty, D., Pilkenton, A., & Powell, E. (2018). Adaptive learning: An Innovative Method for Online Teaching and Learning. *Journal of Extension, 56*(5), 1-9.

Toetsbeleid PO (juni 2014). Retrieved on 05-03-2020 from: https://wetten.overheid.nl/BWBR0035216/2019-08-01

Van Damme, J., Opdenakker, M. C., & Van Landeghem, G. (2006). *Educational Effectiveness. An Introduction to International and Flemish research on schools, teachers and classes.* Leuven: Uitgeverij Acco.

Van der Kleij, F., Vermeulen, J., Schildkamp, K., & Eggen, T. (2015). Integrating data-based decision making, assessment for learning, and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy and Practice, 22*(3), 324–343, DOI: 10.1080/0969594X.2014.999024

Weltman, H., Timchenko, V., Sofios, H., Ayres, P. & Marcus, N. (2019). Evaluation of an adaptive tutorial supporting the teaching of mathematics. *European Journal of Engineering Education, 44*(5), 787-804, DOI: 10.1080/03043797.2018.1513993

Wet Nederlandse taal en rekenen. (2010, April 29) Retrieved from https://wetten.overheid.nl/BWBR0027679/2020-08-01

Yenmez, A. & Özpinar, I. (2017). Pre-service education on differentiated instruction: elementary teacher candidates' competences and opinions on the process. *Journal of Education and Practice, 8*(5), 87-93.

Ysseldyke, J. & Bolt, D. (2007). Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement. *School Psychology Review, 36*(3), 453-467.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C. & Lemkuil, A. (2003). Using a Curriculum-Based Instructional Management System to Enhance Math Achievement in Urban Schools. *Journal of education for students placed at risk, 8*(2), 247-265.

Appendices

Table 1. Distribution of methods used

Method	Control group N(%)	Experimental group N(%)
Wereld in getallen	85 (59,86%)	60 (42,25%)
Alles telt	15 (10,56%)	23 (16,20%)
Pluspunt	26 (18,31%)	32 (22,54%)
Rekenrijk	14 (9,86%)	10 (7,04%)
Snappet	0 (0%)	2 (1,41%)
Exova Math	0 (0%)	3 (2,11%)
Wizwijs	0 (0%)	1 (0,70%)
Reken zeker	2 (1,41%)	2 (1,41%)
Onbekend	0 (0%)	9 (6,34%)
Total	142 (100%)	142 (100%)