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 ($0.08\sigma$). 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.