The effect of individualized digital practice at home on math skills Evidence from a two-stage experiment on whether and why it works

Computers & Education 113 (2017) 119–134

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

The effect of individualized digital practice at home on math skills—Evidence from a two-stage experiment on whether and why it works

Carla Haelermans*, Joris Ghysels [1]

Top Institute for Evidence-Based Education Research (TIER), Maastricht University, PO Box 616, 6200 MD, The Netherlands

ARTICLE INFO

Article history:
Received 27 June 2016
Received in revised form 11 April 2017
Accepted 18 May 2017
Available online 24 May 2017

JEL-Classification:
I21
H75
C93

Keywords:
Field experiment
Digital practice tool
Individualization
Numeracy
Secondary education

ABSTRACT

This paper analyses an experiment on the effect of an individualized, digital practice tool on numeracy skills for 337 seventh grade students. The first stage of the experiment shows that offering students the opportunity to practice numeracy digitally at home (intent-to-treat) leads to a substantial and significant increase in numeracy performance growth. The second stage reveals that the effectiveness of the tool mainly stems from its individualized nature. With good implementation prospects and relatively low costs, the consequences are discussed to be potentially large.

1. Introduction

Basic math and language skills (numeracy and literacy) are generally recognized as major components of human capital. Furthermore, they are important in the daily functioning of potentially every human being and they are well documented to contribute to labor market success (see e.g. Chiswick, Lee, & Miller, 2003; Hanushek, Schwerdt, Wiederhold, & Woessman, 2013; Vignoles, De Coulon, & Mercenaro-Gutierrez, 2011). Despite this seemingly self-evident statement, many students are observed not to have sufficient basic math and language skills (Funnekotter, 2012; KNAW, 2009; OECD, 2013). In the Netherlands, for example, both a special parliamentary commission (Commissie Meijerink, 2008) and the Royal Netherlands Academy of Arts and Sciences (KNAW, 2009) concluded some years ago that urgent action was required, because an increasing number of students lack the necessary numeracy and literacy skills. This call for action has led to the introduction of a compulsory numeracy section in the already existing national graduation exam program for secondary education. In turn, the

* Corresponding author.
E-mail address: Carla.Haelermans@maastrichtuniversity.nl (C. Haelermans).

[1] The authors would like to thank Jan-Hein de Wit and Dorien Stals from Dendron College for making this field experiment possible. Additionally, they express gratitude for the contributions of Freek Weeda and Theo Schijf. This paper has greatly benefited from the comments and feedback from Trudie Schils and Anders Sternberg.

http://dx.doi.org/10.1016/j.compedu.2017.05.010
0360-1315/

introduction of these exams has motivated schools to formulate policy documents on how to improve numeracy and literacy skills in the most efficient way.

The policy plans draw on the scientific consensus that individual differentiation is the key to higher student performance (e.g. Hattie, 2009), where traditional classroom settings only partly allow schools to differentiate their teaching between individual students. The combination of the increase in computer use in education, the need for individualization in the learning process and the decrease in basic math and language skills has led to the development of individualized Information Technology (IT)-tools aimed at developing these skills, especially in K-12. Accordingly, in their search to improve students' numeracy and literacy skills, many schools started using individualized IT-tools, often outside regular school hours (at home). This seems comparable to the digital tutoring in the out-of-school-time program, although many Dutch schools use it for *all* students and not just disadvantaged students (see e.g. Heinrich et al., 2014). Individualized IT-tools focus on an individual learning path for the student, adapting the exercises available for the student to the skills that he or she is lacking.

However, the existing literature contains only few experimental studies on the effect of IT-tools on numeracy and literacy performance (Arroyo, Park Woolf, Royer, Tai, & English, 2010; Barrow, Markman, & Rouse, 2009; Borman, Benson, & Overman, 2008; Pilli & Aksu, 2013; Rouse & Krueger, 2004). Furthermore, it is unclear how best to use these IT-tools, e.g. with respect to where and when to practice or the amount of teacher involvement. Hence, it is unclear whether these schools chose an effective teaching program and use it in an effective way.

Therefore, the purpose of this paper is, first of all, to analyze the effect of individualized educational software developed in the context of the above described policy change in the Netherlands. Furthermore, the purpose is to analyze *why* this software is effective and under what circumstances it is most effective. We conduct an experiment with an interactive digital practice tool and analyze the effects of offering this tool to students (intent-to-treat) on numeracy performance of students in 7th grade (age 12, first year of secondary school in the Netherlands). In this experiment, we are able to also take into account the intensity of treatment and the influence of the teacher. In the first stage, we show that the effect of the digital practice tool is about 0.40 of a standard deviation. Furthermore, we show that the non-compulsory training complements class-based training and is effective regardless of the math class and the teacher and the teachers' attitude towards the tool. The second stage shows that the effectiveness of the digital practice tool is due to the individualized differentiation (similar to the finding in Barrow et al., 2009).

In the literature, we see that many of the previous evaluations of information technology (IT) are wide in scope, as they evaluate, for example, increased budgets for IT either for schools or for households. They rely on the assumption that users have sufficient skills to implement and use IT to their benefit and that it does not matter how IT is used, in order to benefit educational outcomes. Yet, in practice, these general evaluations offer mixed results (e.g. no significant effect of IT: Goolsbee and Guryan (2006); positive effect of IT: Machin, McNally, and Silva (2007); negative effect of IT: Angrist and Lavy (2002), Leuven, Lindahl, Oosterbeek, and Webbink (2007)).

A second part of the literature on IT focuses on the comparison of computer directed versus traditional classroom teaching. A couple of meta-analyses apply strict selection criteria with respect to methodology used in the individual studies (Cheung & Slavin, 2012, 2013; Kulik & Kulik, 1991; Means, Toyama, Murphy, Bakia, & Jones, 2010) and show that in general, computer directed instruction does have small positive effects on student performance, compared with traditional classroom teaching, for both math and language.

Thirdly, IT proves particularly suited to provide individualized differentiation (from now on: individualization), with its algorithms allowing for individual learning paths. Incorporating the differences in level, interests and learning styles between students is shown to improve students' motivation (Tomlinson, 2004), and neglecting these differences might lead to decreased performance of certain students (Tomlinson & Kalbfleisch, 1998).

Evaluations of IT-based individualization programs in math and numeracy range from general teaching to remedial programs and cover both general student audiences and students with learning disabilities. In general, evaluation outcomes tend to be positive. Burns, Kanive, and DeGrande (2012) show that significantly fewer of the students at risk for math difficulties before, were still at risk after using a computer delivered math fact intervention. Similar results are found by Pilli and Aksu (2013). Banerjee, Cole, Dufflo, and Linden (2007) report on the positive outcomes of an experiment with an IT-based math remedial program, introduced in public schools of two cities in India, which illustrates that the benefits of IT-individualization are not confined to students from highly technologized societies.

The before mentioned three studies all analyzed 3rd and/or 4th grade students. There are only a few academic publication using a similar age group as in the study at hand. Arroyo et al. (2010) analyzed 250 7th and 8th grade students that used a digital skill drill method, or traditional practicing on paper, 15 min per day next to math classes, for four days, and find a significant positive effect of digital practicing. Barrow et al. (2009) also perform a randomized experiment, under 1605 middle and high school students, and show that treated students score significantly higher on pre-algebra and algebra skills than their counterparts who received traditional instruction.

Although it is possible that publication bias distorts the conclusions on specific programs more than in the case of the general IT evaluation, the former group of evaluations offers a range of positive experiences to build on. However, they do not go into detail in why these tools work or how to implement them. Furthermore, a potential hindrance to the rapid expansion of educational innovation through IT is acceptance by teachers. On the one hand, teachers often do not want interference in their classroom, and especially elder teachers often do not believe in the benefits of IT training. On the other hand, interventions and innovations are often imposed by the management, without consulting the teachers, which also might lead to resistance by teachers. As we study the introduction of a type of software that does not require a large teacher investment,

Table 1
Descriptive statistics of first year students of Dendron College in the school year 2012/2013.

	Obs.	*Average*	*St. Dev.*	*Min*	*Max*
Primary school ability test: numeracy	323	56.83	24.30	4	100
Primary school ability test: literacy (mother tongue only)	322	56.34	24.25	3	100
Primary school ability test: world studies	290	55.18	25.73	0	100
Primary school ability test: total score	328	538.28	6.21	517	550
Age (in completed years)	337	12.28	0.47	11	14
Student diagnosed with dyslexia (number of students)	21				
Female student (number of students)	189				
Oldest child in her/his household (number of students)	196				

called "Mousework[2]", we present results with exceptionally promising implementation prospects. The Mousework program we put to the test is a digital practice program that students are supposed to use at home and that complements math classes without being tightly linked to the pace of teaching.

Based on the literature, we hypothesize that allowing students to use the adaptive digital practice program by giving them access will increase student performance and will make more students reach the legally required reference level.

In analyzing this tool, this paper contributes to the literature in four ways. First, the experimental design allows for studying the effect of practicing with an interactive digital tool on numeracy skills of students in 7th grade, since we can use the random assignment over classes as exogenous variation that can be used for identification, while controlling for a bunch of student and teacher characteristics. However, since randomization took place at the class level with only 13 classes, we have to explicitly control and correct for this, as well as for the teacher, by using the relatively new method of bootstrapped clustered standard errors. Furthermore, the results should be interpreted with some caution. Second, due to the two different stages of the experiment, we can not only study if this interactive digital tool is effective, but also why it is effective. Third, we show effects of non-compulsory training that complements class-based training. This implies that the training can take place, and that students can gain in performance, regardless of the math class and the teacher attitude. Last, in this experiment, we can include extremely rich information on these students, their teachers and the context of the experiment, which means we also have information on the intensity of the treatment; i.e. we do not just have a dummy variable to measure who had access to the tool (although that is the main focus of this paper) but we also know if, when, what, how often and how long students used the practice tool.

The remainder of this paper is structured as follows: Section 2 presents the context of the experiments, e.g. the purpose, contents and organization of the digital practice tool, the identification strategy, the use of the digital practice tool, and measuring the numeracy skills of secondary students. Section 3 presents the empirical model, the baseline results and the regression results of the first stage of the experiment. In Section 4 we discuss the robustness checks of the first stage. Section 5 discusses the regression results of the second stage of the experiment. In Section 6, we conduct a cost benefit analysis and Section 7 concludes the paper and discusses the findings.

2. Context of the experiment

2.1. The school under study

The school under study, Dendron College, is - to Dutch standards - a mid-sized school for secondary education (junior high and high school). Dendron College offers secondary education in all tracks[3] and is tracking students from the first year on in several prevocational, general and pre-university tracks. Compared with the average Dutch secondary school, Dendron College has about 2000 students (national average M = 1473, SD = 1142), about 137 fte teachers employed (national average M = 130, SD = 101), a graduation percentage of 92 percent (national average M = 90, SD = 5), an average national exam grade of 6.5 (on a scale from 1 to 10) (national average M = 6.4, SD = 0.2) and a higher share of students that need additional support, namely 14 percent (national average M = 11 percent, SD = 7).[4] These statistics indicate that this school is representative for the average Dutch secondary school as it is within half a standard deviation of the average of all variables.

In school year 2012/2013, 13 first year classes (equivalent to seventh grade in the US) were part of the experiment (337 students), ranging from the more theoretical prevocational track to the pre-university track. The age of the students in the experiment ranges from 11 to 14 (differences are mainly due to grade repetition), 6 percent of students are diagnosed with dyslexia by an external organization (21 students) and 56 percent, 189 students, are girls (see Table 1). Furthermore, students

[2] The Dutch name is "Muiswerk".

[3] Dutch secondary education has a tracking system from 7th grade on, with 3 different tracks (prevocational education, which consists of 4 sub tracks where level 1 is the lowest (mainly practical) track and level 4 the highest (mainly theoretical) track, general higher education and pre-university education).

[4] The data are from 2012, and are obtained from the governmental website containing the Dutch open education data (https://www.duo.nl/open_onderwijsdata/databestanden/vo/).

attended 25 different primary schools. As such the school is a typical representative of schools outside of the highly urbanized, central region of the Netherlands (the “Randstad”).

2.2. Purpose, contents and organization of the interactive digital practice tool

The purpose of the interactive digital practice tool is to help students practice their numeracy skills, while being able to individualize, and give users direct feedback (Muiswerk, 2013). Although the program is mainly being used in the Netherlands, it also has an international version and is used by several international schools both in Europe and other parts of the world. In the Netherlands, around half of the schools use the program Muiswerk in some way, although only a small share of the schools uses the program in the way it is supposed to work best, which is the way Dendron College applies the program. However, there are other adaptive computer programs in the Netherlands, e.g. Gotit?! (see e.g. De Witte, Haelermans, & Rogge, 2015), that have similar features, as described below. The results of this study therefore also apply to other IT-tools in education that focus on numeracy skills, while offering a fully individualized practice track, based on regular tests, and direct feedback.

The program is interactive and person specific. Students work at their own level and get those exercises that will help them improve the sub-aspects of numeracy they are not knowledgeable in yet, while some exercises are meant to keep up their already gathered knowledge. Students have a certain set of exercises available, covering all domains of numeracy, where they choose from when they log in to the system. The school uses this tool to make sure each student achieves the highest possible level of numeracy, given his/her abilities, and maintains the level achieved. It offers all students online access to the tool for use after school hours, at home. Currently, most math teachers at Dendron College are not using the tool in class, although the tool is also developed to that end.

A numeracy pretest determines students' level of different sub-aspects of numeracy, which in turn determines the types of exercises they have to start practicing with at home.[5] At regular intervals (supposedly biweekly, but in practice once every three to four weeks), students make a short computer test at school to determine for which exercises their skills are still lacking and for which exercises their knowledge level is good enough for the moment. After every test, the type and level of exercises a student can choose from are adjusted to their new skill level. Apart from that, adjustment is also based on performance while practicing in the tool. The individualization therefore makes sure that the right exercises are selected for the student, but in the end, until the next adjustment, the student decides in which order he practices the exercises, and whether he repeats an exercise or not. If he performs badly at an exercise, but does not choose to repeat it, it will remain in his selection of exercises, even after the adjustment.

The program functions in a highly individualized manner, as it starts with explanation screens (digital instruction), offers feed-back and it provides the student with either repetition or new learning modules on the basis of previous performance of the individual student. It works without teacher interventions, but teachers have access to a reporting module and some may incorporate knowledge of “Mousework” performance in their interaction with the students.

Math teachers are supposed to motivate students to practice with ”Mousework” at home and for checking students' practicing behavior. However, not all math teachers agree with the management to use this interactive digital practice tool school-wide and some of these teachers refuse to act in accordance with the responsibility to check the students' practice behavior. Therefore, we will control for the teachers' attitude towards “Mousework” when studying the impact of digital practicing.

2.3. The field experiment

Fig. 1 shows the timeline of the field experiment, which consists of a pre-experiment period and stages I and II. In spring of their final year in primary education, students register at their school of choice for secondary education. The secondary school uses the results of the standardized national exit exam and the recommendation made by the primary school teacher to assign students to the first year classes before the summer break of 2012. At the school under study, the assignment of students to classes is done randomly within the boundary of the ability grouping that forms part of the Dutch system of secondary education (“early tracking”). Assignment of teachers to these new first year classes is fairly random as well, given that they can teach a certain amount of classes each year, and it has to fit their (part-time) schedule. In summer, week 29/2012, the researcher assigned classes randomly to treatment and control group. Only two types of first year classes (5 prevocational classes and 8 higher general/pre-university classes) took part in the experiment. Two classes of each type were assigned randomly to the control group (107 students), whereas the other 9 classes are the treatment group (230 students).[6] These classes were taught by 7 math teachers, of whom 2 teachers have both a control and one or more treatment classes, whereas one teacher only has a control class, and the other teachers only have treatment classes. In week 33/2012 the school year started, and in the second week of the school year all students and their parents were informed about the experiment by

[5] A student questionnaire in spring 2013 shows that only 5 students do not have a computer at home to practice with. However, IP address data shows that these students have practiced with the tool at school, where there are computers available for students that do not have one at home.

[6] Note that the school did not allow more than 4 classes to be control group (i.e. to be excluded from practicing), since they were going to compensate those classes in the next semester, and having more classes in the control group was financially unfeasible, hence the unequal division of classes.

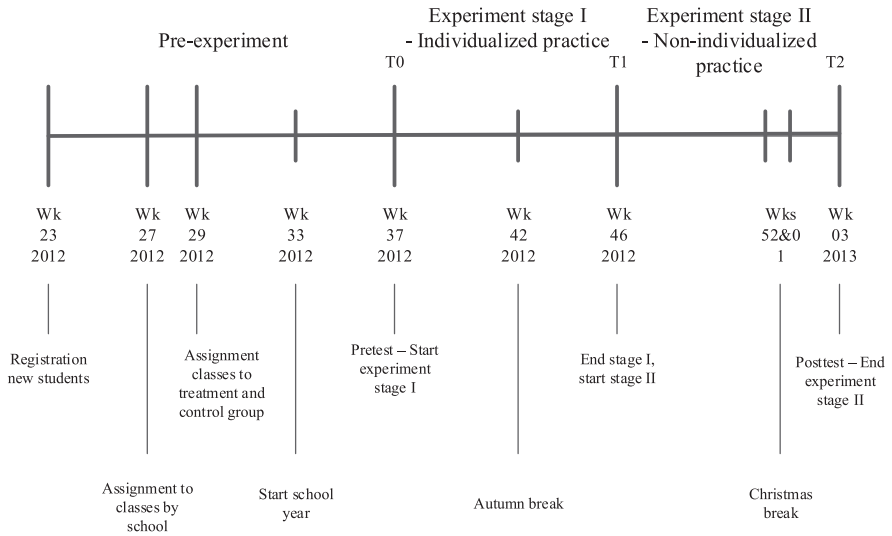

Fig. 1. Overview of the field experiment.

Table 2a
t-statistics, Mann-Whitney Statistics and Chi-squared statistics of Treatment and Control Group.

Variable	*Control group*			*Treatment group*			*T-statistic*	*P-value*
	N[a]	Average	Std. Dev.	N[a]	Average	Std. Dev.		
Primary school ability test numeracy	104	53.53	23.95	219	58.39	24.36	−1.69	0.09
Primary school ability test literacy	104	55.38	24.14	218	56.81	24.35	−0.49	0.62
Primary school ability test world studies	96	54.17	25.04	194	55.69	26.11	−0.47	0.63
Primary school ability test total score	106	537.73	6.29	222	538.56	6.17	−1.13	0.36
Age	107	12.27	0.47	230	12.29	0.47	−0.29	0.77
Pretest in September (T0)	107	56.05	13.60	230	54.18	13.91	1.15	0.25
Variable	N	Rank Sum	Expected	N	Rank Sum	Expected	*Mann-Whitney Z-score*	*P-value*
Primary School Advice	107	17 433.0	18 083.0	230	39 520.0	38 870.0	−0.84	0.39
Variable	N			N		Df	Pearson Chi2	*P-value*
Female	107			230		1	0.05	0.81
Dyslexia	107			230		1	0.10	0.74
Oldest Child	107			230		1	1.54	0.21
Country of Birth	107			230		3	2.92	0.40
Nationality	107			230		3	3.08	0.38
Religion	107			230		3	1.03	0.79
Situation at home	107			230		2	1.59	0.43

[a] Note that not all students wrote the primary school ability test and that not all primary schools delivered detailed information on the subparts of that test to the secondary school. Therefore, a small number of observations is missing for these variables.

means of a letter. In the fourth week of the school year, one of the researchers was present at the information evenings for parents to provide them with additional information regarding the experiment. Because the school provides children in the control group with extra lessons in numeracy and literacy in the second half of the school year, all parents agreed to their child's participation in the experiment.

The pretest took place in week 37/2012. Table 2 shows that there is no significant difference in performance between the treatment and the control group at the pretest. The first stage of the experiment lasted 8 weeks and the first output test (T1) took place in week 46. The second stage of the experiment lasted 6 weeks, but due to Christmas holidays T2 only took place in week 3 of 2013. The treatment and control classes were the same classes and the same students throughout both the stages of this study. The two stages are different in the sense that they have a different treatments state (as will be explained below), but are the same from a methodological point of view.

2.3.1. Content of the experiment

In experiment stage I, the digital practice tool is used as described above in Section 2.2. Treated students practice with the tool at home, and have an individualized learning path with selected exercises. In experiment stage II, we studied the effect of non-individualized digital practice with the tool as the school had decided, without deliberation, to give all students access to all possible exercises, regardless of whether those exercises fitted in their individualized learning route. Although this was not

Table 2b
t-statistics and Chi-squared Statistics of Teacher Characteristics of Treatment and Control Group.

Variable	*Control group*			*Treatment group*			*T-statistic*	*P-value*
	N	Average	Std. Dev.	N	Average	Std. Dev.		
Full time equivalent of teacher	4	0.97	0.06	9	0.92	0.14	0.77	0.12
Teacher experience at this school	4	10.2	2.7	9	9.3	4.2	0.44	0.67
Teacher age	4	54.25	10.78	9	57.44	4.42	−0.78	0.45
Variable	N			N		df	Pearson Chi2	*P-value*
Salary scale	4			11		1	0.41	0.52
Teaching degree	4			11		2	1.26	0.53

planned, this allowed us to make a comparison between individualized and non-individualized practice, and explore whether this could be the reason of the effectiveness of the tool. The crucial (and only) difference with experiment stage I was that the exercises available were no longer individualized. This means that students were no longer offered an exercising program tailored to their skill level, but were forced to make their own selection out of a much wider offer than they had previously. A priori it is not clear what motivated the selection of exercises they chose to practice with. They may have chosen whatever exercises they judged as most adequate, liked most or felt they could excel in easily.

The control group did not practice with the tool, in neither of the two stages. A student questionnaire shows that in the control condition, students simply invested less time in homework, given that practicing in the tool was additional homework to students.

2.3.2. Identification strategy

The main problem with determining the effect of a practicing tool is the potential correlation of unobservable factors with both the practicing behavior and the outcome variables, such as numeracy performance. In this study we use exogenous variation in the possibility and circumstances to practice through an experimental set-up, as explained above, although the experiment is not perfect due to the randomization at the class level.

Table 2a presents the observable characteristics of the treatment and control group in experiment stages I and II, for all students that wrote the pretest (N = 337), as well as the *t*-statistics/Mann-Whitney statistics/Chi-squared statistics on the differences between the groups. We use T-test for continuous variables, Mann-Whitney statistics for ordinal categorical variables and Chi-squared statistics for nominal categorical variables. Table 2b presents the *t*-statistics/Chi-squared statistics on the differences on teacher characteristics, between treatment and control group. These latter results should be viewed with caution, given the small numbers, but they do show that the absolute values for teacher age and experience at the school, as well as teaching degree and part time workers are very similar for teachers that teach the treatment and control group. Apart from the randomization, these statistics indicate that we can trust (with a significance level of 5%) the treatment and control group to represent the same population, and to not have very different teachers. However, these are simple bivariate analyses in which we do not take into account the multilevel nature of our study (i.e. the randomization at the class level). If we regress the treatment dummy on individual student characteristics, while clustering standard errors at the class and teacher level, none of the characteristics are significant either, not even at the 10% level.

Note in this respect, that we will evaluate the provision of access (intent-to-treat effect), rather than the effective use of the tool. We do so, because the former is more policy relevant. A school can provide access, but cannot force students to use the tool. We do, however, take into account how much the student has practiced.

2.4. Measuring numeracy skills and practice behavior

The numeracy skills are measured using digital standardized numeracy tests,[7,8] which are written by all 7th grade students at T0, T1 and T2 (see Fig. 1). These are standardized validated tests developed by the company of the tool, and these tests are based on other nationally validated tests. The reliability (Cronbach's alpha scores of between 0.79 and 0.92) and validity of these tests are analyzed yearly by the tool developer, based on norm data of several participating schools (Schijf & Schijf, 2014). Although the pre and posttest are digital tests that are developed by the same company as the tool and are administered in the same digital environment as the tool, the tests themselves are external to the practice exercise tool and do not

[7] Ideally, we would also like to see if there is an effect on regular math performance, next to the potential effect on performance on this specific numeracy test. However, regular math tests are not comparable (with respect to topics and skills required) with the numeracy skills that are being practiced in the tool. Therefore, we cannot make a valid comparison of the regular math performance of treatment and control group. We can, however, look at a numeracy test written in class during the first treatment stage. Although this numeracy test only covers the topics that were discussed in class (whereas the tool lets students practice with all topics, depending on the students' skill level) it can give an indication that the results of this study are not dependent on the tests used. We analyze this as a robustness check.

[8] Note that this school uses quite some digital material and digital tests for other courses as well, implying that control group students had a lot of exposure to digital teaching materials as well, and are therefore not disadvantaged in this respect.

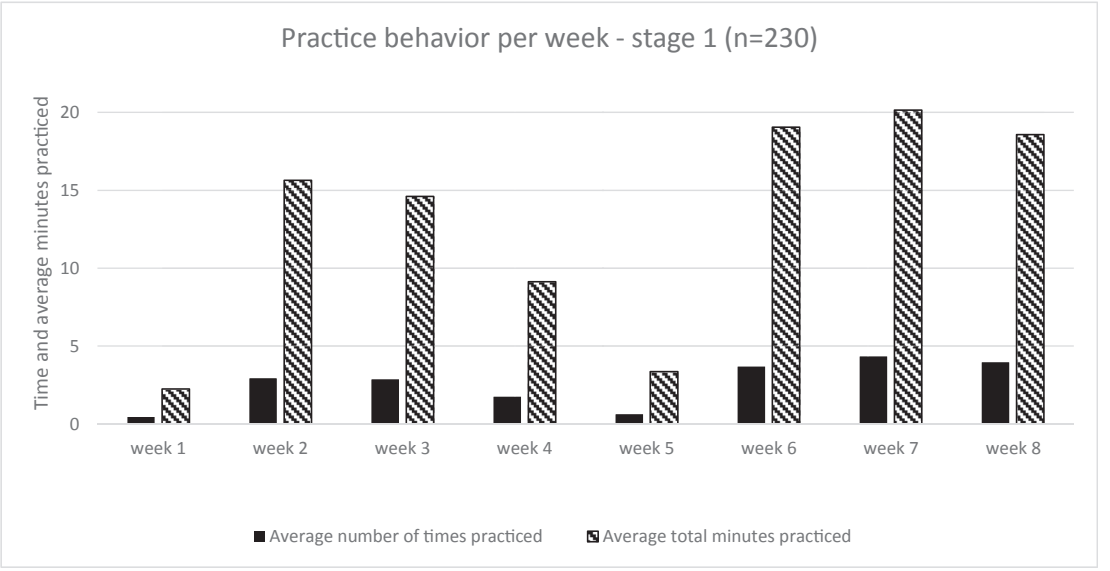

Fig. 2. Variation in practice behavior per week.

contain any of the exercise questions. The tests measure whether students have mastered the required national numeracy level (called 'reference level') they are supposed to have, given their age and given the fact that they finished primary school.

The numeracy test contains of relatively simple multiplication or addition questions, but also contains special understanding questions, where the student sees an unfolded shape and is asked to select the figure that could create the unfolded shape. Or the student is asked to calculate the volume of a sphere, or is asked to quickly make calculations by heart.

All students, both treatment and control group, practiced with digital multiple choice numeracy assignments in the testing program in the week before the pretest (T0) was administered, to make sure they knew what to expect when writing the pretest and to get acquainted with the testing environment. The test contains multiple choice questions and students were allowed to use scrap paper for their calculations, but no digital calculator. The tests lasted for about 20 min. Test scores can range from 0 to 100, 100 being the absolute maximum. For the analysis of experiment stage I, we use three different outcome measures: 1) the absolute score at T1, 2) the absolute growth in score between T0 and T1, and 3) and an indication whether the student has the same reference level in T0 and T1 or has improved by one reference level (dummy indicator, which is 1 if you have performed at the next reference level in the posttest, and 0 if you are still at the same level[9]). In stage II we cannot use the absolute scores at T2 as outcome measures, for the obvious reason that starting level is not comparable anymore, because of the previous stage of the experiment. Therefore, we use the absolute growth between T1 and T2. For the sake of simplicity and space we do not report the results on the references levels for experiment stage II. Another reason for this stems from the results of this second stage, which will become clear below in Section 5.

In addition to studying the effect of access to the digital practice tool, it is also descriptively interesting to include the intensity of the treatment, especially in stage II, where the variation in practice behavior is rather large. Below, we use the number of minutes practiced per week, which is determined by subtracting the weeks of school holidays from the total number of regular school weeks during the experimental period (i.e. excluding test weeks). Minutes practiced is registered by the digital practice tool for each time the student logs into the system, and for each exercise the student practices with. Students could still practice at home during school holidays, but the data shows that this was hardly the case. This leaves us with 7 weeks between T0 and T1 and 5 weeks between T1 and T2 (see Fig. 1).

Fig. 2 and Table 3 show the descriptive statistics of the practice behavior of students. Fig. 2 shows the spread of number of times practiced and the practice minutes over the weeks, and Table 3 shows the average practice minutes per week over the full period. Fig. 2 shows that the average minutes practiced per week is downsized by the first week, when students still had to start up and did not practice much. The fifth week was the autumn break, which was not taken into account in the calculation for average minutes per week. The practice numbers are pretty comparable for the other weeks.

As said before, there are (very) large differences in practice behavior in stage II, with 25 students not practicing at all. In stage I, only 21 of the treatment students did not practice, and the average amount of time students practiced was 13 min when those 21 students are included, ranging from an average of 0 min per week to 35 min per week[10], and 15 min when we only include the students that actually have practiced. In stage II, we see that the average amount of practice per week for all

[9] Note that, at the pretest, most students do not yet perform at the reference level they are supposed to have achieved by the end of primary school.
[10] Note that practicing minutes only count when at least one exercise is finished.

Table 3
Descriptive statistics of practice behaviour.

	Obs.	*Average*	*St. Dev.*	*Min*	*Max*
Minutes practiced per week stage I	230	13.34	9.76	0	40.25
Minutes practiced per week stage II	230	16.56	14.27	0	66.38
Minutes practiced per week stage I (if practiced at all)	209	14.68	9.22	0.56	40.25
Minutes practiced per week stage II (if practiced at all)	205	18.58	13.81	0.29	66.38

treatment students is 17, with larger standard deviations and maxima than in stage I. This amount increases to 19 when we only include students that actually practiced. Given the previously mentioned average amount of 45–60 min of math homework students write per week, an additional 15 min is an increase of math homework time of 25–33 per cent. So although the 15 min is only half of the supposed practice, it is still a considerable increase in math homework time for students.

2.5. Measuring teacher attitude towards the tool and teachers' use of the tool

In the beginning of the experiment a short questionnaire was handed out among math teachers, to gather information on their attitude towards the digital tool and on the way they use the digital tool (if they use it at all), for example in class or to check up on students. The questionnaire consisted of 21 multiple choice questions, among which 18 statements which are measured on a 5-point Likert scale. Using factor analysis, we found that of these 18 statements, 6 considered the attitude towards the digital tool, and we use the average of these 6 statements to measure “math teacher attitude”. This combined measure has a Cronbach's alpha of 0.78, where an alpha of 0.7 or more is acceptable (Field, 2013). Another 7 statements all considered the use of the tool by the teacher (based on experience from the previous school year), and these were combined into the new variable “math teacher use of tool” (this combined measure has a Cronbach's alpha of 0.84).[11]

2.6. Consideration on validity of the experiment

The school uses the interactive digital practice tool in addition to their math classes. All students, both in the treatment and control classes, are being taught mathematics using the math method that has been used over the previous years. For first year secondary students (7th grade) there are 4 math classes of 50 min each week, and students make an average of between 45 and 60 min of homework per week.[12] Practicing the basic skills with the digital tool is an additional activity that takes place outside the school. As these 13 classes are being taught by 7 different teachers (and the 9 treatment classes by 6 different teachers), there is large variation in practice behavior between classes in the treatment group, which, as we will show later, highly correlates with the teachers' attitude towards the tool.

Contamination is not likely to bias our results for various reasons. Students have their personal account with login information to practice at home and are only allowed to make the small tests at school. Therefore, it is hardly possible for students from the control group to gain access to the digital tool. Classes that were not allowed to work with the program simply did not have access, and the teachers did not mention the program at all during class. Because this is the first year of these students at this school (coming from primary education, which are always separate schools in the Netherlands), students are not used to work with the program, and in most cases do not even know of the existence of the program. Apart from that, most treatment students complain about the tool being ‘boring’, indicating that they do not practice with is because they want to, but because they feel a certain pressure from their teacher/the school, which makes it even more unlikely that there will be contamination effects of students in the control group gaining access to the digital practice tool. It is also highly unlikely that there are spill-over effects between classes, because the experiment takes place at the class level, and in 7th grade, students do about everything in school at the class-level. They are in the same class during *all* classes they have at school.

Furthermore, it is important to note that there is no inherent interaction between remedial teaching during tutoring classes and the use of the digital tool. This may seem contradictory, but the reader should bear in mind that the digital tool is geared towards the new skills requirements of numeracy (see introduction) of *all* students, which are defined apart from the regular math study program in secondary school. Remedial teaching focuses on the regular math program and does not follow from signals of the digital tool. Therefore it is not surprising that comparison of data on the attendance of tutoring classes of this cohort with earlier cohorts that did not work with this digital numeracy tool, shows that the share of students who follow additional tutoring lessons for math, is similar over the years.

[11] Questionnaire and data files are available upon request from the corresponding author.

[12] Information on homework gathered via a student questionnaire in which a multiple choice question was used to ask them how much time they spend on homework for math, *excluding* the time they spend on “Mousework”.

Table 4
Baseline Results: *t*-test of the effect of the experiment on various outcome indicators.

Variable	*Control group (n = 107)*		*Treatment group (n = 230)*			
	Average	Std. Dev.	Average	Std. Dev.	*t*-statistic	p-value
Absolute test score T1	58.691	14.099	61.605	14.606	−1.722	0.08
Absolute test score T2	60.851	13.900	63.309	14.691	−0.966	0.79
Absolute growth in test score T0-T1	2.645	10.971	7.422	11.533	−3.594	0.00
Absolute growth in test score T1-T2	2.168	10.69	1.704	10.088	1.047	0.40
Absolute growth in test score T0-T2	6.813	11.78	9.126	11.873	−1.983	0.09
Difference in reference level between T0 and T1 (dummy indicator, 1 = changed ref level, 0 = same level)	0.009	0.366	0.100	0.358	−2.100	0.03

3. Empirical analysis and results

3.1. Methodology

To identify the intent-to-treat effect of access to the digital practice tool on test scores and growth in test scores we use the exogenous variation that is created by randomizing at the class level which classes have access to the tool and which do not. We observe a student i in class j's (with teacher t) (growth in) test score y_{ij} and the treatment, a students' access to the interactive digital practice tool, determined at the class level, d_{ij}. The linear regression, in which also student characteristics are included, is estimated as follows:

$$y_{ij} = \alpha_i + \tau_2 d_{ij} + \beta_i X_i + (\varepsilon_i + u_j), \quad (1)$$

where X_i are the students' observable characteristics, such as ability variables, age, gender and situation at home, which are independent of the treatment, ε_i are the residuals at the student level, u_j are the residuals at the class level and teacher level. Because of the randomization at class level, we cluster the standard errors at the class and teacher level throughout the analyses presented in this paper. However, the regular clustered standard errors procedure is based on a minimum of around 30 clusters, and because we have a relatively small amount of classes (13), it is likely that our results are biased if we use the regular clustered standard errors (Angrist & Pischke, 2009; Wooldridge, 2010). Therefore, as also suggested in Angrist and Pischke (2009) and in Wooldridge (2010), in the regressions we use the wild cluster bootstrap-t procedure, as described in Cameron, Gelbach, and Miller (2008) to bootstrap the clustered standard error of the coefficient of interest, the treatment variable. By doing this we correct for the low amount of clusters in our analysis.

3.2. Baseline results experiment stages I and II

The first results we present are the simple *t*-statistics of the effect of treatment on the three outcome measures from experiment stage I, and the outcome measures of stage II. In the remainder of Section 3 we focus on the results of experiment stage I, because this is the main experiment.

Table 4 presents the absolute score at T1 and T2, the growth in test scores between T0 and T1, between T1 and T2, and between T0 and T2, and the growth in reference level between T0 and T1, as well as the T-statistics of the simple binary independent sample comparison.[13] In Table 4, we see that the treatment group has a significantly higher absolute score at T1 and growth between T0 and T1 than the control group, and that this difference more or less remains in the second stage of the experiment. There are also significantly more students that, during the first stage, increased in reference level in the treatment group than in the control group. This implies that practicing at home with the individualized digital tool is beneficial, but not anymore once the individual nature of the tool is disabled, despite more practice minutes, as we saw before in Table 3.

3.3. The returns to digital practice – experiment stage I

The next step is to analyze the returns to practicing in the digital environment using regression analysis. The results of these analyses of the returns to practicing online are presented in Tables 5 and 6. Table 5 presents the result for outcome measure 'growth in test score between T0 and T1'. Table 6 presents the results for outcome measure 'growth in reference level'. In Table 5, we present 6 models: Model 1 gives the basic model in which no covariates are included, estimated by simple OLS, where we control for the clustering of students in classes by using clustering the standard errors at the class and teacher level, as explained above.[14] In Model 2, we add the work time used for the test, to account not only for accuracy but also for speed, because we believe that students that answer the same amount of questions correct in both T0 and T1 but do this twice

[13] Note that in Table 4 we have not corrected for intra-class correlation.
[14] Therefore the results are not exactly the same as in Table 4.

Table 5
Results Experiment stage I – The Returns to Digital Practice on Numeracy Score Growth between T0 and T1.

Dep. Var = Growth numeracy score between T0 and T1	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Treatment (individualized practice)	4.78	4.79	4.75	5.24	4.35	4.32
Bootstrapped clustered p-value	0.02 ***	0.02 **	0.02 **	0.00 ***	0.00 ***	0.00 ***
Covariates ($X_{j\ i}$)		Work time test	Work time test, score text comprehension	Work time test, score text comprehension, student ability, student characteristics	Work time test, score text comprehension, student ability, student characteristics, teacher attitude towards tool	Work time test, score text comprehension, student ability, student characteristics, teacher fe
Observations (n)	337	337	337	328	286	328
Clusters (classes)	13	13	13	13	11	13

significant at the 5% level; * significant at the 1% level.
Standard errors clustered at the class and teacher level. Note that we lose 2 classes when including the teacher questionnaire, because these teachers did not fill it out.

Table 6
Experiment stage I – The Returns to Digital Practice on Difference in Reference Level between T0 and T1.

Dep. Var = Difference in reference level between T0 and T1	Model 1	Model 2	Model 3	Model 4	Model 5
Treatment (individualized practice)	0.09	0.09	0.12	0.12	0.10
Bootstrapped clustered p-value	0.06	0.09	0.05 **	0.03 **	0.05 **
Covariates ($X_{j\ i}$)		score text comprehension	score text comprehension, student ability, student characteristics	score text comprehension, student ability, student characteristics, teacher attitude towards tool	score text comprehension, student ability, student characteristics, teacher fe
Observations (n)	337	337	328	286	328
Clusters (classes)	13	13	13	11	13

significant at the 5% level; * significant at the 1% level.
Standard errors clustered at the class and teacher level. Note that we lose 2 classes when including the teacher questionnaire, because these teachers did not fill it out.

as fast have more automated basic math skills, and are most likely already practicing at a higher level, not tested in the tests. In Model 3, we include the test score for the text comprehension test of T0, as the numeracy questions can be very linguistic, and students who have problems with text comprehension might also score lower for numeracy. In Model 4, we add variables that account for student ability and past education, such as the total score for the standardized exit exam of primary education and an indicator for dyslexia, and we add student specific characteristics such as age, gender, oldest child, religion, family type and primary school. Note that we lose some observations because not all students wrote the primary school ability test. In the fifth model we add the attitude of the math teacher towards the digital practice tool and the extent to which the math teacher uses the tool, for example for checking up on students. Note that two teachers did not fill out this questionnaire, leading to a lower amount of observations. Therefore, in model 6, those two teacher variables are replaced by teacher fixed effects, to see if the teacher influence is more than use and attitude of the tool. As such, models 5 and 6 combine measures of student ability with indicators of past education (e.g. primary school) and current education (e.g. class group, math teacher).

In Table 6 we have the same models as in Table 5, except that in Table 6 we do not correct for test time, which implies that we only have 5 models instead of 6, because model 2 of Table 5 is excluded. Obviously, work time is not taken into account in models 3–5 either.

The results presented in Table 5 show that the growth in score between T0 and T1 is around 4.8 absolute points higher for treatment students (scale 0 to 100), compared with control students, when we only control for class. The growth in score decreases slightly (to 4.3 points) in the fifth and sixth model, where we control for all the covariates including the teacher. The latter corresponds to a small to medium effect of 0.40 of a standard deviation, given the interpretation of Cohen's d (Cohen, 1988). Hence, the significance and magnitude of the effect proves robust to adding different types of student specific information. In other words, even when taking into account various student and teacher characteristics which may contribute

Table 7
Robustness Analyses Experiment stage 1, growth score between T0 and T1 – Model 6 from Table 5.

Robustness analyses	Smaller sample	Absolute score T1 as outcome measure	Only prevocational	Only higher general and pre university	Controlled for minutes practiced per week	Class random effects	Regular Arithmetic test in class
Treatment (individualized practice)	4.66	2.88	8.34	5.82	4.72	4.32	1.41
Bootstrapped clustered p-value	0.00 ***	0.06	0.22	0.02 **	0.01 ***	n/a	0.01 ***
Covariates ($X_{j\,i}$)	Work time test, score text comprehension, student ability, student characteristics, teacher fe	Work time test, score text comprehension, student ability, student characteristics, teacher fe	Work time test, score text comprehension, student ability, student characteristics, teacher fe	Work time test, score text comprehension, student ability, student characteristics, teacher fe	Work time test, score text comprehension, student ability, student characteristics, teacher fe	Work time test, score text comprehension, student ability, student characteristics, teacher fe	score_ tekstbegrip_t0, student ability, student characteristics, teacher fe
Observations (n)	286	328	116	212	328	328	325
Clusters (classes)	11	13	5	8	13	13	13

significant at the 5% level; * significant at the 1% level.
Standard errors clustered at the class and teacher level. Note that we lose 2 classes when including the teacher questionnaire, because these teachers did not fill it out.

to numeracy learning, the intervention is shown to add to numeracy performance. Detailed results of the regressions presented in Table 5 can be found in Appendix 1.[15]

The results in Table 6 show that the treatment (i.e. individualized practice) has a significant effect of around 0.10 on the growth in reference level. This implies that students in the treatment group are 10% more likely to increase in reference level than students in the control group. Therefore, practice with the digital tool does not only significantly increase the absolute growth in test score, but also significantly influences the growth in reference level. Given that the tool is mainly implemented to guarantee that all students end up at least at the reference level that is expected of them (given their track) in graduation year, we can conclude that the tool seems effective in doing that, already in the short time period of experiment stage I. Detailed results of the regressions presented in Table 6 can be found in Appendix 2.

4. Robustness analyses experiment stage I

4.1. Robustness analyses growth in test score

Table 7 presents the robustness analyses for our main analysis, namely model 6 of Table 5, where growth in score between T0 and T1 is the outcome measure. The robustness analyses are presented in a random order. The first robustness analysis is done using a smaller sample, where we exclude the two classes of the 'higher general' type from the treatment group, as these classes are not represented in the control group. We do so to check whether the grouping of students in class types influences their behavior, even though the personal characteristics of the students in the original control and intervention groups were comparable as we already showed in Table 2 After exclusion of the 'higher general' classes, the sample only consists of prevocational classes and mixed higher general/pre-university classes. The second robustness check is done by using the absolute score at T1 as outcome measure, instead of the growth between T0 and T1. The third and fourth robustness analyses are the separate analyses for prevocational students (column 3) and higher general/pre-university students (column 4).[16] In the fifth robustness check we include minutes practiced per week, to show that it does not make a difference whether we include the minutes practiced per week or not. Next, we include class random effects instead of clustered standard errors to control for the exogenous variation at the class and teacher level. Lastly, we analyze if the results still hold if we use the regular numeracy test that was written in class. This is indeed the case, although it is important to mention that the numeracy classes and the test written in class are not necessarily directly related to what is being practiced in the tool. Especially because all

[15] As practice behavior differed largely among students, we checked whether the effect was driven by the group of students that had practiced the most during this period (on average more than 10 min per week), vs. the group that practiced the least (less than 5 min per week) and the group in between. The results of the analysis of these separate groups show that this is not the case, independently of the group with respect to minutes practiced, experiment group students perform significantly better than control group students. Furthermore, we analyze details of the practice behavior as potential mechanisms in Section 5.2.

[16] Separate *t*-statistics/Chi-squared statistics for pre-vocational students and for higher general/pre-university students show that treatment and control group are still not significantly different on observable characteristics after splitting the sample into these two groups.

Table 8
Robustness Analyses Experiment stage I, growth in reference level – Model 5 from Table VI.

Robustness analyses	Smaller sample	Only higher general and pre university	Controlled for minutes practiced per week	Class random effects
Treatment (individualized practice)	0.12	0.14	0.10	0.10
Bootstrapped clustered p-value	0.02 **	0.08 *	0.05 **	n/a
Covariates ($X_{j\ i}$)	score text comprehension, student ability, student characteristics, teacher fe	score text comprehension, student ability, student characteristics, teacher fe	score text comprehension, student ability, student characteristics, teacher fe	score text comprehension, student ability, student characteristics, teacher fe
Observations (*n*)	286	212	328	328
Clusters (classes)	11	8	13	13

Note that only 4 prevocational students had a change in their reference level, which makes the regression of model 5 impossible to run.
significant at the 5% level; * significant at the 1% level.
Standard errors clustered at the class and teacher level. Note that we lose 2 classes when including the teacher questionnaire, because these teachers did not fill it out.

students have an individually adapted practice program in the tool and might practice with all domains deepening on their starting level, whereas the test only covers the topics that were discussed in the weeks before the test. Therefore, it is unclear what exactly is driving this result and we do not want to put too much weight on this finding, although it does give confidence for the generalizability of the results of our study.

Table 7 shows that the analysis is very robust to using different samples, other outcome measures, using random effects instead of clustered standard errors and the inclusion of minutes practiced per week, with respect to significance levels, and the magnitude of the coefficient for the relevant columns. The magnitude of the coefficient is different for the second column, where we use a different outcome measure, and for the separate analyses for pre-vocational and higher general/pre-university students. Both the two separate analyses give higher coefficients than the combined analysis, showing that the effect of individualized practice works differently for the two groups and is differently influenced by the inclusion of student characteristics.

4.2. Robustness analyses difference in reference level

Table 8 presents the robustness analyses for model 5 of Table 6, where growth in reference level is the outcome measure. Table 8 contains robustness analyses on the smaller sample, the subgroup higher general/pre-university students, inclusion of minutes practiced per week and using class random effects instead of clustered standard errors. Again, the results are very similar to the results in Table 6, both in magnitude and in significance.

5. Additional evidence

5.1. The returns to non-individualized digital practice (experiment stage II)

Table 9 presents the results from the second stage of the experiment, in which the same four classes were excluded from practicing with Mousework as in the first experiment. However, in this stage of the experiment the exercises available in Mousework for the treatment group students were not individualized. Instead, all exercises were available for practice for all students. This setup requires a higher responsibility by the treatment students, as they do not only have to practice, but also select the exercises that are relevant for them to practice with.

The results in Table 9 show that the treated students have experienced less increase in their test score than control group students. However, this finding is not or hardly significant. This implies that there is no effect of non-individualized digital practice, despite the fact that on average students practiced more in period II than they did in period I. Taking together experiment stages I and II we conclude that practicing with a digital practice tool is only effective when there is an individual learning route for the student. The latter suggests that the 7th grade students lack the skills or motivation to select the most effective exercises in the digital environment. Unfortunately, this cannot be verified with the data, as there is only data available from the practiced exercises, and not from the available exercises for each student. Detailed results of the regressions presented in Table 9 can be found in Appendix 3.

Table 9
The Returns to Practicing Online on Numeracy Score Growth between T1 and T2.

Dep. Var = Growth numeracy score between T1 and T2	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Treatment (Non-individualized practice)	−2.46	−2.23	−2.27	−2.25	−2.32	−2.47	−2.15
Bootstrapped clustered p-value	0.08	0.11	0.11	0.15	0.16	0.21	0.16
Covariates ($X_{j\,i}$)		Average minutes practiced per week	Average minutes practiced per week, work time test	Average minutes practiced per week, work time test, score text comprehension	Average minutes practiced per week, work time test, score text comprehension, student ability, student characteristics	Average minutes practiced per week, work time test, score text comprehension, student ability, student characteristics, teacher attitude towards tool	Average minutes practiced per week, work time test, score text comprehension, student ability, student characteristics, teacher fe
Observations (n)	337	337	337	337	328	286	328
Clusters (classes)	13	13	13	13	13	11	13

significant at the 5% level; * significant at the 1% level.
Standard errors clustered at the class and teacher level. Note that we lose 2 classes when including the teacher questionnaire, because these teachers did not fill it out.

5.2. Students' practice behavior

The literature suggests that difference in practice behavior might account for differences in the outcome measure (e.g. Biagi & Loi, 2013). Although practice behavior is never completely exogenous, it is still interesting to explore these correlations. Bartelet, Ghysels, Groot, Haelermans, and Maassen van den Brink (2016), who use the same dataset, but focus on different groups of students and different outcome measures, have already extensively discussed the five different practice indicators that are collected from the Mousework program and the potential mechanism that lies in each indicator. The tool registers (a) how many seconds the student has worked in the tool, (b) how many exercises/tests a student has completed, (c) what the score was on the exercises/tests, (d) how many explanation screens the student asked for during the exercise/test (note that during the test, explanation screens were not on content but on the instruction of the test), and (e) the time wasted, which counts the time that the student is logged onto the tool but no activity takes place, with the counting starting after 3 min of nonactivity by the student.

Similar to Bartelet et al. (2016) we look into the potential mechanisms of the found effect by analyzing the correlations of the outcome measure with four indicators of practice behavior, namely whether a student has practiced at all, how many minutes the student has practiced, the number of explanation screens per minute and the number of exercises per minute.[17] We estimate a regression in which we look at the increase in test score as outcome measure, and explain differences in the outcome with the practice indicators. We also control for student characteristics and teacher dummy variables. Table 10 shows the results of this analysis. The indicator whether a student has practiced at all make the largest difference, though not at the expense of speed, as the number of exercises per 10 min is negatively significant. The latter implies that students who click through the exercises too fast do not seem to learn too much.

Another interesting aspect is that the number of minutes practiced does not have any additional explanatory power once accounted for whether a student has practiced at all (which represents the average number of minutes practiced). Of course, once the practice dummy is taken out of the analysis, the number of minutes becomes positive and significant (not visible in Table 10). In our data, it also appears that there is not a quadratic relationship of the number of minutes practiced and the outcome, most likely due to the limited range of minutes practiced, combined with a relatively low average.

The mechanisms that are at play when practicing with the tool are therefore not necessarily as obvious as one would expect, because all types of practice behavior are related, but certain (practice) behavior is negatively related with the outcome (practicing too fast), whereas other behavior is positively related (practicing at all).

6. Cost effectiveness

Determining the cost effectiveness of this digital practice tool can be done from the schools' point of view, but also from the society's point of view. The costs of the digital practice tool for numeracy are approximately 25 euros per student per year.

[17] Note that we also do not include time wasted, as it is unclear how this indicator would influence performance.

Table 10
Correlates of growth in math skills level and practice behavior.

Dep. Var = Growth numeracy score between T0 and T1	Coefficient	St. error.	p-value
Practiced dummy	7.62	3.78	0.045
Minutes practiced	−0.00	0.02	0.839
Explanation screens	−0.14	0.25	0.584
Nr of exercises per 10 min	−1.87	0.97	0.054
Covariates (Xji)	student ability; student characteristics; teacher dummies		
Observations(N)	336		
Groups (classes)	13		

Standard errors clustered at the class and teacher level.

The total costs for this tool for the group of 337 first year students for the school is around 9000 euros. For the school, an alternative measure to foster numeracy skills, could be the introduction of an additional math class (for the school under study, this was the alternative they considered). The additional costs of hiring a teacher who practices numeracy skills with the student for at least an hour a week would be a lot higher than using this digital tool. Given that there are 13 first year classes, one would need an additional full time teacher each year, which will bring about costs of at least 30 000 euros per year.

With respect to gains for society, we see that in the test at T0, about 85 percent of the first year students in our dataset performed lower on numeracy than they are supposed to, according to the national reference levels (Commissie Meijerink, 2008), having finished primary education successfully. The results of experiment stage I already show that in the treatment group significantly more students increase in reference level during the experiment. If practicing with the digital tool would increase the average numeracy level such that the majority of these students would perform at the expected level in their graduation year, the societal cost saving could be very large. Each student that does not fail the national exam at the end of secondary education because they fail their numeracy exam, saves the government 7000 euros, which is the average cost per student per year (Teule, 2012). Furthermore, retention in grade will delay the student for at least one year in entering either vocational education or higher education, which in turn delays labor market entry by at least one year. The opportunity costs of the student will therefore be a lot higher than the 7000 euros for the government. In any case, to be cost effective from society's point of view, the introduction of "Mousework" in the starting year of secondary education only needs to allow two students to graduate in time instead of delaying their graduation with one year. Given that it can be expected to help 35 students across the threshold (10% of 355), the latter seems highly likely, although future research following-up on students throughout secondary education should confirm that expectation before any solid statements of this type can be made.

7. Conclusions and discussion

7.1. Conclusions

The aim of this paper was to evaluate various ways to foster a crucial aspect of human capital, basic math skills (numeracy) and, more specifically, to look into the promise of computer assistance. We analyze a two-stage experiment on the effect of an individualized interactive digital practice tool on numeracy skills for 7th grade students. The results from the first stage of the experiment show that there is a significant effect of (being able to) practice individualized with the digital tool on both the absolute numeracy score and the growth in numeracy score, as well as on the share of students increasing in reference level for numeracy. The effect on the growth in score is 4.8 points, which corresponds to a small to medium effect of 0.40 of a standard deviation. Given the increase in math homework time, due to the tool, of 25–33 per cent, this is not surprising. However, this effect might not solely be due to practicing, but could partly be caused by the small tests that were written 3 or 4 times during at school, during experiment stage I (relates to the testing effect, where usually large effects are found, see Roediger and Karpicke (2006)), and possibly also to the increased teacher attention for the treatment classes. In comparison with benchmarks on average gains in effect size, which is around 0.32 for math (Hill, Bloom, Black, & Lipsey, 2007), the effect we find is quite comparable. Some caution is needed with these results, as the randomization of the experiment took place at the class level, and although we do control for that in the analyses, the number of classes is not high enough to be absolutely sure that these effects are causal.

This effect is robust to adding different types of student specific information that may influence practice behavior and incorporating information on the use of the tool and attitude towards the tool of the teacher. Furthermore, several robustness checks confirm this finding.

In the second stage of the experiment, the treatment group did not have an individualized learning path in the Mousework tool, but instead had access to *all* existing numeracy exercises in Mousework. The results from the second experiment show no effects, although students practiced on average *more* minutes per week during this period, implying that the effectiveness of the tool found in experiment stage I, is mainly due to its individualized and personal nature (similar to the finding in Barrow et al., 2009).

Furthermore, analysis of the practice behaviour reveals that the mechanisms that are at play when practicing with the tool are not necessarily as obvious as one would expect. All types of practice behavior are related, but certain (practice) behavior is negatively related with the outcome (practicing too fast), whereas other behavior is positively related (practicing at all).

Lastly, a cost benefit analysis shows large potential gains, since this tool might prevent many students from failing their national graduation exams and consequently having to repeat the last grade.

The results from the experiment and the cost benefit analysis lead to a dual conclusion: 1) Individualized digital practice tools are (cost)effective to improve numeracy performance of secondary students, and 2) individualization of exercises makes digital practice tools effective.

7.2. Discussion

A potential problem with respect to the generalization of the results is that the experiment was conducted at only one school in the Netherlands with no disadvantaged students. However, test results of numeracy tests at all other schools in the Netherlands that use the same digital practice tool with an individualized path (where no experiment was conducted) show a similar increasing trend in results after students have practiced with the tool, which suggests that the results may not be dependent on this one school where we conducted the experiment. It should be noted though, that, although the literature shows that the previously found effects of these types of practice programs are not limited to economically and educationally privileged students, we cannot be certain that these results can be generalized to disadvantaged students as well.

The data on practice behaviour shows that the average amount of minutes practiced per week does not differ much between the two periods (and if anything, is lower in the first period), which indicates that the effect found in the first stage of the experiment is not driven by student motivation caused by the fact that they were part of an experiment. Furthermore, the student questionnaire shows that students do not like to practice with the tool. Informal conversations with students reveal that initial control group students were actually *happy* that they were selected as control group. These two observations indicate that we are measuring the effect of digital practice and that we do not observe the so-called Hawthorne effect.

The innovation evaluated in this paper offers a dual conclusion regarding the role of teachers. First, using a digital practice tool that is independent of teaching in class gives positive outcomes, even if the teachers have mixed feelings about its use (results from experiment stage I). Second, the additional analyses also illustrate, however, that teachers with a positive attitude towards the tool tend to contribute to the success of the tool itself, by fostering the intensity of its use (results from all parts of the experiment). It is therefore important to manage the implementation of the tool in such a way that teachers feel involved and want to do their best to make the use of the tool a success. Although it also works without a positive teacher attitude, it works *better* when teachers are positively involved.

All-in-all, the results of this experiment offer promising perspectives for policy makers wanting to increase numeracy among students. We show that a digital practice tool which is used by students at home without a close link to the math teaching at school can be an effective and efficient way to improve numeracy skills, provided the tool is individualized. Lastly, barriers to implementation seem relatively low, because students use it at home and the tool does not require a large learning effort by teachers, because they do not need to adapt their teaching to the tool.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compedu.2017.05.010.

References

Angrist, J. D., & Lavy, V. (2002). New evidence on classroom computers and pupil learning. *Economic Journal, 112*, 735–765.
Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics - an Empiricist's companion* (1st ed.). Princeton, NJ: Princeton University Press.
Arroyo, I., Park Woolf, B., Royer, J. M., Tai, M., & English, S. (2010). Improving math learning through intelligent tutoring and basic skills training. In V. Aleven, J. Kay, & J. Mostow (Eds.), *ITS 2010, Part i, LNCS 6094* (pp. 423–432). Berlin Heidelberg: Springer-Verlag.
Banerjee, A. V., Cole, S., Dufflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics, 122*(3), 1235–1264.
Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's Edge: The Educational Benefits of Computer-Aided Instruction. *American Economic Journal: Economic Policy, 1*(1), 52–74.
Bartelet, D., Ghysels, J., Groot, W., Haelermans, C., & Maassen van den Brink, H. (2016). The differential effect of basic mathematics skills homework via a web-based intelligent tutoring system across achievement subgroups and mathematics domains: A randomized field experiment. *Journal of Educational Psychology, 108*(1), 1–20.
Biagi, F., & Loi, M. (2013). Measuring ICT use and learning outcomes: Evidence from recent econometric studies. *European Journal of Education, 48*(1), 28–42.
Borman, G. D., Benson, J. G., & Overman, L. (2008). A randomized field trial of the fast ForWord language computer-based training program. *Educational Evaluation and Policy Analysis, 31*(1), 82–106.
Burns, M. K., Kanive, R., & DeGrande, M. (2012). Effect of a computer-delivered math fact intervention as a supplemental intervention for math in third and fourth grades. *Remedial and Special Education, 33*, 3.
Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Boostrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics, 90*(3), 414–427.
Cheung, A. C. K., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*(3), 198–215.
Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88–113.
Chiswick, B. R., Lee, L. Y., & Miller, P. W. (2003). Schooling, literacy, numeracy and labour market success. *The Economic Record, 79*(245), 165–181.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Commissie Meijerink. (2008). *Over de drempels met taal en rekenen - hoofdrapport van de expertgroep doorlopende leerlijnen taal en rekenen Enschede*.
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.
Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4 ed.). London: Sage.
Funnekotter, B., (2012, 11 juni 2012). Middelbare scholieren scoren massaal onvoldoende bij rekentoets. NRC.
Goolsbee, A., & Guryan, J. (2006). The impact of internet subsidies in public schools. *Review of Economics and Statistics, 88*(2), 336–347.
Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessman, L. (2013). *Returns to skills around the world: Evidence from PIAAC*. NBER Working Paper 19762. Cambridge (MA): NBER.
Hattie, J. (2009). *Visible learning - a synthesis of over 800 meta-analysis relating to achievement*. New York, NY: Routledge.
Heinrich, C. J., Burch, P., Good, A., Acosta, R., Cheng, H., Dillender, M., … Stewart, M. (2014). Improving the implementation and effectiveness of out-of-school-time tutoring. *Journal of Policy Analysis and Management, 33*(2), 471–494.
Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*.
KNAW. (2009). *Rekenonderwijs op de basis school*. Advies en sleutels tot verbetering. KNAW Amsterdam.
Kulik, C. L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*, 75–94.
Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economic and Statistics, 89*(4), 721–736.
Machin, S., McNally, S., & Silva, O. (2007). New technologies in schools: Is there a payoff? *The Economic Journal, 117*, 1145–1167.
Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Center for Technology in Learning Washington D.C.
Muiswerk. (2013). Retrieved from www.muiswerk.nl.
OECD. (2013). *Skills outlook 2013: First results from the survey of adult skills*. OECD Publishing Paris.
Pilli, O., & Aksu, M. (2013). The effects of computer-assisted instruction on the achievement, attitudes and retention of fourth grade mathematics students in North Cyprus. *Computers and Education, 61*, 62–71.
Roediger, H., & Karpicke, J. D. (2006). The power of testing memory - basic research and implications for educational practice. *Perspectives on Psychological Scienc, 1*(3), 181–210.
Rouse, C. E., & Krueger, A. B. (2004). Putting computerized instruction to the test: A randomized evaluation of a 'Scientifically-Based' reading program. *Economics of Education Review, 23*(4), 323–338.
Schijf, G. M., & Schijf, T. (2014). *Handleiding online Testprogramma's*. Hoorn: Muiswerk Educatief.
Teule, P., (2012, 25 July 2012). Wat kost zittenblijven nou echt?. Retrieved from: http://sargasso.nl/wat-kost-zittenblijven-nou-echt/.
Tomlinson, C. A. (2004). Research evidence for differentiation. *School Administrator, 61*(7), 30.
Tomlinson, C. A., & Kalbfleisch, M. L. (1998). Teach me, teach my brain: A call for differentiated classrooms. *Educational Leadership, 56*(3), 52–55.
Vignoles, A., De Coulon, A., & Mercenaro-Gutierrez, O. (2011). The value of basic skills in the British labour market. *Oxford Economic Papers, 63*, 27–48.
Wooldridge, J. (2010). *Econometric analysis of cross-section and panel data* (2nd ed.). Cambridge, MA: MIT Press.