The effect of adaptive versus static practicing on student learning - evidence from a randomized field experiment

van Klaveren, Chris; Vonk, Sebastiaan; Cornelisz, Ilja

published in
Economics of Education Review
2017

DOI (link to publisher)
10.1016/j.econedurev.2017.04.003

document version
Publisher's PDF, also known as Version of record

Economics of Education Review 58 (2017) 175–187

Contents lists available at ScienceDirect

Economics of Education Review

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

The effect of adaptive versus static practicing on student learning - evidence from a randomized field experiment

Chris van Klaveren[a,1,*], Sebastiaan Vonk[b], Ilja Cornelisz[a,1]

[a] Faculty of Behavioral and Movement Sciences at Vrije Universiteit Amsterdam and the Amsterdam Center for Learning Analytics, Netherlands
[b] Amsterdam School of Economics at University of Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 1 July 2016
Revised 31 March 2017
Accepted 11 April 2017
Available online 18 April 2017

JEL classification:
A20
I21
L86

Keywords:
Personalized education
Adaptive practice software
Field experiment

ABSTRACT

Schools and governments are increasingly investing in adaptive practice software. To date, the evidence whether adaptivity improves learning outcomes is limited and mixed. A large-scale randomized control trial is conducted in Dutch secondary schools to evaluate the effectiveness of an adaptive practice program relative to a static program. Learning theories predict that adaptive practicing is more effective, but this experimental evaluation provides a more nuanced picture. Relative to the static software environment, students working in the adaptive software environment receive more difficult exercises, practice longer and answer fewer questions correctly. Takeup and usage of the software program is, overall, modest, but varies considerably within and between classrooms. The outcome differences between both environments are more pronounced in classrooms with higher practice intensity. On average, no test score effects are found, but static practicing does improve test scores for higher ability students (0.08σ). Caution is thus warranted when adaptive practice software is implemented to address individual learning needs, as static formative test preparation can be more effective in improving test scores.

1. Introduction

Educators and educational policy makers have long argued that the educational needs of students can be better accommodated when the educational process is more personalized (Lou et al., 1996; Miliband, 2006; Reezigt, Houtveen, & Van de Grift, 2001). Schools endorse this point of view, but point out that teachers cannot possibly develop fully personalized programs; due to both a lack of time and knowledge (Coubergs, Struyven, Gheyssens, & Engels, 2015). Therefore, computerized adaptive practicing is considered a viable alternative for offering personalized education and huge increases are observed in the percentage of teachers who use computers to offer education (OECD, 2015).

Schools and governments invest heavily in adaptive practice software, but the lack of solid evidence that this improves learning outcomes is worrisome (Bulman & Fairlie, 2015; Slavin, 2002; 2004). Generally, the didactical and technical foundations underlying these software programs are not explicitly presented, such that it remains unclear why these products should lead to improved learning outcomes in the first place (OECD, 2015). In order to structurally improve adaptive practice software a solid didactical and technical underpinning should be provided and empirically evaluated.

This study evaluates the relative effectiveness of two computerized practice environments that can potentially improve learning outcomes. The learning outcomes considered are summative test scores, practice time, the number of completed exercises, and the number of correct answers given. Both practice environments rely on well-known theories of learning and motivation (e.g. Bloom's Taxonomy (Bloom, of College, & Examiners, 1956), self-determination theory (Deci & Ryan, 1985), effective feedback (Hattie & Timperley, 2007)), but each has separate strengths. The strength of the adaptive environment is that the practice process is tailored around previous performance. The adaptive process relies on mastery learning (i.e. students receive exercises of higher knowledge types only when mastery has been demonstrated), and the zone of proximal development (i.e. offered exercises should not be too difficult) (Tomlinson et al., 2003). Thereby, the adaptive environment has the objective to personalize the practice process, such that it better accommodates the individual educational needs of students. The strength of the static practice environment is that, while practicing, students are essentially offered formative tests that are valid and representative with respect to the upcoming summative test. As such, students do not receive a 'tailor-made' practice process, and exercises can be considered too easy or too

* Corresponding author.
E-mail addresses: c.p.b.j.van.klaveren@vu.nl (C. van Klaveren), s.j.j.vonk@uva.nl (S. Vonk), i.cornelisz@vu.nl (I. Cornelisz).
[1] (*ACLA*, acla.amsterdam).

http://dx.doi.org/10.1016/j.econedurev.2017.04.003
0272-7757/

difficult. But, by providing an equal amount of questions across the entire spectrum of topics, knowledge types and difficulty levels, students may be effectively prepared for the test. Effective adaptive practicing requires that the assumptions underlying the adaptation process are correct when tailoring the process, as students otherwise run the risk of being exposed to questions that do not effectively accommodate their individual educational needs.

The contributions of this study to the current evaluation literature on adaptive software are the following. First of all, the algorithms that underlie both conditions are outlined formally and in detail. In this way, we ensure that both conditions are not a black box, as is generally the case with current practice software programs (one positive exception is Klinkenberg, Straatemeier, & Van der Maas, 2011). Secondly, both algorithms are directly linked to theories of learning and motivation, such that the educational mechanisms underlying the mathematical processes can be interpreted. Finally, a randomized field experiment is conducted, for the duration of one school year, such that the relative effectiveness of the adaptive practice environments is rigorously evaluated. In total, 1021 children in Dutch secondary schools are, within classes, randomly assigned to one of the two practice environments. Importantly, and empirically evaluated in this study, students were unaware of the practice environment they were assigned to. By conducting a large-scale randomized field experiment, this study addresses the lack of systematic evidence (i.e. based on randomized evaluations and large-enough samples) regarding the effectiveness of interventions aimed towards computerized (adaptive) learning (e.g. see West, 2011).

This paper proceeds as follows. In Section 2, an overview of the recent international empirical literature on the effectiveness of digital learning is provided, together with its implications for personalized learning. Section 3 provides a technical explanation of the two different practicing algorithms. This is followed by a discussion of the experimental design in Section 4. Descriptive statistics on the practicing process of both versions are presented in Section 5, and the empirical findings on the algorithms' relative effectiveness are presented in Section 6. The paper concludes by providing an analysis of student experiences during practicing in Section 7 and a discussion of the findings in Section 8.

2. Empirical results on the effectiveness of ICT and personalized learning

Ever since the 1960s, computer-based instruction programs have been developed to augment schooling and improve learning outcomes. Early meta-analyses conclude that these programs hold the potential to increase test scores, on average by 0.3 standard deviations, but also that results vary depending on context, implementation, length of the program and features of the outcome test (Kulik & Kulik, 1991, 1987). A more recent meta-analysis by Cheung and Slavin (2013) on the effectiveness of educational technology applications point to modest, but positive, results when compared to traditional methods. Importantly, the results tend to vary by technology type, with the largest effects found for supplemental computer-aided learning.[2] However, the results of these meta-analyses should be treated with caution, as the estimated effects of supplemental computer-aided learning may (partly) be driven by unobserved factors, such as behavioral effects (e.g. novelty, Hawthorne), selective teacher assignment and publication bias.

In recent years, some (quasi-) experimental evaluations have been conducted to gain more insights into the effects of (i) the availability of computers, (ii) the level of ICT expenditure, (iii) computer-aided learning, and (iv) specific educational software products. The results of these studies are presented in Table 1. Column two of this table refers to the type of intervention, column three indicates the subjects for which effects were evaluated, Columns four and five refer to, respectively, the targeted student population and the country-context, and column six summarizes the empirical results. With respect to the latter column, it thus implies that one single minus sign indicates that only negative effects were found, while +0− indicates that positive, negative and non-significant effects were found. This points out that the empirical findings are rather ambiguous and that no general conclusions can be drawn with respect to the effectiveness of ICT in education.

Studies focusing on the effectiveness of CAL and educational software are most relevant for this study, since these applications are specifically designed to make learning adaptive. The results of these ICT applications are also ambiguous, but differ from the results of studies focusing on the effectiveness of computers and ICT funding. In particular, the effects are either positive or not statistically significantly different from zero, whereas studies focusing on ICT funding and computers tend to find more negative results. Studies with negative results argue that, at least in the short-run, this might be due to disruption and implementation issues (Angrist & Lavy, 2002; Campuzano, Dynarski, Agodini, & Rall, 2009; Dynarski et al., 2007). Leuven, Lindahl, Oosterbeek, and Webbink (2007) and Barrow, Markman, and Rouse (2009) emphasize that replacement of instruction time by ICT-related activities can have negative, as well as positive effects, depending on the relative quality of instruction and the ICT application evaluated. Moreover, the results indicate that introducing ICT in already highly developed educational settings seem to yield only weakly positive effects (Machin, McNally, & Silva, 2007). More promising results are found in developing contexts, when relatively poorly performing students are targeted and when expanding the usage of ICT is aimed at augmenting the existing curriculum (Banerjee, Cole, Duflo, & Linden, 2007; Rouse & Krueger, 2004). This is corroborated by a recent experimental evaluation of adaptive software in India, which also finds relatively large positive effects for students in secondary education that are significantly behind their grade-appropriate standard (Muralidharan, Singh, Ganimian et al. (2016)).

3. Static and adaptive practice algorithms

The total set of exercises is defined as N where $e^{t,k,d} \subset N$ with labels t, k and d. The subset-labels refer to the topic, t (with $t \in \{1,\ldots,T\}$), the knowledge type level, k (with $k \in \{1,\ldots,K\}$) and the level of difficulty, d (with $d \in \{1,\ldots,D\}$). It follows that there are $T \times K \times D$ subsets and that $e_n^{t,k,d}$ refers to exercise n in subset $\{t, k, d\}$.

Static practice algorithm

The static practice environment distinguishes between three knowledge-type and difficulty levels, such that each topic t encompasses nine subsets (i.e. $K \times D$ subsets). This is graphically illustrated in Fig. 1. The three knowledge type levels refer to different levels of the cognitive domain (1 = replication, 2 = application, and 3 = insight). The three difficulty levels are defined based on the responses of students of current and earlier cohorts (i.e. based on several million student answers). Exercises are labeled *easy* when the proportion of accurate responses is among the 33.33% *best* answered exercises. In a similar fashion, exercises are labeled *medium* (*hard*) if the proportion of accurate responses is between the 33.33 and 66.67% *best* answered exercises (is among the 33.33% *worst* answered exercises). $h^{t,k,d}$ in the figure indicates the number of exercises in subset $\{t, k, d\}$ and the probability of

[2] We note that Computer aided instruction (CAI), computer aided learning (CAL), and E-learning are used interchangeably in the economics and education literature.

Table 1
Findings on the effectiveness of ICT in education.

Study	Intervention	Subject(s)	Students	Context	Results
Angrist and Lavy (2002)	Computer Availability	Languages/Math	4th graders	Israel	–
Machin et al. (2007)	ICT funding	Language/Science/Math	Age 11	Great Britain	+ 0
Leuven et al. (2007)	ICT funding	Grade 8 subjects	8th graders	Netherlands	–
Rouse and Krueger (2004)	Computer-Aided Learning	Language	3rd–6th graders	United States	0
Barrow et al. (2009)	Computer-Aided Learning	Math	7th–12th graders	United States	+
Banerjee et al. (2007)	Computer-Aided Learning	Math	4th graders	India	+ +
Dynarski et al. (2007)	Software	Reading/Math	1st–12th graders	United States	0
Campuzano et al. (2009)	Software	Reading/Math	1st–12th graders	United States	+ 0 –

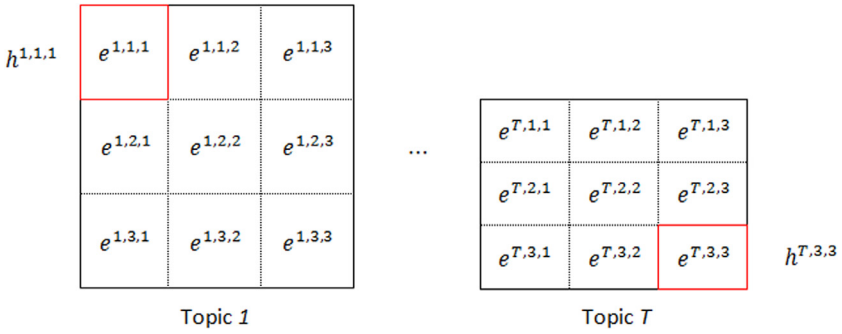

Fig. 1. Exercise subsets.

drawing an exercise from this subset is dependent on the amount of content (exercises) available within this exercise subset.[3]

The static algorithm generates sequences of exercises, similar to the traditional workbook. For each student, *one* exercise is randomly selected for each subset {*t, k, d*}, beginning with $e^{1,1,1}$ and ending with $e^{T,3,3}$. The selected exercises are ordered within each topic by knowledge type and difficulty level and presented one-by-one and in this specific order to students. Randomly drawing exercises ensures that students receive unique practice sessions, and this prevents them to learn to which environment they were assigned. The probability that an exercise is offered to a student depends solely on the available content, which agains emphasizing the importance of having sufficient content. The probability of offering an exercise from subset {*t, k, d*} equals:

$$\Pr(e_n|t,d,k) = \frac{1}{h^{t,d,k}}. \tag{1}$$

We note that the static algorithm does not take into account previous practice performance, but in the process of returning exercises it does take into account difficulty and knowledge type levels. As a result, formative tests are offered to students which are representative and valid for the upcoming summative test.

Adaptive practice algorithm

The adaptive algorithm has the objective to provide students with additional practice material on a certain topic if their understanding level is poor relative to the average understanding level of the class. The algorithm considers relative understanding levels, since absolute levels may reflect the general difficulty of a particular topic, rather than a below-average understanding for a particular student. The understanding level of student i and class c on topic t is determined by the proportion of correct answers of the total answers given. If P denotes the proportion of correct answers given on topic t, then the relative understanding level of topic t for student i is:

$$\frac{P_{it}}{P_{ct}}. \tag{2}$$

Higher values of $\frac{P_{it}}{P_{ct}}$ thus refer to better understanding levels for student i. The probability of receiving a particular topic is inversely related to this relative understanding level. The adaptive module updates the relative understanding levels after each exercise, such that the probability that student i receives an exercise on topic t equals:

$$\Pr(t) = \left(\frac{(\frac{P_{it}}{P_{ct}})}{\sum_{t=1}^{T}(\frac{P_{it}}{P_{ct}})} \right)^{-1}. \tag{3}$$

When the algorithm has selected topic t, parameters μ and p are defined. Parameter μ represents a topic-specific threshold level which determines the knowledge type subset k from which an exercise is offered. The dynamic practice environment considers the same knowledge-type levels as the static practice environment. The threshold level μ is set at 50% , meaning that knowledge type k exercises are offered as long as $\frac{P_{it}}{P_{ct}}$ does not exceed 0.5. The algorithm switches to exercises of knowledge type $k+1$ if and only if $\frac{P_{it}}{P_{ct}} > 0.5$. Let p_{tk} indicate the topic- and knowledge type-specific understanding level. The algorithm then returns an exercise from the subset t, k for which holds that the understanding level of the average student (i.e. $\frac{1}{I}\sum_{i=1}^{I} P_{itk}$) is more difficult and closest to p_{tk}. The understanding level of the average student is approximated by the proportion of accurate answers given by students of current and earlier cohorts (i.e. based on several million student answers). It follows that the algorithm dynamically selects a topic, conditionally on individual levels of relative understanding, whereas the knowledge type and difficulty level of the exercises are generated deterministically.

Practice Environments and Learning Theories

Both the static and adaptive practice procedures are founded on theories of learning. The learning theories underlying the practice

[3] This field experiment is made possible with support of Dedact, a private-entity (Levitt & List, 2009) involved in education. Dedact is a major publisher in the Netherlands, such that the amount of content that could be offered to students was sufficient.

environments are self-determination theory (SDT), Bloom' s taxonomy, Mastery Learning, the zone of proximal development and theories of feedback.

The theory of self-determination theory (SDT) argues that intrinsic motivation is more effective in promoting learning than extrinsic motivation, even though both are considered important in determining a learner's behavior (Deci & Ryan, 1985; Deci, Vallerand, Pelletier, & Ryan, 1991). Since practice is voluntary, SDT predicts that students practice more when the process is experienced as more enjoyable and/or useful. Practice intensity is thus expected to be higher in the practice environment that students' experience as most enjoyable.

Bloom's taxonomy (Bloom et al., 1956) defines different stages of learning, and we note that many adaptations and revisions have since been conjectured (see: Krathwohl, 2002). Both algorithms incorporate Bloom's taxonomy when returning exercises to students and additionally that exercises within each knowledge-type domain can have different difficulty levels. The static algorithm assumes three difficulty levels (easy, medium, hard), as outlined above, and an exercise is randomly drawn for each knowledge type-difficulty combination for each topic. The exercises selected are then ordered within topic by knowledge type and difficulty level (also in that order), and the student will progress through nine exercises on a given topic before moving on to the next topic. Students assigned to the static practice environment receive formative tests which are valid and representative. As such, at least in theory, students are effectively prepared for the upcoming summative test.

The adaptive algorithm adapts the practice process by considering the readiness level of a student in terms of knowledge and difficulty levels (Tomlinson, 2001). Readiness with respect to knowledge type is ensured by imposing mastery learning in addition to Bloom' s taxonomy. Mastery learning has been found to positively affect student achievement (Kulik, Kulik, & Bangert-Drowns, 1990; Slavin, 1987). Regarding the adaptive practice process, mastery learning implies that students must first demonstrate mastery in remembering content (replication), then must demonstrate mastery in actively using knowledge (application), and only then move on to exercises assessing whether students comprehensively understand a particular topic (insight).

The adaptive algorithm returns exercises on a particular topic with a higher probability when the understanding-level of the student is poor relative to the average understanding level of the class. Readiness with respect to exercise difficulty is ensured by offering exercises in the zone of proximal development, such that exercises are not too easy or too difficult (Tomlinson et al., 2003). Practically, this means that exercises with difficulty levels closest to the individual's level of mastery are offered to students.

Mastery learning and the zone of proximal development point out the adaptive practice process as the more effective approach to increase learning outcomes. But, at the same time, this is not evident as the static practice environment may prepare students more effectively for the upcoming summative tests. Indeed, students in the static practice environment may often receive exercises that are too easy or too difficult, but at least these type of exercises are repeatedly addressed (if they are too easy) or not avoided (if they are too difficult). This can be of utmost importance in order to perform well in the upcoming summative test, or even to maintain the motivation to practice.

Finally, feedback is often accredited to be one of the most important factors in facilitating learning (Hattie & Timperley, 2007). It has also been found that the optimal feedback strategy is contingent on a student's current performance level (Shute, 2008). While practicing, students immediately receive feedback on the accuracy of their answer, together with cues on where information can be found that can contribute to a better level of understanding. For students in the static condition, it implies that they relatively often receive feedback on exercises that deviate, in terms of difficulty and knowledge-type level, from their own mastery level. Students in the adaptive practice environment always receive feedback on exercises that are assumed to be in their zone of proximal development. Even though the feedback strategy is similar for students in both environments, the results in Shute (2008) would predict that the feedback procedure in the adaptive condition contributes more effectively to student learning.

4. Experimental design

At the beginning of the school year, teachers and their students fill out an electronic survey to collect information on background characteristics. Students of each participating school are randomly assigned *within classes* to one of the two practice environments, as to control for school-, class- and teacher-specific characteristics. Randomization occurs using a specific PHP-command (i.e. array_rand-function), which assigns for each class one half of the students to the static group and the other half of the students to the adaptive group. More specifically, this command selects n random elements out of an $Nx1$ vector. The vector N contains all students in a given class, and n is defined to be a random half of that number. If a class holds an odd number of students, one student is randomly assigned to either the static or the adaptive group, and the remaining (even) number of students are allocated following the procedure described before.[4]

With the actual implementation of the experiment, students in the two experimental groups are exposed to a different algorithm. It is of key importance here that students in both groups observe the same interface which only allows them to choose a subject and a chapter to work on. After this decision, students have no further control over the content they are exposed to by the program. Regardless assignment status, the exact sequence of exercises offered will always differ from that of their peers. The static algorithm draws exercises with predetermined probabilities, but is therefore still yielding different strings of exercises (even for students within the control group). Similarly, the adaptive algorithm draws exercises based on the aforementioned decision rules (e.g. relative level understanding), yielding individualized sequences of exercises.

Each practice session for a student is linked to the corresponding test for which the student is preparing. In general, students have a few weeks of classes, with corresponding homework, before a test is administered. After this period, the next practice period starts, in which new practice sessions are generated for each of the participating subjects and with exercises linked to the next test for that particular subject. During the year of the experiment, four subjects participate in the experimental sample: Dutch, Biology, History and Economics.

Throughout the experiment, all participating classes extensively use the digital learning environment. During instruction hours, students use a hybrid between a book and the digital learning environment. Homework is made completely in the digital environment, such that students are familiar with the software due to exposure in class and through homework sessions. Instruction hours and homework are exactly the same for both groups. As such, the intervention actually only kicks in after instruction and after finishing mandatory homework. The optional "practice mode", therefore, is where students are differentially assigned to either the static or adaptive practice environment. All teachers are asked to inform their students to freely practice at least a quarter of an hour a week, but students ultimately decide themselves how much they will use this practice environment. As such, practice intensity is an

[4] See also: http://php.net/manual/en/function.array-rand.php.

Table 2
Descriptive statistics: schools.

			Subjects (# students)				School Grades (# students)				
School	ISCED Level	# Students	# Classes	Dutch	Economics	Biology	History	7th	8th	9th	School size
1	2/3	105	4	0	105	0	0	0	0	105	1379
2	3	26	1	0	26	0	0	0	0	26	1578
3	3	147	7	105	0	0	67	26	121	0	358
4	3	180	7	0	0	180	0	180	0	0	1200
5	3	132	5	0	132	0	0	0	0	132	1081
6	2/3	81	3	29	52	0	0	29	0	52	998
7	2/3	350	19	322	0	.0	165	326	24	0	1327
Total		1021	46	456	315	180	232	561	145	315	

Table 3
Descriptive statistics: teachers.

		N	Mean	Std. dev.	Min	Max
Age (years)		13	43.535	13.747	26.06	60.84
Man		15	0.571	0.514	0	1
Educational competence	First degree	13	0.307	0.480	0	1
Education level	University of AS*	9	0.600	0.507	0	1
	University	4	0.267	0.458	0	1
	Education unknown	2	0.133	0.351	0	1
Experience years		13	7.307	6.395	1	25

*AS: Applied Sciences

important learning outcome to consider. For example, it is to be expected that students will practice more extensively when they perceive practicing to be interesting, enjoyable, or facilitating their learning process.

5. Descriptive statistics

Table 2 gives descriptive statistics for the experimental sample, disaggregated by the seven participating schools. The second column shows the International Standard Classification of Education (ISCED) for these schools. Schools in the Dutch system frequently offer heterogeneous classrooms in the first year(s) of secondary education and can also choose to offer all or only a subset of educational tracks. The first years of pre-vocational education relates to ISCED-level 2, whereas pre-university and upper secondary education represent ISCED-level 3. However, schools 3 and 5 do differ from schools 2 and 4, as the former offer only pre-university education, while the latter offer both pre-university and upper secondary education. Schools 1, 6 and 7 offer all education levels, as indicated by ISCED-level 2/3. As such, the experimental sample offers a variety of schools and educational offerings. A detailed description of the Dutch educational system is given in Appendix A.

Within these schools, students of the 7th, 8th and 9th grade participate in the experiment. Schools did not give consent to conduct the experiment in higher grades, as these students are in their final exam year, or already take school exams that count toward the final exam grades. Therefore, the stakes were considered too great for these grades. The bottom row indicates that 1021 students participate in the experiment. These students were randomly assigned within classes to both practice environments, which yields a statistical power 0.89 using a theoretical effect size of 0.2 SD. The statistical power is in fact somewhat higher, because students frequently practice for multiple subjects and receive several cognitive tests during the school year (in the experimental sample this is on average 8.5).

The learning outcomes of students in the experimental sample can be importantly influenced by the characteristics of their teacher and whether or not (s)he stimulates using the digital practicing module. Table 3 shows descriptive statistics of the 15 teachers participating in the experiment. The information on teachers characteristics was obtained by conducting surveys among teachers.[5] We note that the within-class randomization ensures that differences in learning outcomes between students in the adaptive and static environment cannot be influenced by (un)observed teacher characteristics (i.e. under the assumption that teachers do not cause spillover effects, a point to which we return empirically in Section 6.5).

Teachers are on average 43.5 years old, predominantly male, and 4 teachers (i.e. 30.7% of 13 teachers) possess a first-degree teaching qualification, while 9 teachers possess a second-degree teaching qualification. Teachers with a second-degree license are qualified to teach the first three grades in upper secondary education or pre-university and all grades in pre-vocational education. Teachers with a first-degree license are qualified to teach all classes in upper secondary, pre-university and pre-vocational education. Teachers with a first-degree license successfully finished university, while teachers with a second-degree qualification successfully finished a university of applied sciences. Teachers, on average, have 7.3 years of experience teaching at the current school. The standard deviation, minimum and maximum experience year further indicate that there are substantial differences between teachers.

Table 4 shows student background characteristics for both practice environments. This table categorizes the highest attained education level of parents using the classification of Statistics Netherlands.[6] Information on student characteristics was obtained by conducting a survey among students, for which the response rate was 91.48% (i.e. 87 of the 1021 students did not fill in the survey, which is reflected by the variable 'non-response'). The table shows that none of the mean differences between both environments are statistically significant at the 95% confidence level. A test of joint significance shows that both groups are not significantly different when the background characteristics in Table 4 are jointly considered.

[5] Unfortunately, two of the teachers did not fill in these questionnaires and for these teachers the information available is limited to just their gender.

[6] A low education level represents parents with either no, a primary- or a pre-vocational secondary education level. A middle education level represents parents with upper secondary, pre-university or secondary vocational education. A high education level represents parents with university of applied sciences, university or higher.

Table 4
Compare students characteristics between adaptive and static group.

	Static			Adaptive				
	N	Mean	Std. dev.	N	Mean	Std. dev.	|Δ|	p-values
Boy	472	0.504	0.023	462	0.511	0.023	0.007	0.841
Age	472	13.629	0.049	462	13.691	0.055	0.061	0.407
Home language is Dutch	472	0.941	0.011	462	0.942	0.011	0.001	0.954
Country of birth mother: Dutch	472	0.833	0.017	462	0.816	0.018	0.017	0.505
Mother's education: Low	472	0.100	0.014	462	0.087	0.013	0.013	0.495
Mother's education: Middle	472	0.189	0.018	462	0.238	0.020	0.050	0.065
Mother's education: High	472	0.318	0.021	462	0.312	0.022	0.006	0.841
Mother's education: Unknown	472	0.386	0.022	462	0.355	0.022	0.031	0.333
Country of birth father: Dutch	472	0.845	0.017	462	0.820	0.018	0.025	0.306
Father's education: Low	472	0.106	0.014	462	0.091	0.013	0.015	0.442
Father's education: Middle	472	0.165	0.017	462	0.195	0.018	0.030	0.240
Father's education: High	472	0.350	0.022	462	0.355	0.022	0.005	0.863
Father's education: Unknown	472	0.375	0.022	462	0.348	0.022	0.027	0.400
Non-response	41	0.080	0.012	46	0.091	0.013	-0.011	0.543
Test of joint significance				Prob>F = 0.66				

Table 5
Descriptive statistics on practicing outcomes.

	N	Mean	Std. dev.	5th percentile	95th percentile
All students:					
Practicing time per session (minutes)	1021	14.376	14.722	0.00	46.658
Exercises Practiced per Session	1021	28.98	48.989	0.00	119.00
Success rate	926	0.729	0.085	0.588	0.870
Students who practice:					
Practicing time per Session (minutes)	926	15.617	19.804	1.216	47.685
Exercises Practiced per Session	926	31.954	50.511	1.636	121.80

Dutch is the home language for 94% of the students and somewhat over 80% of the parents are born in the Netherlands. Students are on average 13.6 years old, which is the result of considering students in grade 7, 8 and 9. Almost 40% of the students were not able to indicate the highest education level of their parents. Of the students that could indicate their parents' education level, a relatively large proportion answered that the highest attained education level of their parents was university of applied sciences or higher.

Table 5 presents descriptive statistics for several practice outcomes per session. The average duration of each session is 19.5 days.[7] The table shows that students practice per session, on average, 14.4 min in which they finish approximately 29 exercises. The average success rate indicates that students answer 73% of the exercises correctly. Also, 95 students did not practice at all during the experimental window, which is why we observe a success rate for 926 students. The lower panel shows that these 926 students practice per session, on average, 15.6 min in which they make around 32 exercises.

To better characterize the difference between the two practice environments, Fig. 2 illustrates the difference between both environments, as expressed by the average difficulty level of the exercises received by students.[8] If environments do not differ in terms of difficulty, exercises of each level of difficulty would be assigned equally often to students in both groups. This situation is reflected by the horizontal dash-dotted line. Exercises of a certain difficulty level are relatively more frequently received by students in the adaptive (static) condition if the plotted line is above (below) this dash-dotted line.

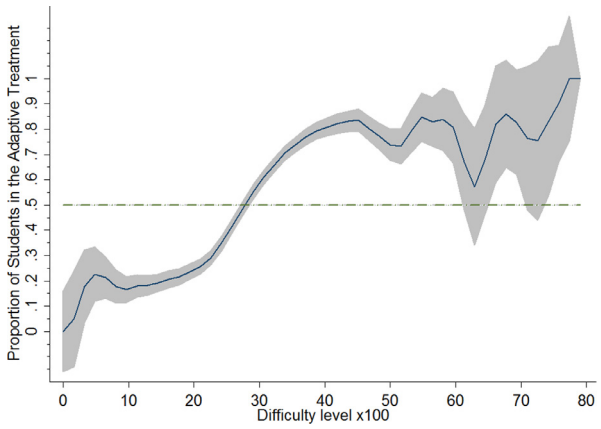

Fig. 2. Proportion of students in adaptive treatment by difficulty level.

The average difficulty level per session does not exceed 0.8 and the confidence interval reveals that the average difficulty level per session lies between 0.1 and 0.45. The static practicing algorithm generates valid and representative practicing tests, by each time generating a sequence of questions containing all topics, knowledge types and difficulty levels. By contrast, students in the adaptive condition receive, on average, more difficult exercises per session (and with a relative oversampling of topics that they relatively poorly understand). If students in the adaptive practice environment indeed receive exercises that better suit their learning needs, the pattern in this figure suggests that, on average, receiving more difficult exercises better accommodates the learning needs of students.

[7] The actual (maximum) duration of each "session" could be determined by using the dates of subsequent summative tests. This approach could not be used for first sessions (i.e. no starting point), for periods in which there were school holidays, and for when there was a discontinuity in the offering of a subject (not all subjects are taught all year round). As a result, the average presented here is the exact duration for 60% of all sessions recorded.

[8] We note that the difficulty level of each received exercise is determined based on historical practicing data of, on average, 3000 students. The difficulty level represents one minus the proportion of correctly answered exercises.

6. Estimation strategy & results

6.1. Estimation strategy

The effect of an adaptive versus static practice process is estimated by fitting the following model, using ordinary least squares (OLS):

$$Y_{is}^{k} = \alpha + \beta A_i + X_i'\gamma + \varepsilon_{is}, \quad (4)$$

where Y_{is}^{k} represents learning outcome k for student i in session s. The k learning outcomes considered in this study are (1) achieved test scores, (2) practicing time, (3) the number of exercises made, and (4) the number of correctly answered exercises. A indicates if the student was assigned to the adaptive ($A = 1$) or the static ($A = 0$) treatment, and X represents a vector of background characteristics. ε_i is assumed to be a random error term (i.e. $Cov(\varepsilon_i, \varepsilon_{-i}) = 0$), but because $Cov(\varepsilon_{is}, \varepsilon_{i-s}) \neq 0$ we cluster standard errors at the student level.

The practice intensity of classes differ between sessions and when the practice intensity in a class is low there will be no differences in learning outcomes due to the differential effectiveness of both algorithms. A unique feature of experimental set up is that the test moments and practice sessions of all participating classes were synchronized, which allows us to rank classes in each session according to their practice intensity, as measured in average practice time per session. Since students are randomly assigned within classes, we can exploit this session-specific variation in rankings to estimate the differential effectiveness of both algorithms on learning outcomes. For this, we create three sub-samples by identifying, per session, the top, middle, and bottom 33% classrooms in terms of practice intensity. The proportion of students who practice a substantial amount of exercises is higher for classes in a higher practice-intensity classroom, such that the differential effectiveness of both algorithms is likely to be larger for these classes. Put differently, we expect that a differential effect between the two algorithms will be larger (smaller) in high (low) intensity classes. To gain insight in the differential effectiveness of using one of both algorithms, instead of merely *offering* adaptive versus static practice software, Eq. (4) is estimated for these three sub-samples.[9]

6.2. Effects on achieved test scores

Table 6 show the estimation results for four model specifications with achieved test scores as the dependent variable. The possible test scores that students can achieve on a tests taken (in all classes and for all subjects) ranges from 1 to 10 ($M = 6.47$, $SD = 1.89$) and the dependent test score variable used in the empirical analysis reflects the actual achieved test scores. It is not necessary to use standardize test scores, because students are randomly assigned to both practice environments within classrooms. The reported number of student observations and student clusters indicate that students are tested, on average, 8.5 times during the school year. The estimation results in all four model specifications show that the test scores achieved by students exposed to the adaptive practice environment are not significantly different from those assigned to the static practice environment. The inclusion of control variables does not result in large efficiency gains in the estimated coefficient.

The effect of the adaptive practice environment may vary with a student's position in the test score distribution. Therefore, Fig. 3 plots the proportion of students in the adaptive group by test score bins. If relative effectiveness is independent from the test score distribution, the proportion of students in the adaptive group must be 0.5 for each test score bin, since students are randomly assigned to both practice environments.[10] The figure shows that for students who achieved test scores higher than or equal to 1, the proportion of students exposed to adaptive practicing is not significantly different from 0.5 for all consecutive test score bins. In 180 of the 8682 observed test scores, students score a test score lower than one, and for these cases the proportion of students exposed to the adaptive environment is slightly, but statistically significantly, higher than 0.5. This helps to explain the small non-significant negative coefficients reported in Table 6.

The practice intensity of classes differs between sessions and differential effectiveness of both practice environments should be increasing with the practice intensity in a classroom. As explained in Section 6.1, the synchronization of test moments and practice sessions allows us to rank classes in each session according to their practice intensity and to exploit this session-specific variation. First, we constructed a sample in which classes that did not practice in particular sessions were removed (with test scores $M = 6.59$, $SD = 1.65$). Then, we constructed a session-specific ranking and constructed three additional sub-samples by selecting the top, middle and bottom 33% classes in terms of practice intensity. Students in the highest intensity quintile of classes practice on average for 32.9 min per session (with test scores $M = 5.95$, $SD = 2.18$), for the middle quintile this is 13.8 min per session (with test

Table 6
Baseline estimation results: test scores.

	(1)	(2)	(3)	(4)
Adaptive Practice	−0.073	−0.071	−0.090	−0.068
	(0.068)	(0.064)	(0.063)	(0.059)
Constant	6.505***	4.913***	5.273***	8.428***
	(0.045)	(0.150)	(0.634)	(0.816)
Pre-test scores	*No*	*Yes*	*Yes*	*Yes*
Student-level controls	*No*	*No*	*Yes*	*Yes*
Other control variables	*No*	*No*	*No*	*Yes*
N	8682	8682	8682	8682
# Student clusters	1021	1021	1021	1021
R^2	0.000	0.047	0.059	0.143

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The student control variables are gender, age, home language, mother's and father's education level. The other control variables included are classroom-, subject- and grade dummies. Pre-test scores are available only for students who start practicing after the first test is taken mid October (i.e. 35 of the 46 classes). Therefore we assigned a value of −10 to the missing pre-test scores and include a dummy variable that indicates 1 if the pre-test score was missing, and zero otherwise.

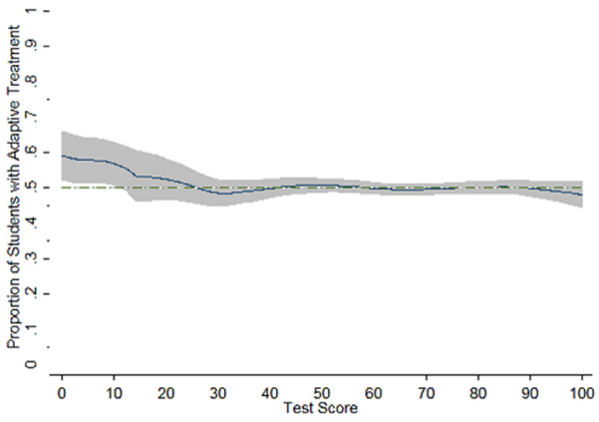

Fig. 3. Proportion of students in adaptive treatment by test score bins.

[9] We verified that the conclusions do not change when we construct sub-samples using different practice-intensity quintiles.

[10] Test scores are measured accurately to one decimal place on a ten-point scale. We generated test score integers multiplied by 10, such that the size of each bin is one-tenth of a test score, or 1 point in the figure.

Table 7
Estimation results: test scores by practice intensity.

			Quintile analysis:		
	(4)	(# exercises > 0)	0-33%	33-67%	67-100%
Adaptive Practice	−0.068	−0.088	−0.095	−0.035	−0.146
	(0.059)	(0.063)	(0.082)	(0.081)	(0.111)
Constant	8.428***	8.226**	5.090***	8.856***	9.077***
	(0.816)	(0.941)	(1.208)	(0.953)	(1.337)
Pre-test scores	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Student-level controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Other control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
$\bar{y}$-control	6.50	6.50	6.78	6.70	6.02
N	8682	7525	2475	2536	2514
# Clusters	1021	1021	752	791	679
R^2	0.143	0.138	0.155	0.104	0.171

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The included student- and other control variables are similar to those reported in Table 6.

scores $M = 6.68$, $SD = 1.77$), and for the lowest intensity quintile this is only 3.24 min per session (with test scores $M = 6.73$, $SD = 1.69$).[11] The estimation results for effects on test scores for these four sub-samples are shown in Table 7. In order to compare the results with the results presented in Table 6, we again show in Column 2 the Table 6 results of model specification (4). The table shows that removing classes that do not practice during a certain session reduces the total number of student observations by 1157. We note that the number of student clusters remains 1021, indicating that we observe practicing periods for all students during the school year. After removing non-practicing classrooms, the estimation effect of the relative effectiveness of the adaptive practice environment becomes slightly, but not significantly, more negative compared to the result presented in column 2. The estimation results for the three practice-intensity sub-samples indicate that the effect of adaptive practicing becomes more negative if classrooms practice relatively more intensively, but effect sizes remain statistically insignificant.

6.3. Effects on practicing outcomes

Of all 1021 students participating in the experiment, 90.7% take up the possibility to use the voluntary practice module. There is no statistically significant difference between students in the adaptive (91.5%) and static (89.9%) environment (p-value: 0.18). This makes sense, as students, at least initially, do not know to which group they have been randomly assigned and should, therefore, exhibit a similar propensity to start using the practicing module. However, while experiencing the actual (differential) process of practicing, and its potential consequences regarding test results, practicing behavior can diverge between both environments. Therefore, this section focuses on differential effects on practicing time in minutes per session (Table 8), the number of exercises made Table 9), and the number of correctly answered exercises (Table 10). Columns 2 and 3 of Table 8 show that students in the adaptive group practice longer. The last three columns show that this effect is magnified as classroom-level practice intensity increases. However, the estimation results of Tables 8 and 9 also reveal that, while students in the adaptive environment practice longer, they do not complete more exercises. Thereby, the estimation results seem to indicate that students do not practice longer, because adaptive practicing motivates them more than static practicing. Instead, the results suggest that students need more time for the same amount of exercises, probably because the difficulty level of these exercises is higher. This reasoning is supported by the estimation results in Table 10, which show that the number of correct answers given by students in the adaptive group is smaller than those given by students in the static group (given that students in both groups make the same number of exercises).

6.4. Heterogeneous treatment effects

A recent meta-analysis on educational technology applications in K-12 education indicates that improvements were markedly more positive for boys and for students of relatively low ability (Cheung & Slavin, 2012). Therefore, we examine whether the effects of an adaptive practice process on learning outcomes are heterogeneous with respect to student gender and ability. Table 11 displays that boys perform worse and do not spend more time practicing than girls. Boys are able to finish more exercises than girls and, correspondingly, also answer more exercises correctly. However, there is no differential effect between environments by gender for any of the learning and practicing outcomes considered.

In Dutch schools, students fail a test when they score below 5.5 points. In Table 12, the findings are disaggregated by whether or not the student for that particular subject failed the last test, prior to when the class started using the digital practice module. The results show that students who failed this test, which in the experimental sample is true for approximately 17% of the students, used the digital practice module more intensively (each session lasted approximately 2.76 min longer). Furthermore, students who had passed the prior test perform statistically significantly worse on test scores when assigned to the adaptive practice environment (around −0.15 points, or -0.08σ). For students who had failed the prior test, no such negative effect is found. The aforementioned finding that students in the adaptive practice environment answer fewer questions correctly also seems to be driven by those who had passed the prior test. This could indicate that they, relative to those who had failed the prior test, are exposed to more difficult questions (more quickly) when assigned to the adaptive practice environment.

6.5. Substitution bias and spillover effects

In this section, we address the potential concern of substitution bias and spillover effects (i.e. students and/or teachers may distort the results on relative effectiveness with respect to test scores). If, once the first test scores become available, the teacher effectively redirects his/her attention to relatively low-performing students in the class, this might obscure any differential effectiveness of the algorithms. The same would occur if a student, as a result of the first test score, starts to increase effort and/or seek additional help

[11] Please note that test scores cannot be easily compared between quintiles, as these differ in grade, level, subject and test of the classes they contain.

Table 8
Estimation results: practicing time.

			Quintile analysis:		
	Full Sample	(# exercises > 0)	0–33%	33–67%	67–100%
Adaptive Practice	2.153***	2.489***	0.322	2.027**	5.368**
	(0.805)	(0.805)	(0.413)	(0.843)	(2.424)
Constant	−30.21***	−30.09**	−14.70	10.94	−38.25
	(8.466)	(12.86)	(10.64)	(10.56)	(23.89)
Pre-test scores	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Student-level controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Other control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
$\bar{y}$-control	13.58	15.67	3.13	13.10	30.28
N	8682	7525	2475	2536	2514
# Clusters	1021	1021	752	791	679
R^2	0.225	0.216	0.070	0.134	0.236

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The included student- and other control variables are similar to those reported in Table 6.

Table 9
Estimation results: log (# exercises).

			Quintile analysis:		
	Full Sample	(# Exercises > 0)	0-33%	33-67%	67-100%
Adaptive Practice	0.023	0.026	0.057	0.061	−0.024
	(0.047)	(0.052)	(0.053)	(0.075)	(0.986)
Constant	−3.207***	−2.353***	−1.667**	0.762	−0.428
	(0.628)	(0.678)	(0.654)	(0.840)	(23.89)
Pre-test scores	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Student-level controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Other control variables	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
$\bar{y}$-control	1.74	2.00	0.65	2.09	3.22
N	8682	7525	2475	2536	2514
# Clusters	1021	1021	752	791	679
R^2	0.168	0.171	0.049	0.073	0.073

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The included student- and other control variables are similar to those reported in Table 6.

Table 10
Estimation results: number of correct answers.

			Quintile analysis:		
	Full Sample	(# exercises > 0)	0–33%	33–67%	67–100%
Adaptive Practice	−5.112***	−5.808***	−0.476	−2.157	−15.06***
	(1.852)	(2.054)	(0.532)	(1.889)	(5.187)
Constant	−129.1***	−144.0***	−15.13*	−16.32	−76.04
	(25.67)	(27.61)	(8.683)	(22.72)	(53.31)
Pre-test scores	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Student-level controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Other control variables	*Yes*	*Yes*	*No*	*Yes*	*Yes*
$\bar{y}$-control	25.76	29.73	4.32	22.58	61.28
N	8682	7525	2475	2536	2514
# Clusters	1021	1021	752	791	679
R^2	0.175	0.187	0.072	0.104	0.221

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The included student- and other control variables are similar to those reported in Table 5.

as to overcome a relative poor performance on this first test. If this happens, one should see a shift over time in the performance difference between the two experimental groups.

To investigate whether this indeed is the case, test scores for each subject are numbered chronologically. Fig. 4 displays the effect of the adaptive practice process on test scores for each test number separately (the black line) and the surrounding dotted line represents the confidence interval. The dotted horizontal line in this figure represents the estimated non-significant baseline effect (i.e. −0.068 points).

The results indicate that none of the consecutive estimates are statistically significantly different from the estimated effect for test number one. Put differently, the estimation results do not vary over the course of the year. This is in line with an interpretation in which students and teachers do not come to learn that students of a particular experimental condition are structurally performing differently from their class peers and/or are not able to effectively overcome this difference by increasing effort and seeking additional help (i.e. students) or by redirecting attention (i.e. teachers). As such, we interpret the non-significant results on test scores as that, on average, both versions of the practice module are equally effective in preparing students for the test.

7. Practice experiences

A follow-up questionnaire was taken in late June, which provides information about students' experiences using the practice module. The questionnaire was administered in 33 of the 46

Table 11
Heterogeneous treatment effects: gender.

	Model 4: Practice Intensity of the Class > 0			
	Test scores	Practice Time	Log(# Exercises)	# Correct Answers
Adaptive Practice (*AP*)	−0.0943	2.512***	0.0207	−5.992***
	(0.0637)	(0.918)	(0.0527)	(2.086)
AP·Boy	−0.0256	0.0968	−0.0202	−0.790
	(0.0198)	(0.251)	(0.0196)	(0.543)
Boy	−0.288***	−0.361	0.170***	8.849***
	(0.0669)	(0.961)	(0.0563)	(2.244)
Constant	8.239***	−30.14**	−2.342***	−143.6***
	(0.931)	(12.85)	(0.676)	(27.56)
Pre-test scores	*Yes*	*Yes*	*Yes*	*Yes*
Student-level controls	*Yes*	*Yes*	*Yes*	*Yes*
Other control variables	*Yes*	*Yes*	*No*	*Yes*
$\bar{y}$-control	6.50	15.67	2.00	29.73
N	7525	7525	7525	7525
# Clusters	1021	1021	1021	1021
R^2	0.138	0.171	0.222	0.187

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The included student- and other control variables are similar to those reported in Table 6.

Table 12
Heterogeneous treatment effects: ability.

	Model 4: Practice Intensity of the Class > 0			
	Test scores	Practice Time	Log(# Exercises)	# Correct Answers
Adaptive Practice (*AP*)	−0.146**	2.790**	0.0229	−7.877**
	(0.0718)	(1.106)	(0.0620)	(2.530)
AP·prescore_fail	0.251*	−1.287	0.0103	9.128***
	(0.147)	(1.575)	(0.107)	(3.321)
prescore_fail	−0.266**	2.755*	−0.0591	−4.359
	(0.128)	(1.572)	(0.102)	(3.414)
Constant	8.429***	−33.04***	−2.278***	−143.9***
	(0.946)	(12.69)	(0.686)	(27.30)
$\bar{y}$-control	6.50	15.67	2.00	29.73
Pre-test scores	*Yes*	*Yes*	*Yes*	*Yes*
Student-level controls	*Yes*	*Yes*	*Yes*	*Yes*
Other control variables	*Yes*	*Yes*	*Yes*	*Yes*
N	7525	7525	7525	7525
# Clusters	1021	1021	1021	1021
R^2	0.139	0.172	0.222	0.188

Note: */**/*** denote significance at 10/5/1% level (two-sided). SEs are clustered at the student level. The included student- and other control variables are similar to those reported in Table 6.

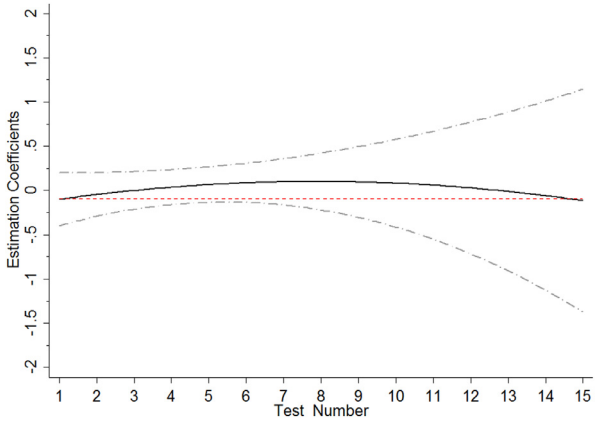

Fig. 4. Estimation results by test number.

classes, and in these classes the response rate was 69% (i.e. 509 of the 739 students filled in the follow-up questionnaire). Unfortunately, the questionnaire was not administered in 13 classes because of extracurricular activities planned before the summer holiday.[12] Even though 739 students in the 33 classes were randomly assigned within classes to both practice environments, it may be that the practice environment influenced the probability of responding to the follow-up questionnaire. If students are, for example, more (less) satisfied with the adaptive practice process, it follows that students assigned to this condition may be relatively more (less) likely to respond to the request to participate in the follow-up questionnaire. Table 13 shows that the proportion of non-response to the follow-up questionnaire is not correlated with the practice environment. As a result, response itself may be selective, but in the examination of difference between outcomes on the follow-up questionnaire, these selection effects are accounted for through the initial randomization.

In the follow-up questionnaire, students learned that they had been assigned to either a static or an adaptive practice environment. The questionnaire literally states: *"Perhaps you did not know, but there were two different practice environments. You have been practicing with one of them"*. Both practice environments are referred to as version 1 and 2, and the underlying differences between the two conditions are intuitively explained. This question

[12] The estimation results for these 33 classes were similar to those presented in Section 6. The results are available on request.

Table 13
Non-response follow-up questionnaire by treatment status.

	Static (N = 371)		Adaptive (N = 368)			
	Mean	Std. dev.	Mean	Std. dev.	|Δ|	p-values
Non-response Follow-up	0.313	0.024	0.310	0.024	0.003	0.933

Table 14
Follow-up questionnaire I.

		Static (N = 255)		Adaptive (N = 254)		
		Mean	Std. dev.	Mean	Std. dev.	p-values
I practiced with version 1		0.639	0.030	0.685	0.029	0.275
Computerized practice is fun	No	0.278	0.028	0.244	0.027	0.379
	Neutral	0.361	0.030	0.406	0.031	0.300
	Yes	0.361	0.030	0.350	0.030	0.807
Practiced in the week when subject matter was taught	Sometimes	0.431	0.031	0.504	0.031	0.101
	Often	0.255	0.027	0.240	0.027	0.701
	Always	0.314	0.029	0.256	0.027	0.149
Practice for test preparation	Sometimes	0.612	0.031	0.610	0.031	0.972
	Often	0.239	0.027	0.248	0.027	0.817
	Always	0.145	0.022	0.138	0.022	0.814

Table 15
Follow-up questionnaire II.

		Static (N = 255)		Adaptive (N = 254)		
		Mean	Std. dev.	Mean	Std. dev.	p-values
CP causes better understanding	Disagree	0.282	0.028	0.299	0.029	0.676
	Neutral	0.412	0.031	0.402	0.031	0.815
	Agree	0.298	0.029	0.299	0.029	0.977
CP causes quicker understanding	Disagree	0.322	0.029	0.331	0.030	0.826
	Neutral	0.392	0.031	0.417	0.031	0.564
	Agree	0.278	0.028	0.248	0.027	0.437
CP offers too difficult exercises	Disagree	0.435	0.031	0.390	0.031	0.298
	Neutral	0.420	0.031	0.472	0.031	0.231
	Agree	0.133	0.021	0.134	0.021	0.986

*CP: Computerized Practicing

was asked at the end of the questionnaire, such that students could not change the answers given before.[13] The first row with descriptive statistics in Table 14 shows the proportion of students who thought they were assigned to the static condition separately for students in both conditions. Roughly 65% of the students in both practice environments indicated that they were assigned to the static conditions and the mean difference between the two students groups not statistically significant. This result is interesting, as students did not know to which practice environment they were assigned at the beginning of the experiment, and, apparently, did not come to learn this during the course of the experiment either. The majority of students in both practice environments indicated that they were assigned to the static group, which may reflect that neither two practice environments were experienced as particularly adaptive. The student opinions with respect to whether computerized practicing is fun, and how often they practice, are very much equally divided. None of the mean differences between both conditions are statistically significant. Thus, students in both practice environments have experienced, and value, the practice module similarly and also state similar frequencies of using this option.

Table 15 shows if students think that Computerize Practice (*CP*) helps them to understand subject matter better or faster, and if students experienced the exercises received as being too difficult. If an adaptive practice process better suits the learning needs of students, we would expect students in the adaptive environment, more so than students assigned to the static environment, to state that *CP* helps them to understand subject matter better or faster. Table 15 reveals that there are no significant mean differences between the students in both conditions. This implies that, in terms of understanding and efficiency, the stated practice experience of students in the adaptive environment is not different from that of students in the static environment. An interesting outcome is that there is no difference between environments in whether students express that exercises were too difficulty, while in reality students assigned to the adaptive condition receive, on average, more difficult exercises (see Fig. 2) and on topics that are relatively poorly understood. On the one hand, this may indicate that the adaptive condition returned exercises to students that better suited their learning needs. On the other hand, students do not value both practice environments differently, and both practice environments seem to affect learning outcomes similarly, indicating that both versions are equally successful in addressing the learning needs of students.

8. Conclusion and discussion

This study evaluates if an adaptive practice process, relative to a static practice process, improves learning outcomes. The outcomes considered are test scores, practice time, number of exercises made, and number of correctly answered exercises. The static practice process implies that students receive a predetermined path of exercises (and feedback) in terms of topics, question types and difficulty levels. The adaptive practice process implies that students receive a sequence of exercises (and feedback) conditional on the student's performance during the practice pro-

[13] Students filled in a digital questionnaire which prevented students to go back to earlier questions once the last question is reached.

cess. Theories of learning (i.e. self-determination theory, mastery learning and zone of proximal development) predict the adaptive practice process is more effective in promoting learning outcomes. This study evaluates the relative effectiveness of two versions of a digital practice module by conducting a randomized field experiment, with a duration of one year, in which 1021 secondary school students were randomly assigned to either a static or an adaptive practice environment.

The experimental findings show that the achieved test scores by students do not depend on whether they were assigned to the static or to the adaptive practice environment. However, students of higher ability (i.e. who had passed the last test taken before their class started to actively participate in the experiment) achieve lower test scores on the corresponding test when practicing adaptively (-0.15 points, or -0.08σ). More generally, students working with the adaptive version receive more difficult exercises, practice longer, but answer fewer questions correctly. The practice experience, as self-reported by students, with respect to effectiveness, efficiency and satisfaction is similar for both groups. This, together with the notion that students have remained unaware of their treatment status throughout the year, and the fact that estimation coefficients remain constant over time, supports the argument that there have been no significant distortions due to potential substitution bias and spillover effects. As such, the conclusion is that both algorithms considered in this study are equally effective in fostering student learning.

An interesting discussion with respect to the effectiveness of both practice environments arises from the argument that effectively preparing students for a test might well be something different than addressing their learning needs. It is well-documented that repetition and retrieval of content, for example by means of a formative assessment, is essential for the retention of this knowledge (Johnson & Mayer, 2009; Karpicke & Roediger, 2008). Even though the static practice environment will sometimes offer student exercises that they do or do not master, it does prepare them effectively for the upcoming summative test, by providing an equal amount of questions across the entire spectrum of topics, knowledge types and difficulty levels. It moreover provides them each session with valid formative tests. The adaptive practice environment aims to adapt to the student learning needs, but implicitly it assumes that the underlying decision rules (such as mastery threshold levels, probability distributions) are valid and accurate throughout the practice process. Therefore, it is important that these assumptions underlying the adaptive practice environment are empirically tested, with the aim to make personalized practicing environments more effective and to better understand the mechanisms for an effective practice process.

A research avenue worth pursuing, which follows from the recognition that personalized education not only allows for different (practice) paths, but also for different targets, would be to introduce and evaluate student-specific learning goals and dynamic feedback in adaptive software programs. Effective goal-setting requires that the learning objectives are not considered too easy, yet feasible (Locke & Latham, 1990). Empirical studies show that goal-setting can increase self-discipline and (academic) performance (Duckworth, Kirby, Gollwitzer, & Oettingen, 2013; Latham & Brown, 2006), enhances self-regulation (Oettingen, Hoenig, & Gollwitzer, 2000), and that students are less distracted in the process of achieving their learning objectives (Kruglanski et al., 2002). An important and necessary condition for effective goal-setting is that students also receive feedback during the learning process. Regarding feedback, there is a vast body of research confirming its importance with respect to student learning, although not providing uniformly positive results, which is considered to be the result of the wide variety of types of feedback and the way it is offered (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Promising for the development of effective adaptive practice software, however, is that providing cues or reinforcement related to learning goals through a computer are among the most effective examples of feedback that have been evaluated (Hattie & Timperley, 2007). Research suggests that the optimal feedback strategy, in terms of type (e.g. positive/negative, directive/facilitative), amount offered and timing (e.g. immediate/delayed), is contingent on learner characteristics and learning outcomes, such that adaptive formative feedback systems are required (Shute, 2008). Current computerized practicing tools, however, generally do not lead to individual learning goals, nor do they include personalized adaptive feedback. Therefore, by incorporating goal-setting and feedback in an adaptive practice environment, students should be better informed on the specific objectives to be achieved and more in control of their own learning process.

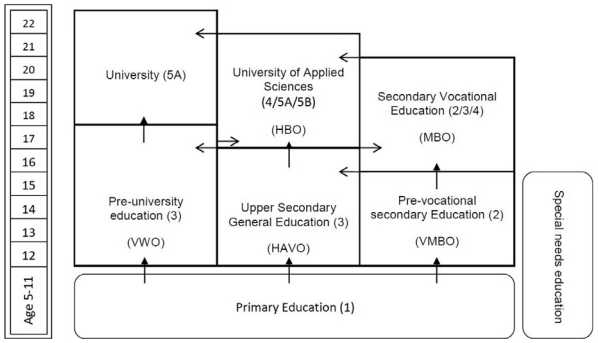

Fig. 5. Dutch education system.
Note: ISCED levels are shown in parentheses.

Acknowledgments

Van Klaveren acknowledges financial support for this research project from NWO (405-14-510). Support for the collection of data has been provided by Dedact, a developer of digital learning materials. The authors bear full responsibility for the analysis and interpretation of these data. The authors thank Henry Levin and Eric Hanushek for their valuable comments.

Appendix A. The Dutch education system

Fig. 5 displays the general structure of the Dutch educational system.

Children in the Netherlands can already enroll in primary education when they are four years old, but are required to start with primary education in the year they turn five years old. Generally, these children finish primary education at the age of twelve (grade 6), after which they are tracked into different education levels in secondary schools, based on the results of a standardized test in grade 6 and the primary school teacher's advice. The secondary education levels in which children are tracked are: (1) Pre-vocational education (4 years), (2) General secondary education (5 years), and (3) Pre-university education (6 years). Pre-vocational education prepares children for vocational education, secondary general education prepares children for universities of applied sciences, and pre-university education prepares children for universities.

Schools can either offer the entire spectrum of tracks available, a specific subset of tracks, or just one track (e.g. pre-vocational education). This study focuses on children enrolled in the first three years of secondary education. In these years, schools offering different tracks frequently combine students of different performance levels into heterogeneous classrooms. Generally, the first year (grade 7) is then used to make an updated decision on

whether a student should go onwards to pre-vocational, upper secondary or pre-university education. Teachers in these classrooms are expected to effectively differentiate their instruction as to accommodate the various levels of ability. Some schools still allow for heterogeneous classrooms in the second year (grade 7) and a few still do so in the third year (grade 8).

References

Angrist, J., & Lavy, V. (2002). New evidence on classroom computers and pupil learning. *The Economic Journal, 112*(482), 735–765.

Banerjee, A., Cole, S., Duflo, 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.

Bloom, B. S., of College, C., & Examiners, U. (1956). *Taxonomy of educational objectives*: 1. David McKay New York.

Bulman, G., & Fairlie, R. W. (2015). *Technology and education: Computers, software, and the internet*: (vol.5. North Holland, Amsterdam. forthcoming).

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of reading and mathematics software products: Findings from two student cohorts. *Technical Report*. Mathematica Policy Research.

Cheung, A. C., & 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., & 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.

Coubergs, C., Struyven, K., Gheyssens, E., & Engels, N. (2015). Het bkd-leer-kracht model: binnenklasdifferentiatie realiseren in de klas. *Impuls, 45*(3), 151–159.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3–4), 325–346.

Duckworth, A., Kirby, T., Gollwitzer, A., & Oettingen, G. (2013). From fantasy to action: Mental contrasting with implementation intentions (mcii) improves academic performance in children. *Social Psychological and Personality Science, 4*(6), 745–753.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., … Javitz, H., et al. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort*. National Center for Education Evaluation and Regional Assistance. Report to congress.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.

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

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218.

Kruglanski, A., Shah, J., Fishbach, A., Friedman, R., Chun, W., & Sleeth-Keppler, D. (2002). A theory of goal systems. *Advances in Experimental Social Psychology, 34*, 331–378.

Kulik, C.-L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*(1), 75–94.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265–299.

Kulik, J. A., & Kulik, C.-L. C. (1987). Review of recent research literature on computer-based instruction. *Contemporary Educational Psychology, 12*(3), 222–230.

Latham, G., & Brown, T. (2006). The effect of learning vs. outcome goals on self-efficacy, satisfaction and performance in an mba program. *Applied Psychology, 55*(4), 606–623.

Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *Review of Economics and Statistics, 89*(4), 721–736.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review, 53*(1), 1–18.

Locke, E., & Latham, G. (1990). A theory of goal setting and task performance.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*(4), 423–458.

Machin, S., McNally, S., & Silva, O. (2007). New technology in schools: Is there a payoff?*. *The Economic Journal, 117*(522), 1145–1167.

Miliband, D. (2006). Choice and voice in personalised learning. *Schooling for Tomorrow*, 21–30.

Muralidharan, K., Singh, A., Ganimian, A. J., et al. (2016). Disrupting education? Experimental evidence on technology-aided instruction in India. *Technical Report*.

OECD (2015). Students, computers and learning: Making the connection. *PISA*.

Oettingen, G., Hoenig, G., & Gollwitzer, P. (2000). Effective self-regulation of goal attainment. *International Journal of Educational Research, 33*(7–8), 705–732.

Reezigt, G., Houtveen, A., & Van de Grift, W. (2001). *Vormgeving en effecten van adaptief onderwijs*. Groningen: RUG/GION.

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.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research, 57*(2), 175–213.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21.

Slavin, R. E. (2004). Education research can and must address "what works" questions. *Educational Researcher, 33*(1), 27–28.

Tomlinson, C. A. (2001). *How to differentiate in mixed-ability classrooms*. Alexandria, VA: Association for Supervision & Curriculum Development.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., … Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2–3), 119–145.

West, D. M. (2011). *Using technology to personalize learning and assess students in real-time*. Washington, DC: Brookings Institution.