Delaying Access to a Problem-Skipping Option Increases Effortful Practice: Application of an A/B Test in Large-Scale Online Learning

Alexander O. Savi[a,*], Nienke M. Ruijs[b], Gunter K. J. Maris[c,a], Han L. J. van der Maas[a]

[a] *Department of Psychology, Psychological Methods, University of Amsterdam, The Netherlands*
[b] *Research Institute of Child Development and Education, Research Priority Area Yield, University of Amsterdam, The Netherlands*
[c] *Cito Institute for Educational Measurement, The Netherlands*

Abstract

We report on an online double-blind randomized controlled field experiment (A/B test) in Math Garden, a computer adaptive practice system with over 150,000 active primary school children. The experiment was designed to eliminate an unforeseen opportunity to practice with minimal effort. Some children tend to skip problems that require deliberate effort, and only attempt problems that they can spontaneously answer. The intervention delayed the option to skip a problem, thereby promoting effortful practice. The results reveal an increase in the exerted effort, without being at the expense of engagement. Whether the additional effort positively affected the children's learning gains could not be concluded. Finally, in addition to these substantial results, the experiment demonstrates some of the advantages of A/B tests, such as the unique opportunity to apply truly blind randomized field experiments in educational science.

Keywords: teaching/learning strategies, evaluation of CAL systems, evaluation methodologies, elementary education, interactive learning environments

1. Introduction

One of the main challenges in education research is to unravel causal relations. Randomized controlled trials are widely viewed as the gold standard in

*Correspondence concerning this manuscript should be addressed to Alexander Savi, University of Amsterdam, Department of Psychology, Psychological Methods, Postbus 15906, 1001 NK, Amsterdam, The Netherlands. Email: o.a.savi@gmail.com. Funding by Netherlands Organisation for Scientific Research, grant number 314-99-107. The final version of this manuscript is published in Computers & Education: `https://doi.org/10.1016/j.compedu.2017.12.008`.

studying causal effects (Athey & Imbens, 2016; Borghans et al., 2016; Slavin, 2002). However, the use of RCTs in education research is not uncontroversial. The main critiques are that they are expensive, take long to conduct, and only provide answers to narrowly defined questions. Moreover, while double-blinding is deemed essential to avoid experimenter effects in medical research, so far, this turned out to be near-impossible in education research (Deaton & Cartwright, 2016; Olson, 2004).

In this paper, we show that large-scale experiments in online learning environments, also referred to as A/B tests, can be used to solve some of these issues. We report on a successful application of an A/B test in a large-scale online computer-adaptive practice system for Dutch primary schools (Math Garden, with over 150.000 active users). In the A/B test we delayed the option to skip a problem. This option was used by some children to skip difficult problems and practice with minimal effort, and the delay was thus aimed at promoting more effortful practice. Before describing the experimental details, we first aim to build a basic understanding of A/B tests in relation to traditional educational experiments, introduce the online practice system that is central to the experiment, and then discuss our motivation for this particular intervention.

1.1. A/B Tests

A/B tests, the online equivalent of randomized controlled field experiments, are widely used by internet companies. In this section, we shortly dedicate some specific attention to the method of A/B testing, as there are relatively few applications in the field of online learning, especially in comparison to the thousands of A/B tests that large internet companies perform on a yearly basis, while the methodology has opened up massive opportunities for learning research.

Because of the huge scale of online learning, A/B tests enable mass experimentation that is virtually free of charge. There are no recruitment and data-collection costs, as participants already use the system and responses are tracked. Also, randomization is effortless, and adjustments to the environments can be made readily, precisely, and homogeneously. This is the reason A/B tests are sometimes said to be minimally invasive (Heffernan & Heffernan, 2014) and enable iterative improvement (Williams et al., 2014). Using A/B tests, we can successively test changes to learning environments to find out which components are effective. One might thus argue that A/B tests combine the scale and ecological validity of RCTs and the precision of laboratory experiments.

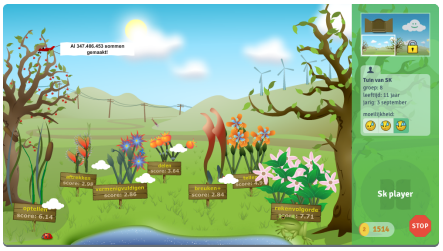
Importantly, a profound criticism of educational experiments, the practically near impossibility to satisfy a double-blinded procedure (e.g., Olson, 2004), does not pertain to A/B tests. Interventions in online learning environments neither need to rely on teacher instructions, nor need to be necessarily noticeable for the students. Let alone that the hypotheses that drive those changes need to be known to either teachers or students. A/B tests thereby have the power to effectively eliminate experimenter effects from educational experiments. This is not to say that A/B tests are a panacea. A/B tests only suit large-scale online education and are restricted to single platforms, consequently problems like external validity still require attention.

1.2. Math Garden

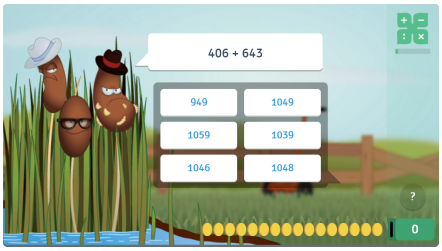
In this paper, we illustrate the method of A/B testing in the online learning environment Math Garden. We aim to reduce problem skipping and promote effortful practice. Before discussing the experiment, we first introduce the environment. Math Garden is an online environment for adaptive practice of math and math-related domains, spanning from addition and multiplication to logical reasoning and working memory. The system is used in over $1,500$ primary schools in The Netherlands, and currently has over $150,000$ active users, that collectively respond to more than 6 million items on a weekly basis. Such scale, its numerous sister systems for languages, typing, and statistics, and the symbiotic relationship between research and practice, provides an ideal basis for scientific research and continues to result in both methodological and substantive papers. Only some of the most recent research concerns topics ranging from the development of typewriting skills (van den Bergh et al., 2015), non-formal mechanisms in cognitive development of arithmetic (Braithwaite et al., 2016), and number transcoding in a language with inversion (van der Ven et al., 2016), to self-adapting success rates in math practice (Jansen et al., 2016).

Children that use Math Garden maintain a virtual garden, with different plants representing different domains and the health of a plant reflecting the frequency of practice. By selecting a plant, the child starts to practise a set of items within that domain. The system uses item response theory to estimate child abilities and item difficulties, and uses the Elo rating system to adaptively match children to items in real-time (Klinkenberg et al., 2011). In order to aid the accurate estimation of abilities and difficulties, a scoring rule with a speed-accuracy trade-off is employed (Maris & van der Maas, 2012). A response must be given within a certain time limit, which is visualized by a diminishing number of virtual coins at the bottom of the screen. Correct responses are rewarded with the remaining coins, whereas incorrect responses are punished by subtracting the remaining coins. Failing to give a response before the deadline results in neither a reward nor a punishment. After each item, the correct answer is shown, and the child proceeds to the next item.

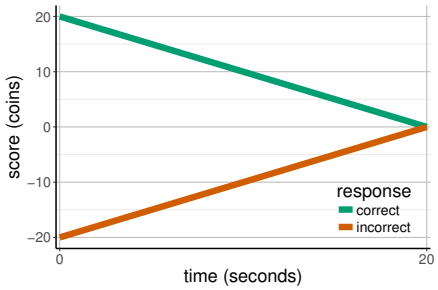
Each successful completion of a set of items within a domain earns the child some additional coins. The collected coins can be used to buy different kinds of virtual trophies. To cater individual differences with respect to desired difficulty, children may select the difficulty level themselves. This is reflected in the expected proportion correct (0.9 for easy items, 0.75 for medium items, and 0.6 for hard items). The rewarded/subtracted coins are doubled when using the hard level, and halved when using the easy level. Children may skip items that they deem too difficult to answer by hitting a question mark button. They are shown the correct answer and neither earn nor lose coins using this strategy. However, the adaptivity of Math Garden should generally prevent matching a child with an item that is too difficult. In Figure 1 we show some of the above elements.



(a) Example of a virtual garden. Plants represent domains. The smileys with different numbers of drops of sweat represent the difficulty levels.



(b) Example of an item from the addition domain. The remaining time (i.e., number of remaining coins) and the question mark button are shown on the bottom.



(c) The scoring rule. Rewards and punishments decrease linearly with time.

(d) A full trophy cabinet.

Figure 1: Math Garden.

1.3. Effortful Practice

A major aim of adaptive practice systems like Math Garden is to present problems at the level of the student. Other than in a traditional classroom environment, where in its most extreme case all students work through the same problems in the same pace, adaptive practice systems function as individual tutors. Vygotsky's theory of the zone of proximal development (Vygotskii, 1978) is central to this practice (Murray & Arroyo, 2002), and adaptive practice systems can be viewed as systems that seek to explore what a student can do with instruction or the outer boundary of what a student can do without instruction (depending on the level of instruction in the system). Specifically, Math Garden exploits the estimated difficulties of the problems, and makes sure each student receives little to no problems that are either too easy or too hard, and thus by balancing on the boundary of what a student can do without instruction.

By exploiting the zone of proximal development and delivering individual tutoring, adaptive practice systems seek to optimize learning gains. In return, this requires a serious and continuous effort from the student, as they are performing on the edges of their abilities. Not only does this take a great deal of motivation from the student (e.g., Pintrich, 1999), students do not always recognize the importance of effort for effective learning, and sometimes even falsely assume that easy problems are better for learning (Bjork et al., 2013). Therefore, Math Garden aids students directly in their motivation to practice problems by means of the virtual coin incentive, and indirectly by giving students the option to move closer towards or further away from the edge of their ability by means of the difficulty level selection.

1.4. Problem Skipping

In spite of these motivational aids, students still find ways to avoid difficult items, and for the current study Math Garden's question mark button is of particular interest. We noticed that some children use the question mark relatively often and relatively fast, which is probably best explained as strategic behaviour. Children that aim to maximize their earned coins pursue fast correct responses, and benefit from quickly skipping those items that they cannot spontaneously answer.[1] This strategy moves the child out of the zone of proximal development and severely reduces the amount of exerted effort.

Two reasons justify the aim to prevent this strategic behaviour. From a learning perspective, the behaviour relates at most, if at all, to surface learning. Those children do practise, but primarily by repeating known problems. Although this benefits memorization of those problems, it obscures the learning of

[1] Math Garden already utilizes one prevention for fast incorrect or question mark responses. Children are logged off from a domain if they submit x or more incorrect and/or question mark responses within the first 3.5 seconds, where x equals the number of items in the set, divided by 3, rounded to the nearest integer, and with a minimum of 3 and a maximum of 9 of such responses.

new problems. The biggest learning gain is to be found in the zone of proximal development, and will require active and effortful learning.

Also, from a measurement perspective, the accuracy of the obtained ability and difficulty estimates increases when the children behave according to the scoring rule. Question marks do not provide clear information about children's abilities. The most accurate ability estimates can be computed for children that put in as much effort as they can and respond as soon as they think they have come up with the correct answer. Ultimately, accurate ability estimates benefit the adaptivity of the system.

1.5. Minimum Toil Time

In order to prevent question mark misuse, we designed an A/B test to test whether a straightforward delay on the appearance of the question mark button would promote more effortful learning. For children that do not directly know the correct answer to an item, this delay can be seen as the minimum required toil time. We expect that those children will resort to more effortful strategies. After all, fast guesses are relatively expensive (an incorrect guess results in a punishment), and effortless waiting until the question mark becomes available costs time and decreases the potential reward. Following this reasoning, we expect children in a toil time condition to use the question mark button less frequently.

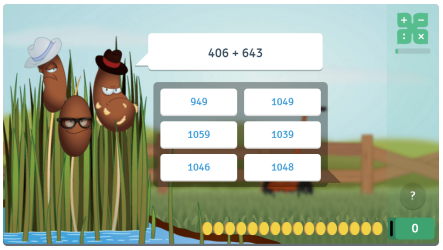
2. Methods

2.1. Experimental Domains

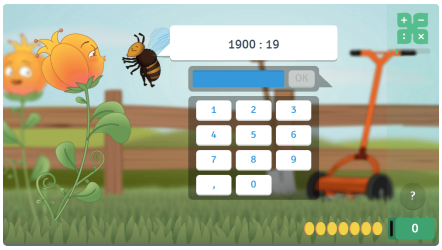
The experiment was performed in three separate game domains: addition, division, and one-two-three. One-two-three is an implementation of the popular logical reasoning game Set (e.g., Nyamsuren & Taatgen, 2013). Figure 2 shows an example item from each of these domains. In all three domains, a one game session contained ten items, after which the child was given the opportunity to choose the same or a different domain. In the addition and division domain each item had a deadline of 20 seconds, whereas one-two-three had a deadline of 30 seconds. Also, by default the former domains were available to all children, whereas the latter only became available after a child had sufficiently practised the base domains (this default setting could be changed by individual teachers for individual children). Finally, addition items had a multiple-choice format, whereas division and one-two-three items were open-ended.

2.2. Participants

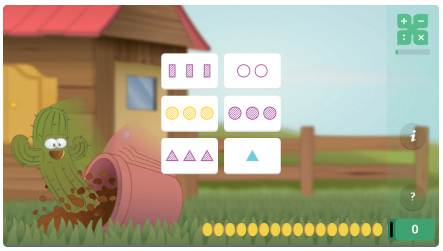
A total of 107,979 Math Garden users participated in the experiment, mostly children aged 4 to 12. Math Garden is used in ecological settings, at school and at home, on different devices, and during the whole day and week, but mostly during school hours. Children that indicated that they did not want to be part of the scientific research done in Math Garden were excluded from the analyses. The procedure was approved by the department's Ethics Review Board.



(a) Example of an item from the addition domain.



(b) Example of an item from the division domain.



(c) Example of an item from the one-two-three domain.

Figure 2: Experimental domains.

2.2.1. Allocation

Participants were randomly distributed across the four conditions[2]: the question mark button was either active (control) or greyed out and inactive for 3, 6, or 9 seconds. Randomization was done separately within each game domain. Figure 3 shows the visualizations of an active and an inactive question mark button.

2.2.2. Exclusion

The intervention relied on the CSS property *pointer-events*[3], which is not supported by some older Internet browsers. In all conditions, we excluded all children that used an incompatible browser ($n = 9,665$). Browsers and browser versions were recovered from the user agent id's that are recorded with each response, using the R implementation of *ua-parser*[4]. Nonetheless, a manipulation check revealed that 39 children with seemingly compatible browsers did have question mark responses before the question mark delay ended. As there is no reason to believe that the responses from children that used an incompatible

[2]The user id's were transformed using a bitwise right shift of 0 in the addition domain, of 2 in the division domain, and of 4 in the one-two-three domain. We then used a modulus to transform each id into one of the four conditions.

[3]`https://developer.mozilla.org/en-US/docs/Web/CSS/pointer-events`

[4]`https://github.com/ua-parser/uap-r`

(a) Active question mark button.

(b) Greyed out and inactive question mark button.

Figure 3: Visualizations of the question mark button.

browser relate in any way to the objective measures of this study (e.g., question mark use), we did not exclude these users from the analyses. We neither expect that the remaining 59 illegal responses (55 in the addition domain, 1 in the division domain, and 3 in the one-two-three domain) from those 39 children will have any substantial effects on the outcomes of the study.

2.2.3. Cross-validation & Peer Review

The huge scale of the experiment allowed us to cross-validate the effects. Moreover, in consultation with the journal editor, we followed a novel procedure to further improve the reliability of the results. We randomly selected half of the participants for the performed analyses (practice set; $n = 50,433$, excluding participants with an incompatible browser). The provisional report, which was solely based on the analyses on the practice set, was then subjected to formal peer review. After acceptance by the editor and reviewers, the results were verified on the other half of the participants (test set; $n = 50,267$, excluding participants with an incompatible browser). In this final report, we additionally report the results from the test set, but only if these deviate from the results from the practice set. This procedure, in the spirit of pre-registration[5], ensures that the methods need to be reviewed and assessed independent of the results, and that possible capitalization on chance during the analysis and review phase is corrected for by the cross-validation.

2.2.4. Distribution

In Table 1 we summarize the number of participants for different selections of the data, excluding participants with an incompatible browser. Be aware

[5]https://www.apa.org/science/about/psa/2015/08/pre-registration.aspx

that children can be in different conditions for different domains.

domain	condition	practice set	test set
addition	no delay	11866	11739
addition	3s delay	11889	11661
addition	6s delay	11600	11794
addition	9s delay	11740	11714
division	no delay	5636	5763
division	3s delay	5696	5584
division	6s delay	5594	5675
division	9s delay	5549	5622
one-two-three	no delay	7160	7012
one-two-three	3s delay	7015	7158
one-two-three	6s delay	7261	7060
one-two-three	9s delay	7134	7040

Table 1: Distribution of participants across domains and conditions.

2.3. Duration

The experiment was performed in 2016, from March 16 to June 22, spanning a total of 14 weeks. The period is a multiple of weeks to eliminate day-of-the-week effects.

2.4. Software

Analyses were performed using *R* (R Core Team, 2016) and *RStudio* (RStudio Team, 2015). Figures were created with the *R* package *ggplot2* (Wickham, 2009).

3. Results

We used linear regression analyses, with dummy variables for the conditions, to discern the effects of the different question mark delays. First, we evaluated the decrease in question mark use and made sure the delay does not affect engagement. Second, we evaluated the speed and accuracy of substitute responses to the question mark. We report standardized beta's, such that the relative strengths of the effects can be evaluated.

3.1. Question Mark Delay Decreases Question Mark Responses

First, we evaluated the decrease in question mark responses, and thus in problem skipping. In Figure 4 we show the weekly proportions of question mark responses, averaged across participants and difficulty levels. We also show how these differ across the experimental domains.

A visual inspection of Figure 4 clearly reveals a structural decrease in the proportions question mark responses with increased question mark delay. For instance, if in the addition domain the question mark button is not delayed, children tend to skip roughly 10 to 12% of the problems. With a 3 seconds

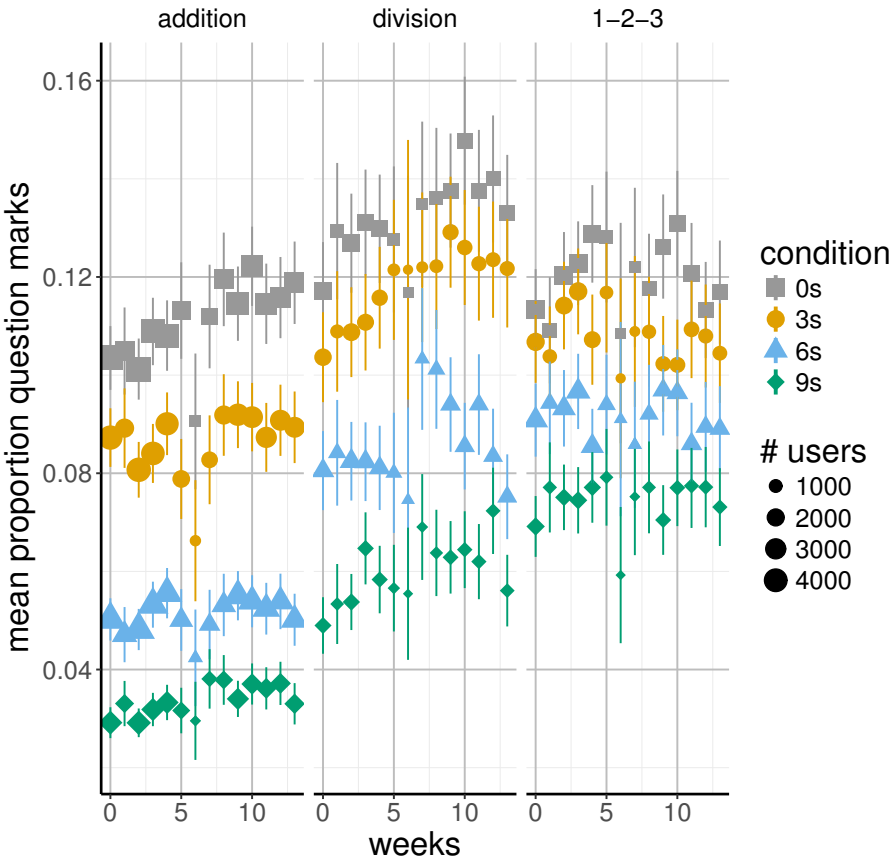


Figure 4: Average proportions of question mark responses across participants and difficulty levels, by week. Panels represent domains. Error bars represent 95% confidence intervals.

delay this percentage is reduced to roughly 8 to 10%, and with a full 9 seconds delay only roughly 3 to 4% of the problems is skipped. Interestingly, these effects seem decidedly smaller in the one-two-three domain. We'll return to this issue in the Discussion section.

We used linear regression analyses with backward difference coding in order to find the effects of the additional increases in question mark delay. Thus, the 3 seconds delay is compared to the control, the 6 seconds delay is compared to the 3 seconds delay, and the 9 seconds delays is compared to the 6 seconds delay. The analyses confirm the differences in question mark use across conditions. Table 2 shows that each additional question mark delay adds up significantly in decreasing the proportion of question marks used (all $p < .01$). In the test set, these results were confirmed, although the 3 seconds delay in the one-two-three domain was found to significantly decrease the proportion of question marks used with $p = .027$.

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.000	0.004	-0.000	1.000
addition	3s delay	-0.061	0.005	-13.214	0.000
addition	6s delay	-0.094	0.005	-17.492	0.000
addition	9s delay	-0.042	0.005	-9.081	0.000
division	(Intercept)	0.000	0.006	0.000	1.000
division	3s delay	-0.039	0.007	-5.512	0.000
division	6s delay	-0.074	0.008	-9.028	0.000
division	9s delay	-0.060	0.007	-8.550	0.000
one-two-three	(Intercept)	0.000	0.005	0.000	1.000
one-two-three	3s delay	-0.018	0.006	-3.002	0.003
one-two-three	6s delay	-0.037	0.007	-5.244	0.000
one-two-three	9s delay	-0.042	0.006	-6.892	0.000

Table 2: Linear regression results for experimental differences in the proportion question mark responses, separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the preceding delay, in order to find the effects of the additional increases in question mark delay. Standardized betas are reported ('estimate').

3.2. Question Mark Delay has No Adverse Effects on Time on Task

Preferably, the question mark delay intervention has no adverse effects on engagement. When children consider the delay annoying, they might decide to practice less. To rule out the possibility of such an adverse effect, we checked whether the question mark delay conditions differed with respect to the readily available proxy-measure time on task. First, time on task (in minutes) was computed by summing the response times separately for each participant during the experimental period. We expected no differences in time on task between conditions, and thus compared each intervention condition (i.e., 3, 6, and 9 seconds question mark delay) directly with the control (no delay).

The results of the linear regression analyses are summarized in Table 3. No significant differences were found, except for the 9 seconds delay conditions in

the addition and division domains. Both effects suggest that with a 9 seconds delay children spend more rather than less time on the addition and division tasks. However, we are reluctant to give these effects too much weight, as the modest standardized beta's of 0.011 and 0.027 point to negligible effects that possibly originate from the huge amount of power of the study. In the test set, these results were confirmed, as no significant differences were found.

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.000	0.005	-0.000	1.000
addition	3s delay	0.005	0.006	0.965	0.335
addition	6s delay	0.006	0.006	1.063	0.288
addition	9s delay	0.011	0.006	1.992	0.046
division	(Intercept)	0.000	0.007	0.000	1.000
division	3s delay	0.015	0.008	1.785	0.074
division	6s delay	0.006	0.008	0.720	0.472
division	9s delay	0.027	0.008	3.327	0.001
one-two-three	(Intercept)	-0.000	0.006	-0.000	1.000
one-two-three	3s delay	-0.005	0.007	-0.647	0.518
one-two-three	6s delay	-0.009	0.007	-1.204	0.229
one-two-three	9s delay	-0.008	0.007	-1.118	0.264

Table 3: Linear regression results for experimental differences in time on task, separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the control condition (no delay), as no differences in time on task are expected. Standardized betas are reported ('estimate').

3.3. Substitute Responses are Primarily Slow

Following up the shown decrease in question mark responses, we investigated how children substitute their responses. Naturally, to know exactly which responses are substitutes for question mark responses requires counterfactual information, but the speeds and accuracies of substitute responses can nonetheless be estimated by assessing the changes to the overall response times and accuracies. In Figure 5 we show the weekly response time means, averaged across participants, difficulty levels, and response types. We also show how these differ across domains.

A visual inspection of Figure 5 reveals structural differences between conditions. In the addition and division domains, the mean response times clearly increase with increased question mark delay. For instance, if in the division domain the question mark button is not delayed, children respond in roughly 7.3 to 7.5 seconds. With a 3 seconds delay the responses slow down to roughly 7.5 to 7.8 seconds, and with a full 9 seconds delay children respond in roughly 8 to 8.2 seconds. Interestingly, this increase is decidedly less clear in the one-two-three domain.

We used linear regression analyses with backward difference coding in order to find the effects of the additional increases in question mark delay. The analyses confirm the observed differences. Table 4 shows that each additional

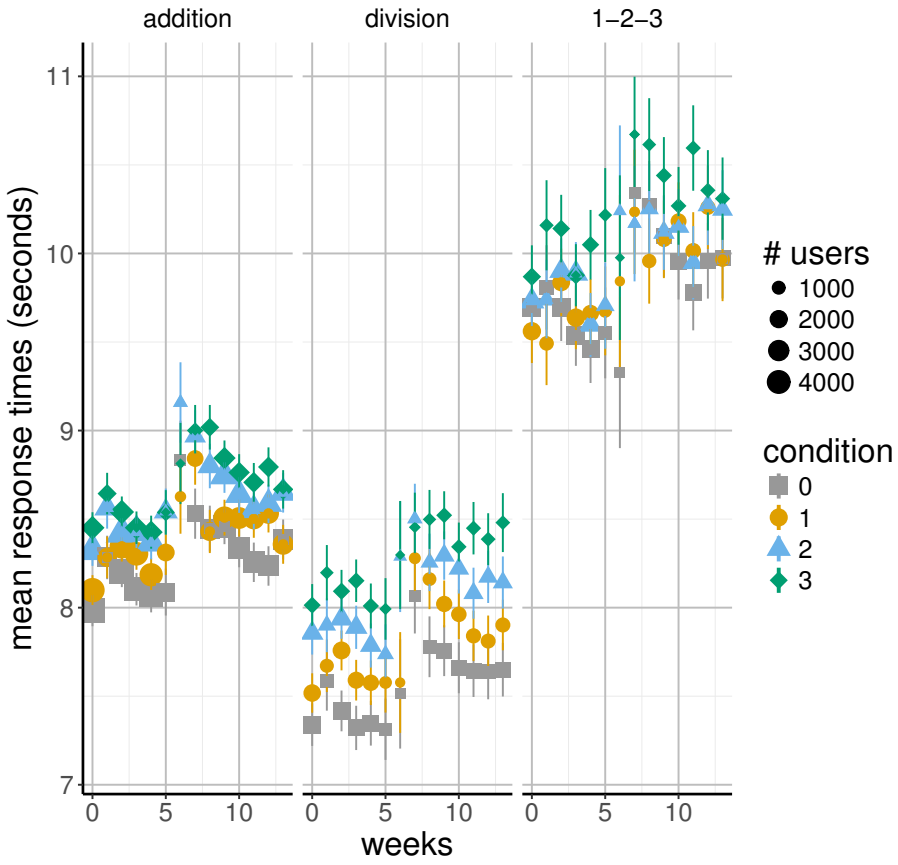


Figure 5: Average response times across participants, difficulty levels, and response types, by week. Panels represent domains. Error bars represent 95% confidence intervals.

question mark delay adds up significantly in increasing the response time, except for the 3 and 6 seconds delay in the one-two-three domain, and a decrease in response times for the 9 seconds delay in the addition domain. This finding provides some evidence that the question mark delay is indeed, at least partly, used for toil time, and that fast question marks are not solely substituted by fast guesses.

In the test set, these results were largely confirmed. In all domains, each additional delay contributed to an increase in response times (all $p < .001$, except for the 3 seconds delay in the one-two-three domain, with $p = .040$). Contrary to the results in the practice set, the 9 seconds delay in the one-two-three domain resulted in a small decrease in response times ($p < .001$).

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.022	0.001	-38.339	0.000
addition	3s delay	0.002	0.001	3.534	0.000
addition	6s delay	0.015	0.001	19.164	0.000
addition	9s delay	-0.003	0.001	-4.552	0.000
division	(Intercept)	-0.161	0.001	-206.909	0.000
division	3s delay	0.005	0.001	5.018	0.000
division	6s delay	0.019	0.001	17.051	0.000
division	9s delay	0.021	0.001	21.630	0.000
one-two-three	(Intercept)	0.131	0.001	138.387	0.000
one-two-three	3s delay	-0.001	0.001	-1.011	0.312
one-two-three	6s delay	0.001	0.001	0.565	0.572
one-two-three	9s delay	0.020	0.001	17.021	0.000

Table 4: Linear regression results for experimental differences in response times (in seconds), separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the preceding delay, in order to find the effects of the additional increases in question mark delay. Standardized betas are reported ('estimate').

3.4. Substitute Responses are Primarily Incorrect

Additionally, we investigated the accuracy of the substitute responses. In Figure 6 we show the weekly response accuracy proportions, averaged across participants and difficulty levels. We also show how these differ across domains. We removed all question mark responses, since a change in question mark responses necessarily changes the proportions correct and incorrect responses with respect to all responses, yet we are interested in the mutual proportions between correct and incorrect responses.

A visual inspection of Figure 6 seems to reveal a decrease in the proportions correct responses with increased question mark delay, at least for the addition and division domains. For instance, if in the addition domain the question mark is not delayed, children tend to solve roughly 70 to 72% of the problems. With a full 9 seconds delay children solve roughly 67 to 69% of the problems. This decrease is much less clear in the one-two-three domain.

We used linear regression analyses with backward difference coding in order to find the effects of the additional increases in question mark delay. The

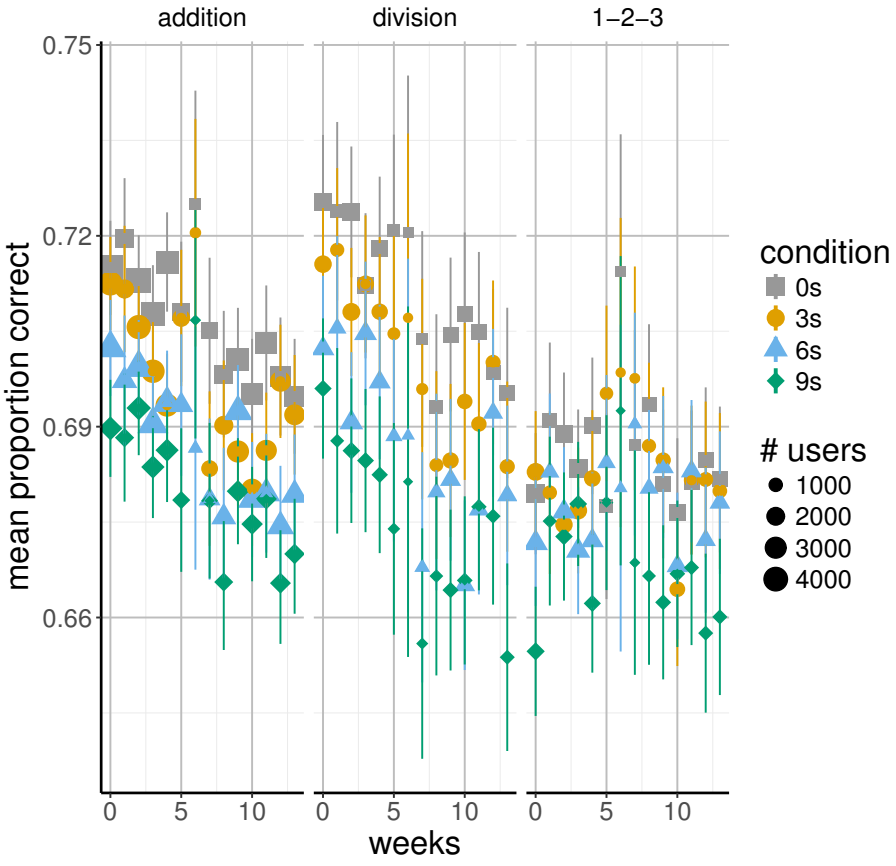


Figure 6: Average proportions correct responses *excluding question mark responses*, across participants and difficulty levels, by week. Panels represent domains. Error bars represent 95% confidence intervals.

analyses confirm the observed diffuse effects. Table 5 shows that each additional question mark delay adds up significantly in decreasing the proportion of correct responses in the addition domain and the 3 and 6 seconds delay in the division domain, but not in the 9 seconds delay in the division domain and in the one-two-three domain. This finding tentatively points out that although children take more time to formulate a response, the response is often incorrect.

In the test set, the tentativeness of these results is further emphasized. The results were confirmed for the addition domain. However, in the division domain, the 3 seconds delay did not differ significantly from the 0 seconds delay ($p = .717$), whereas the 9 seconds delay did differ significantly from the 6 seconds delay ($p = .011$). And in the one-two-three domain, both the 3 seconds delay and 9 seconds delay differed significantly from respectively the 0 seconds delay ($p = .008$) and 6 seconds delay ($p = .001$).

domain	term	estimate	std.error	statistic	p.value
addition	(Intercept)	-0.000	0.004	-0.000	1.000
addition	3s delay	-0.011	0.005	-2.372	0.018
addition	6s delay	-0.028	0.005	-5.065	0.000
addition	9s delay	-0.018	0.005	-3.807	0.000
division	(Intercept)	-0.000	0.006	-0.000	1.000
division	3s delay	-0.019	0.007	-2.609	0.009
division	6s delay	-0.026	0.008	-3.168	0.002
division	9s delay	-0.013	0.007	-1.837	0.066
one-two-three	(Intercept)	-0.000	0.005	-0.000	1.000
one-two-three	3s delay	0.001	0.006	0.143	0.887
one-two-three	6s delay	-0.012	0.007	-1.689	0.091
one-two-three	9s delay	-0.009	0.006	-1.531	0.126

Table 5: Linear regression results for experimental differences in proportion correct responses *excluding question mark responses*, separately for the addition domain, division domain, and one-two-three domain. The question mark button was activated with no delay, 3 seconds delay, 6 seconds delay, or 9 seconds delay. Results show the difference with the preceding delay, in order to find the effects of the additional increases in question mark delay. Standardized betas are reported ('estimate').

4. Discussion

The question mark delay intents to require children to exert at least some minimum amount of effort, and can thus be seen as the minimum amount of required toil time. The results clearly demonstrate that the delay indeed ensures a decrease in the use of the question mark. Rather than waiting for the question mark button to appear, children seem to attempt the item more frequently. Also, the toil time does not seem to diminish engagement as the delay does not affect the amount of time children spent on solving items.

Naturally, whether the question mark delay indeed supports active and effortful learning is not that easily concluded. Children may for instance substitute their fast question mark responses for a fast guessing strategy. From a theoretical point of view this is unlikely however. In Math Garden, a fast

guessing strategy is risky since especially *fast* incorrect answers are punished with a substantial subtraction of coins, and moreover particularly risky in domains with open-ended question such as the division and one-two-three domains. Moreover, to exclude the possibility of fast guesses, we showed that substitute responses are, although primarily incorrect, also primarily slow.

Looking into the decrease in question mark responses across different domains, one thing to notice is the seemingly smaller decrease in the one-two-three domain as opposed to the addition and division domains. Interestingly, also the substitute responses seem to show a different pattern for this domain. As opposed to the responses in the addition and division domains, the response times do not necessarily increase (except for the 9 seconds delay), and the proportion (in)correct responses is not influenced by the delay.

Multiple explanations can account for this possible difference. First, whereas in the addition and division domains children may resort to memorization strategies, in the one-two-three domain, a complex logical reasoning task, more effortful strategies are already demanded. In this case it is expected for the question mark delay to have less of an effect.

Moreover, since by default the one-two-three domain is only unlocked after frequent practice in the base domains, we might be looking at a highly motivated subset of children that are already less likely to quickly resort to effortless strategies. And lastly, the one-two-three domain has a time limit of 30 rather than 20 seconds. Possibly, since the toil time is thus relatively shorter, it could make the effect less pronounced.

Taking the above together, the strength of the intervention is expressed in its broad applicability. The minimum required toil time ensures an increase in more active and effortful practice, regardless of the complexity of the task, the response mode, or the task length. Moreover, it does not invoke other gaming strategies, such as fast guesses. And finally, it is a so-called soft intervention: it does not prevent children from skipping problems and thus from self-regulating their learning, but nudges children towards a more effortful and more effective learning strategy.

5. Conclusions

Delaying the option to skip problems in (online) learning can be beneficial. Especially in cases where students are being challenged and an enduring effort is requested, it can be a helpful nudge to exert at the very least some minimum amount of effort. Of course, to safely and conclusively generalize this finding it must be examined on other platforms and in a variety of situations. Also, establishing whether the increased effort results in actual learning gains is an important question that remains open. Nevertheless, three major strengths of the methodology used in the current study are important to highlight.

First, the current paper demonstrates some of the advantages of the A/B testing methodology in the learning domain. Importantly, it allows researchers to evaluate learning interventions on large groups of learners in their natural

learning environment. We can use experiments to evaluate causal effects of changes to the system. The readily available data taps into many different aspects of the complex dynamic system of learning, and can thus reveal related patterns such as adverse or beneficial side-effects. Successful interventions can have a large and direct impact: on the basis of this study Math Garden implemented a question mark delay of 25% of a domain's deadline (e.g., 5 seconds for domains with a deadline of 20 seconds), potentially benefiting over 150.000 children. Whereas likewise, adverse interventions can be uncovered upfront rather than blindly implemented.

Second, not only does the large scale of online learning drastically improve the reliability and impact of the interventions, it enables cross-validation of the findings. We exploited this fact in order to further increase the reliability of the study, by using a novel procedure in the spirit of pre-registration. As explained in the Methods section, the findings were only verified on the test set after the editor gave formal approval for publication. This way, we ensured that the research is assessed on the basis of the methods, and we prevented capitalization on chance in both the analysis and review phases.

Finally, findings from online experiments may not only help improve online learning, but the obtained insights may as well validate traditional (offline) interventions and feed back into the various sciences they were drawn from. Generalizability may naturally vary from study to study, but A/B tests can be used for triangulation and usually have great ecological validity. Moreover, it tackles many of the problems encountered in traditional educational experimentation, most importantly the often impracticable double-blind procedure.

Acknowledgements

This research was performed in close collaboration with the online computer adaptive practice environment Math Garden. The problem-skipping strategy was noticed by the Math Garden team and contact with Math Garden went through their Chief Scientific Officer (last author). The experiment was designed by the authors and implemented by Math Garden's development team. The principal researchers are independent of Math Garden (first and second author).

References

Athey, S., & Imbens, G. (2016). The state of applied econometrics - causality and policy evaluation. *arXiv*, . arXiv:1607.00699v1.

van den Bergh, M., Schmittmann, V. D., Hofman, A. D., & van der Maas, H. L. J. (2015). Tracing the development of typewriting skills in an adaptive e-learning environment. *Perceptual and Motor Skills*, *121*, 727–745. doi:10.2466/23.25.pms.121c26x6.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. doi:10.1146/annurev-psych-113011-143823.

Borghans, L., de Wolf, I., & Schils, T. (2016). Experimentalism in Dutch education policy. In *T. Burns, & F. Kster (Eds.), Governing Education in a Complex World*. OECD Publishing. doi:10.1787/9789264255364-en.

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition*, *149*, 40–55. doi:10.1016/j.cognition.2016.01.004.

Deaton, A., & Cartwright, N. (2016). *Understanding and Misunderstanding Randomized Controlled Trials*. Technical Report National Bureau of Economic Research. doi:10.3386/w22595.

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*, 470–497. doi:10.1007/s40593-014-0024-x.

Jansen, B. R., Hofman, A. D., Savi, A., Visser, I., & van der Maas, H. L. (2016). Self-adapting the success rate when practicing math. *Learning and Individual Differences*, *51*, 1–10. doi:10.1016/j.lindif.2016.08.027.

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*, 1813–1824. doi:10.1016/j.compedu.2011.02.003.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633. doi:10.1007/s11336-012-9288-y.

Murray, T., & Arroyo, I. (2002). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In *Intelligent Tutoring Systems* (pp. 749–758). Springer Berlin Heidelberg. doi:10.1007/3-540-47987-2_75.

Nyamsuren, E., & Taatgen, N. A. (2013). Set as an instance of a real-world visual-cognitive task. *Cognitive Science*, *37*, 146–175. doi:10.1111/cogs.12001.

Olson, D. R. (2004). The triumph of hope over experience in the search for "what works": A response to Slavin. *Educational Researcher*, *33*, 24–26. doi:10.3102/0013189x033001024.

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, *31*, 459–470. doi:10.1016/s0883-0355(99)00015-4.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc. Boston, MA. URL: http://www.rstudio.com/.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*, 15–21. doi:10.3102/0013189x031007015.

van der Ven, S. H., Klaiber, J. D., & van der Maas, H. L. (2016). Four and twenty blackbirds: How transcoding ability mediates the relationship between visuospatial working memory and math in a language with inversion. *Educational Psychology*, (pp. 1–24). doi:10.1080/01443410.2016.1150421.

Vygotskii, L. S. (1978). *Mind in Society*. Harvard University Press.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: http://ggplot2.org.

Williams, J. J., Li, N., Kim, J., Whitehill, J., Maldonado, S., Pechenizkiy, M., Chu, L., & Heffernan, N. (2014). The MOOClet framework: Improving online education through experimentation and personalization of modules. *SSRN Electronic Journal*, . doi:10.2139/ssrn.2523265.