New evidence on the effect of computerized individualized practice and instruction on language skills

J. Ghysels | C. Haelermans

Top Institute for Evidence-Based Education Research (TIER), Maastricht University, Maastricht, the Netherlands

Correspondence
Carla Haelermans, Top Institute for Evidence-Based Education Research (TIER), Maastricht University, PO Box 616, 6200 MD, Maastricht, the Netherlands.
Email: carla.haelermans@maastrichtuniversity.nl

Abstract
This paper provides new evidence on the effect of computerized individualized practice and instruction on language skills, more specifically on spelling. An individually randomized experiment among 7th grade students in the Netherlands is developed to study the effect of an adaptive digital homework tool on spelling performance. Using an instrumental variable approach to control for actual use of the digital tool, we show that there are small positive effects of practicing with an adaptive digital tool for spelling for 7th grade students. Effects are largest for low-performing students.

KEYWORDS
computerized practice, individualization, language, randomized field experiment, secondary education, spelling

1 | INTRODUCTION

Basic language skills are of major importance in daily life. However, many students do not have sufficient basic language skills (Commissie Meijerink, 2008), and of the language skills, spelling skills are most often lacking. Many scholars conclude that individual differentiation is the key to higher student performance (e.g., Hattie, 2009), especially for easy to automatize topics, but traditional classroom settings only partly allow schools to differentiate their teaching between individual students, which contributes to the deficiency of these skills. The combination of computer use in education, the need for individualization in the learning process, and the decrease in language skills have led to the development of individualized information technology (IT) tools aimed at developing these skills. Accordingly, many schools have started using individualized IT tools to increase students’ language skills. Individualized IT tools focus on an individual learning path for the student, adapting the exercises available for the student to the skills that he or she is lacking. However, schools are introducing IT in very diverse ways, and much of its potential effectiveness depends on the way schools’, teachers’, and students’ use the IT tool (Haelermans, 2017). It is therefore maybe not surprising that the literature shows very mixed results from the studies in which IT tools or IT in education in general are analysed.

Educational or education economics studies of IT tools (e.g., computer-assisted instruction [CAI] or intelligent tutoring systems [ITS]) offer mixed results of the effects of IT tools. It is important to mention that there are not many studies that show causal effects of using digital tools to practice math or language skills, only few studies use experimental or quasiexperimental research methods to study the matter. In general, these causal studies show that CAI and ITS give positive results if applied to mathematics and arithmetics (for an overview, see Haelermans & Ghysels, 2017). However, in the literature on language tools, we only see few experimental studies, and none of them have found positive effects. Borman, Benson, and Overman (2008), for example, conducted a randomized field trial to study the effects of a computer-based training (FastForWord) on language skills and conclude that this program did not help students improve their language skills. Similar conclusions are drawn by Rouse and Krueger (2004) and by Given, Wasserman, Chari, Beattie, and Eden (2008), who both studied the effectiveness of the same program. Potocki, Ecalle, and Magnan (2013) also did not find any significant results when using a randomized controlled trial to study the effects of a computer-assisted comprehension training. Interestingly, the only
source that concludes that small positive effects are likely is the meta-
analysis by Cheung and Slavin (2012). This finding cannot be explained
directly, as the article does not show what the effect sizes are per
study, nor how they estimate the overall small but positive effect.
Furthermore, an earlier meta-analysis by Slavin, Lake, Davis, and
Madden (2011) did not show any results.

However, it is possible that the previous experimental studies
focused too much on language skills in general or on the average
student. A previously carried out pilot study with the same online
practice tool gave the suggestion that CAI or ITS tools might only be
effective for practicing skills that are easy (easier) to automatize.
Furthermore, an earlier study with the same program for mathematics
showed that there were heterogeneous effects (Bartetelet, Ghysels,
Groot, Haelermans, & Maassen van den Brink, 2016), leading us to
believe that this might also be the case for language skills.

Therefore, in this paper, we report the results of an experiment
that also studies the effect of an IT tool for education. We study the
effectiveness of the student’s use of an IT tool as homework for the
language skill spelling. Using an instrumental variable (IV) approach
to control for actual use of the IT tool and to analyze the effect of actually
using the tool while controlling for selection of which student decides
to use the tool, we show that there are small positive effects of prac-
ticing with an adaptive IT tool for spelling for seventh grade students.
We furthermore show that the effects are heterogeneous, as only the
low-achieving students benefit from the program, whereas the middle-
and high-achieving students do not.

The contribution of this paper to the literature is as follows: (a) this
paper is one of the few to use an individually randomized experiment to
study the effect of practicing with an adaptive IT tool for spelling, one of
the easier to automatize aspects of language. It is also the first study to
causally analyze the effects on language skills of this specific IT tool in
the Netherlands. The individually randomized experiment implies that
class and teacher effects can be ruled out, and the randomized experi-
ment implies that the effects are causal; (b) the paper explicitly includes
heterogeneous effects, in which we show that these effects are not nec-
essarily present for all students; and (c) in additional analyses, we look at
the intensity of the treatment; that is, we do not just have a dummy
variable to measure who had access to the tool, but we also know if,
when, what, how often, and how long students used the practice tool.
This allows us to better analyze what is going on in the treatment.

The remainder of this paper is structured as follows: Section 2
presents the program characteristics of both the experiment and
the program under study. Section 3 presents the data and the
identification strategy, and Section 4 presents the empirical strategy.
In Section 5, the baseline results and the regression results are
discussed, and Section 6 presents the additional analyses. Section 7
concludes the paper and discusses the findings.

2 | PROGRAM CHARACTERISTICS

2.1 | The experiment

The intervention in this study is that treatment students practice with
the spelling modules of the computerized adaptive practice homework
tool. The intervention is meant to practice the specific automatization
skill for spelling, but this is a skill that can be more generalized as it is
crucial to other aspects of literacy as well. Control students also
practice with the online tool and are exposed to the tool for the same
average number of minutes each week, but their accounts are techni-
cally disabled for using the spelling modules. They have access to the
vocabulary modules instead. Teachers would not know which students
practiced with which exact modules, so that they will treat all students
the same, and no different treatment is to be expected from parents
either, as all students had to practice with the online tool during the
whole school year, the only difference was that topics practiced were
different for treatment and control group in the first semester, which
was reversed in the second. Besides the intervention, all students have
regular language classes in which all aspects of language are being
taught. All students have access to all other modules of language, such
as text comprehension and grammar. A pilot study with the same pro-
gram gave the indication that the online practice tool would possibly
only be effective for modules that are easier to automatize and are
not so much related to the method used and pace at school. Therefore,
the modules of spelling and vocabulary were chosen. We originally
planned to also study effect of practicing with vocabulary. However,
there was such little variation in the outcome measures that these
analyses would not have been reliable.

2.2 | Characteristics of program under study

The purpose of the computerized adaptive practice tool is to help
students practice their language skills, while being able to individualize,
and give users direct feedback (Muiswerk, 2013). Although the program
is mainly being used in the Netherlands, it also has an international
version and is used by several international schools both in Europe
and other parts of the world. The program is interactive and person
specific. Students work at their own level and get those exercises that
will help them improve the subspects of language they are not knowl-
edgeable in yet, whereas some exercises are meant to keep up their
already acquired knowledge. The school uses this tool to make sure
each student achieves the highest possible level of language, given
his/her abilities, and maintains the level achieved. It offers all students
online access to the tool for use after school hours, at home. In this
sense, it is a supplemental program in the categorization of Cheung
and Slavin (2012).

First year secondary students (seventh grade) make a language
skills pretest at the start of the school year, in September. This test
determines their level of different subspects of language, which in
turn determines the types of exercises they have to start practicing
with at home. Note that a student questionnaire showed that all
students have a computer or device at home that they can use for
practicing. At regular intervals, students make a short computer test
at school to determine for which exercises their skills are still lacking
and for which exercises their knowledge level is good enough for the
moment. After every test, the type and level of exercises a student
can access are adjusted to their new skill level.

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

Language teachers are supposed to motivate students to practice with the computerized instruction tool at home and to check whether students indeed practice the suggested number of minutes. However, not all language teachers are similarly convinced by the usefulness of the tool, and hence, teachers show different behaviour in motivating and checking up on their students. On the one hand, only few of the teachers actually log into the system to check their students’ practice behaviour, but on the other hand, almost all teachers regularly motivate their students to practice the suggested 30 min per week. The program intensity is therefore low, in the categorization of Cheung and Slavin (2012). Despite the instruction of 30 min per week, of which 15 min with spelling (or vocabulary for control students), practice behaviour differs largely among students, as there is no sanction or test they fail if they do not practice.

The language skills are measured using digital standardized tests, which are written by all seventh grade students at T0 and T1 (see Figure 1). These are standardized validated tests, and these tests are based on other nationally validated tests. The reliability (Cronbach’s alpha scores of between .79 and .92) and validity of these tests are analysed yearly by the tool developer, based on norm data of several participating schools (Schijf & Schijf, 2014). The tests are external to the practice exercise tool and do not contain any of the exercise questions. The tests measure whether students have mastered the required national language level they are supposed to have, given their age and given the fact that they finished primary school (called “reference level”). Note that this school uses quite some digital material and digital tests for other courses as well, implying that control group students also had a lot of exposure to digital teaching materials and are therefore not disadvantaged in this respect.

All students, both treatment and control groups, practiced with digital multiple-choice language assignments in the testing program in the week before the pretest (T0) was administered, to make sure they knew what to expect when writing the pretest and to get acquainted with the testing environment. The test contains multiple choice questions. Test scores can range from 0 to 100, 100 being the absolute maximum.

The school under study, Dendron College, has about 2,000 students in total and is—a mid-sized school for secondary education (junior high and high school). Dendron College offers secondary education in all tracks and is tracking students from the first year on in several prevocational, general higher, and preuniversity tracks. Dutch secondary education has a tracking system from seventh grade on, with three different tracks (prevocational education, which consists of four subtracks where Level 1 is the lowest [mainly practical] track and Level 4 the highest [mainly theoretical] track, general higher education, and preuniversity education).

In school year 2013/2014, there were 377 students in 14 first year classes (equivalent to seventh grade in the US), ranging from the lowest prevocational track to the preuniversity track. First, 14 students were excluded from the analysis, as they were ill during either the pre- or the posttest, and therefore did not have test scores for both tests. Next, we lose 13 students, because these students do not have a primary school ability test score, the measure that is our second most important control variable after the pretest, and that we use to study heterogeneous effects. This leaves us with 350 students.

The age of the students in the experiment ranges from 11 to 13 (differences are mainly due to grade repetition), and about 55% is girls. The average score on the primary school ability test is 536, where the minimum is 509 and the maximum is 550. Note that the scores on this
test have a theoretical range from 500 to 550 and that not all students have a score on this test, as primary schools can decide whether they use it or not. However, almost 90% of primary schools issues this specific test. The 14 classes have nine different teachers for language. Furthermore, students have very diverse backgrounds, as they attended 27 different primary schools. As such the school is a typical representative of schools outside of the highly urbanized, central region of the Netherlands (the “Randstad”).

3.2 Identification strategy

The main problem with determining the effect of a practicing tool is the potential correlation of unobservable factors with both the practicing behaviour and the outcome variables. In this study, we use exogenous variation in the possibility to practice through an experimental setup. Students were randomized at the individual level with a random number generator where even numbers where assigned to the treatment group and odd numbers to the control group, using a stratification strategy in which class is taken into account, to ensure an approximately evenly distribution of treatment and control group within a class. The latter is done to ensure that we have a good balance of lower and higher performing students in both treatment and control groups, because the Dutch tracking system places students of similar ability levels in similar classes. Note that we did not impose an exact 50–50 division over treatment and control groups and that some students were removed from the dataset, which in the end resulted in 185 treatment students and 165 control students.

First, we check whether there is balance between treatment and control students on observable characteristics. Although balance does not guarantee that members of both groups had equal expectations regarding the development of their skills at the start of the intervention period, it can at least give a good idea whether the randomization was successful. Using t-statistics for continuous variables and dummies and chi-squared statistics for the comparison of the nominal variables, we found that there is almost perfect balance between the two groups, with no significant differences in the mean values of the observable characteristics. A joint F-test on all the characteristics also does not show a significant difference (F(66, 283) = 0.73, p = .94).

Given these results, we expect the random assignment process to have functioned well, producing an intervention and control groups that

3.3 Compliance with assignment

As it happens, not all students complied with their assignment in the experiment. Although we technically disabled the type of exercise students were not supposed to practice with, students can still decide not to practice at all with the exercise they were supposed to practice with, and some did. This means that we have three groups of students: (a) students that were assigned to the treatment group (A = 1) and also practiced with this type of exercise in the online tool (P = 1), (b) students that were assigned to the control group (A = 0) and therefore did not practice with this type of exercise in the online tool (P = 0), and (c) students that were assigned to the treatment group (A = 1), but who did not practice at all with this type of exercise (P = 0). Comparison of these three groups of students shows that compliance is not a random process. The students that did not comply are “smarter” students (higher primary school ability test score and higher pretest score), mainly boys and often the youngest child at home. Possible explanations could be that smarter children have not felt the need to do their homework properly, as they get good grades anyway and that girls listen better to their teacher if they are being told that they have to practice online as homework.

Because we are interested in the effect of actually using the tool, this selective noncompliance is likely to create (a small) bias in the estimated effect. That is, a simple comparison of the control and experiment groups reveals the intent-to-treat (ITT) effect (showing what the effect is of offering the tool) but not the average treatment effect (showing what the effect is of actually using the tool), because a specific group seems to have self-selected away from treatment. Therefore, we control for the noncompliance using an instrumental variable approach, as will be explained in more detail in Section 4. However, we also acknowledge that the problem of noncompliance is relatively small, so the ITT in itself is also already very informative.

3.4 Practice behaviour

As described above, the setup of the experiment is such that students are relatively free in deciding when, where, with what, and how much to practice, contrary to most other (relatively comparable) studies in the literature (Cheung & Slavin, 2012). Therefore, in addition to studying the effect of access to the digital practice tool, it is also interesting to look at the intensity of the treatment. Apart from that, teachers are mixed in their enthusiasm of the tool, in part because the freedom it offers to students makes it difficult to coordinate home practice learning with a regular, class-based, and fairly homogeneous instruction process. Another reason is that they simply do not like the tool and therefore refuse to remind their students about practicing. However, remember that the tool is not used in class, but as homework, and that students are also reminded to do this homework via their digital calendar, suggesting that the teacher only has a limited influence on the intervention. Nonetheless, different enthusiasm by teachers also leads to differences in usage of the tool by students and also underlines the importance to control for the teacher in our analyses.

Below, we use the number of minutes practiced per week, which is determined by subtracting the weeks of school holidays from the total number of regular school weeks during the experimental period (i.e., excluding test weeks, leaving 16 weeks). We decided to subtract school holidays because students usually do not make homework in these weeks and would most likely also not practice with the tool. However, this means just rescaling, which will not influence the results in itself.

Figure 2 and Table 1 show the descriptive statistics of the practice behaviour of students. Figure 2 shows the spread of number of times practiced and the practice minutes over the weeks, and Table 1 shows...
the number of students who practiced, the total minutes practiced, the number of times students practiced on average, and the average practice minutes per week over the full 16 weeks. Figure 2 shows that the average time dedicated to spelling practice is fairly similar across weeks. It hovers between 5 and 10 min per week, with the exception of Week 5, where nine students practiced more than 20 min per week on average. Furthermore, it can be seen that between 40 and 100 students practiced each week. Although overall almost all students practiced at least once, this figure shows that in the best weeks (Weeks 11–14, when teachers were specifically reminded again to remind their students to practice in the tool), only about 100 out of about 180 students actually practiced. Note that the fifth week was the autumn break, and Weeks 15 and 16 were Christmas break. So nine students were very active in their holidays, but as most students were not active, these weeks were not taken into account in the calculation for average minutes per week. The between student variation of practice behaviour across weeks probably also causes the difference between the average number of minutes practiced per week over the total period. Some students practice a very high number of minutes in 1 week but do not practice each week, whereas the others practice each week but only for a low number of minutes. Note that we do not see any differences in average practice behaviour between boys and girls and low-, middle-, and high-performing students.

As said before, there are (very) large differences in practice behaviour between students. The average amount of time students practiced was 7 min per week for spelling, ranging from an average of 1 to 35 min per week. Note that practicing time is only counted if at least one exercise is finished in the episode. Over the total research period, a student practiced on average about 114 min for spelling, distributed over 40 different episodes, so on average a little more than twice per week. A questionnaire among seventh grade students in an earlier study on the same program showed that students write on average 15 to 45 min of Dutch homework per week. Therefore, an additional 7 min is an increase of Dutch homework time of 15% to almost 50%. So although the 7 min is only (less than) half of the supposed practice, it is still a considerable increase in language homework time for students.

4 | EMPIRICAL STRATEGY

Given the individually randomized setup of the experiment, we want to identify the effect of using the interactive digital practice tool on spelling skills. We do so by using both a simple t-statistic and a linear regression in which we can control for background characteristics of the student, such as age, gender, and ability variables.

However, the experiment provides students with access to the tool but can of course not ensure that students actually use the tool. As we have seen in the section on compliance with the assignment, not all students have used the tool to practice, making it technically an analysis in which we study the intent to treat, so simply the effect of offering the tool, instead of an average treatment effect in which we analyse the effect of actually using the tool.

In order to control for the actual use of the tool, we use a two-stage-least-squares (2SLS) instrumental variable approach. Here, we use the dummy that indicates the random assignment for access to the tool as an instrument for the actual use of the tool. By doing this, we ensure that we can still use the randomization to analyse causal effects, but at the same time, we analyse the more interesting question whether actual use of the tool has a positive effect on performance. For this analysis, it is important that the assignment to treatment or control group is (highly) correlated with the use of the tool, which is represented in the statistically significant and large coefficient of the

![Figure 2](https://example.com/figure2.png) Weekly practice behaviour regarding spelling [Colour figure can be viewed at wileyonlinelibrary.com]
treatment group indicator (access to the tool) in the first stage regression. This first stage regression estimates the probability that students that were randomized into the treatment group actually use the tool. In the second stage of the regression, we use the outcome of the first stage (predicted probability) to estimate the effect of using the tool. Rather than using the observed use of the tool, we now gauge the effect of predicted use, being an indicator that does not entail unobserved reasons for the use of the tool but strictly reflects the effect of the offer of the tool.

In the regressions, we add the following control variables: spelling pretest, primary school ability test score, gender, age, whether the student is the oldest child, whether the child has a stable situation at home, and dummies for religion, country of birth, and for the language teacher. Note that we do not add educational level to our analysis, as this is highly correlated with both the pretest and the primary school ability test score.

In our analyses, we standardize our outcome variables such that all the variables have a mean of zero and a standard deviation of one. This implies that differences between treatment and control group in the T-tests and the regression coefficient of the treatment or usage dummy can be interpreted as standardized effects (i.e., Cohen’s d), where 0.2 is a small effect, 0.5 is a medium effect, and 0.8 is a large effect. Because the standardization of variables is based on the whole group of students, it is possible that we see negative averages if we, for example, split up treatment and control group in Table 5. A negative value here implies that this group scores below the overall average (of 0).

Next, we study heterogeneous effects by adding interaction terms between whether a student has practiced and to which ability group the student belongs. This gives us not one but three instrumented variables (one for each ability group), where we predict the participation status for each group separately.

For the heterogenous effects, we use the primary school ability test score to split the group in three. Although we did strive for about equal groups, there actually was a rational behind the creation of the test score to split the group in three. Although we did strive for about equal groups, there actually was a rational behind the creation of the tool. This gives us not one but three instrumented variables (one for each ability group), where we predict the participation status for each group separately.

For the heterogenous effects, we use the primary school ability test score to split the group in three. Although we did strive for about equal groups, there actually was a rational behind the creation of the tool. This gives us not one but three instrumented variables (one for each ability group), where we predict the participation status for each group separately. These three indicators represent different assumptions about the mechanism underlying a potential effect of the tool. The simple participation dummy assumes that any type of participation can be helpful and focuses on an average effect, whatever the use the student made of the tool. The second indicator, number of practicing episodes, assumes that the effect of the tool varies (in a linear way) by the frequency of practicing. The last indicator stresses the intensity of the experience rather than the frequency. The overall time spent with the tool is analysed, apart from the number of episodes it is constructed of. The three indicators are not completely independent of each other, yet they do represent different dimensions of the learning mechanism underlying the didactical tool evaluated in this project.

5 | RESULTS

5.1 | Baseline results

The first results we present are the simple t-statistics of the effect of access to the tool on multiple outcome measures, for spelling, in Table 2. Note that all coefficients in Table 2 are standardized coefficients, that are to be interpreted in terms of standard deviations with a mean of zero and a standard deviation of one (calculated based on the total group of students). First, we present the test score at T0, where we expect to see no significant difference, given the random school ability score and educational level (as defined by the Dutch tracking system), implying that the three groups that we created represent these educational levels and therefore also have a concrete definition of low, medium, and high.

In additional, noncausal, analyses we also look at two other indicators of usage. The second indicator is how many times the student has practiced in the online tool, and the last indicator is how many minutes the student has practiced per available school week on average. The latter two indicators are both continuous indicators.

We add these noncausal analyses because these three indicators reflect different assumptions about the mechanism underlying a potential effect of the tool. The simple participation dummy assumes that any type of participation can be helpful and focuses on an average effect, whatever the use the student made of the tool. The second indicator, number of practicing episodes, assumes that the effect of the tool varies (in a linear way) by the frequency of practicing. The last indicator stresses the intensity of the experience rather than the frequency. The overall time spent with the tool is analysed, apart from the number of episodes it is constructed of. The three indicators are not completely independent of each other, yet they do represent different dimensions of the learning mechanism underlying the didactical tool evaluated in this project.

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Pre- and posttest score means and standard deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Control group</td>
</tr>
<tr>
<td></td>
<td>Average</td>
</tr>
<tr>
<td>Spelling absolute test score T0</td>
<td>0.09</td>
</tr>
<tr>
<td>Spelling absolute test score T1</td>
<td>−0.08</td>
</tr>
<tr>
<td>Spelling absolute growth in test score T1 − T0</td>
<td>−0.11</td>
</tr>
<tr>
<td>Spelling absolute test score T0—lowest performance group</td>
<td>−0.41</td>
</tr>
<tr>
<td>Spelling absolute test score T1—lowest performance group</td>
<td>−0.74</td>
</tr>
<tr>
<td>Spelling absolute growth in test score T1 − T0—lowest performance group</td>
<td>−0.23</td>
</tr>
<tr>
<td>Spelling absolute test score T0—middle performance group</td>
<td>0.01</td>
</tr>
<tr>
<td>Spelling absolute test score T1—middle performance group</td>
<td>−0.08</td>
</tr>
<tr>
<td>Spelling absolute growth in test score T1 − T0—middle performance group</td>
<td>0.01</td>
</tr>
<tr>
<td>Spelling absolute test score T0—highest performance group</td>
<td>0.60</td>
</tr>
<tr>
<td>Spelling absolute test score T1—highest performance group</td>
<td>0.51</td>
</tr>
<tr>
<td>Spelling absolute growth in test score T1 − T0—highest performance group</td>
<td>−0.11</td>
</tr>
</tbody>
</table>

*representing significance level of 10%.
**representing significance level of 5%.
***representing significance level of 1%.
assignment of the groups. Next, we present the absolute test score at T1, where it can be seen that there is a difference in T1 score between the treatment and the control groups, significant at the 10% level. We do not see a significant difference in the growth between T0 and T1. If we split the sample (approximately equally) by ability, we see that for the lowest performing group, we find a significant difference at the 5% level at T1 and a significant difference at the 1% level for the growth between T0 and T1. We do not see any significant results for the middle- and higher performing groups. But, as discussed before, not all students complied with their assignment to the treatment group, for both domains, and not all students practiced the same amount of time, and therefore, we should explicitly control for this if we would like to analyse the effect of actually using the tool, instead of only having access to the tool.

5.2 The effect of access to and usage of the digital practice tool

Table 3 presents the results of (a) the first stage, in which we analyse whether the randomization into the treatment group indeed highly predicts whether students use the tool, (b) the intent to treat, in which we analyse what the effect is of offering the tool to students, (c) the ordinary least squares (OLS), which is the naive regression in which we analyse the effect of usage of the tool without controlling for the fact that students select themselves into using the tool (as discussed above, this is not a random process), and (d) the 2SLS analyses where we analyse the effect of using the tool while at the same time correction of this selection of usage.

Note that all coefficients in Table 3 are also standardized coefficients with respect to the standard deviation, similar as before. Table 3 shows that the first stage is highly significant, with a high r-squared, and that the assignment to the treatment of the experiment significantly and highly influences the chance to practice with spelling in the digital tool. In the second part of Table 3 (top right), we see the ITT analysis, which does not take into account the actual usage of the tool but looks at offering the tool. This analysis shows that, once controlled for ability by including both the pretest and the primary school ability test, there is a positive significant effect of assignment to the experiment on the spelling posttest score. Students with access to the tool score on average .14 of a standard deviation higher than students that did not have access. In the bottom part of Table 3, we see the OLS and the 2SLS results, where we analyse the effect of

<table>
<thead>
<tr>
<th>TABLE 3</th>
<th>First stage, ITT, OLS, and 2SLS results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First stage</strong></td>
<td><strong>ITT (effect of offering the tool)</strong></td>
</tr>
<tr>
<td>Dependent: Practice with digital tool for spelling</td>
<td>Dependent: Spelling posttest (T1)</td>
</tr>
<tr>
<td>Assignment experiment</td>
<td>(1)</td>
</tr>
<tr>
<td>0.876 (0.026)</td>
<td>0.072 (0.025)</td>
</tr>
<tr>
<td>Spelling pretest</td>
<td>-0.010 (0.014)</td>
</tr>
<tr>
<td>Primary school ability test total score</td>
<td>0.003 (0.003)</td>
</tr>
<tr>
<td>Female</td>
<td>0.029 (0.026)</td>
</tr>
<tr>
<td>Age</td>
<td>0.002 (0.027)</td>
</tr>
<tr>
<td>Oldest child</td>
<td>0.050 (0.026)</td>
</tr>
<tr>
<td>Situation at home</td>
<td>-0.034 (0.041)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.001 (0.019)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
</tr>
<tr>
<td>N = 350</td>
<td>N = 350</td>
</tr>
<tr>
<td>F(2, 347) = 557.25</td>
<td>F(26, 323) = 48.92</td>
</tr>
<tr>
<td>R² = 0.77</td>
<td>R² = 0.80</td>
</tr>
</tbody>
</table>

| **OLS (correlation of using the tool, with selection)** | **IV/2SLS (causal effect of using the tool)** |
| Dependent: Spelling posttest (T1) | Dependent: Spelling posttest (T1) |
| Participation experiment | (1) | (2) | (1) | (2) |
| 0.153 (0.070) | 0.159 (0.071) | 0.162 (0.080) | 0.164 (0.077) |
| Spelling pretest | 0.798 (0.039) | 0.696 (0.048) | 0.798 (0.039) | 0.695 (0.046) |
| Primary school ability test total score | 0.014 (0.008) | 0.014 (0.008) |
| Female | 0.108 (0.072) | 0.108 (0.069) |
| Age | -0.074 (0.073) | -0.074 (0.070) |
| Oldest child | -0.053 (0.072) | -0.054 (0.069) |
| Situation at home | 0.038 (0.112) | 0.039 (0.107) |
| Constant | -0.144 (0.048) | -6.084 (4.564) | -0.148 (0.051) | -6.076 (4.385) |
| Controls | No | Yes | No | Yes |
| N = 350 | N = 350 | N = 350 | N = 350 |
| F(2, 347) = 211.79 | F(26, 323) = 18.41 | F(2, 347) = 211.46 | F(26, 323) = 18.38 |
| R² = 0.55 | R² = 0.60 | R² = 0.55 | R² = 0.60 |

Note. ITT = intent-to-treat; OLS = ordinary least squares; 2SLS = two-stage-least-squares.
Controls = religion, country of birth, and language teacher. Standard errors in parentheses.
Outcome measures are standardized.
actual participation in the experiment on spelling posttest scores. Note that the results of the second (top right; third and fourth analyses [bottom left and right, respectively; ITT, OLS, and 2SLS]) are not significantly different from each other. The OLS analysis (bottom left) shows that students that practiced with the tool score on average .15 of a standard deviation higher on the posttest than students that did not practice. But, as we saw before, this is not a random group of students, and in the OLS analysis, we do not control for that. Therefore, we turn to the 2SLS analysis (bottom right), where we use the assignment to treatment as an instrument for actual participation in the experiment to control for this selection issue (in a way, we use using the tool as a mediator for the effect). The 2SLS analysis shows that the effect is a bit higher than for the OLS analysis, namely, .16 of a standard deviation higher for students that practiced, compared with students that did not use the tool to practice. This corresponds to a small effect (see explanation above on Cohen's d). Note that in all four types of analysis in Table 3, there is hardly a difference between the model without and the model with covariates, also pointing at a successful randomization.

5.3 | Heterogeneous effects

The results in Table 2, presented above, suggest that the effect of the ICT tool may be heterogeneous. Therefore, in these additional regression analyses, we allow the effect estimate to differ by distinguishing three ability groups. In this analysis, we incorporate interaction effects for whether a student scored low, medium, or high on the primary school ability test.

In these specifications, we include interaction effects between the treatment indicators and the performance group of the student, as well as the single terms for treatment and performance groups. This is done for each of the three performance indicators separately.

The results of these analyses are presented in Table 4. The first stages of these analyses were strong, with $r$-squareds of .80, .79, and .92 and highly significant treatment dummies. We find a significant difference in the effect of the treatment indicator for students that scored low on the primary school achievement test and students that scored high on this test. This significant difference is in favour of the low-scoring students, implying that students who displayed a lower ability level at the end of primary school, benefit more from practicing with the digital tool. We see a border significant difference ($p = .10$) between the middle and the lower groups, again in favour of the low-scoring students. Although the upper group is significantly different from the lower performance group, the difference in coefficients results to zero, making us to conclude that there does not seem to be an effect for high-performing students. This leads to the conclusion that the effect of the ICT tool is indeed heterogeneous and that low-performing students benefit the most.

### 6 | ADDITIONAL ANALYSES WITH DIFFERENT USAGE INDICATORS

Table 5 presents the results of the 2SLS analyses for the treatment indicator “number of times practiced” and “minutes per week practiced” (note that the ITT is the same as in Table 3 and therefore not presented). The table contains the same control variables as Table 3 and also has the score for spelling in T1 as outcome measure, again with standardized coefficients. Now, we do not have the dummy whether someone practiced as the treatment, but the number of times (episodes), or the total time (minutes) someone practiced during the intervention period. This is not a fully exogenous treatment, as students have a say in how much or how often they practice, but it is interesting to look at nonetheless. These analyses give us insight in whether the effect differs by intensity of treatment and tests different assumptions about the mechanism underlying a potential effect of the tool.

The coefficient of 0.004 in the multivariate analyses in the left-hand analyses of Table 5 indicates that for each additional practice episode, the posttest for spelling (T1) increases by 0.004 of a standard deviation. This is a highly significant but very small coefficient, as someone should practice more than 50 times to get to an overall small effect size of 0.2. Given the average number of times practiced of 39 times, 50 would imply an increase of more than 100% compared with that average, which also indicates how small the estimated effect is.

The second part of Table 5 also shows highly significant results. The coefficient of 0.023 indicates that for each additional 4 min of practice per school week, the score increases by .1 of a standard deviation. Again, this is an increase of over 50% compared with the average, but given that the instruction was to practice at least 15 min with spelling, if students would practice indeed 15 min, this would make a difference of over .2 standard deviations, and that is comparable in size (larger actually), with the estimated effect of the participation dummy in the bottom part of Table 3.
TABLE 5  Number of times and minutes per week practiced as outcome measures (2SLS)

<table>
<thead>
<tr>
<th></th>
<th>Number of times / minutes per week practiced</th>
<th>Minutes per week practiced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IV/2SLS (causal effect of using the tool)</td>
<td>IV/2SLS (causal effect of using the tool)</td>
</tr>
<tr>
<td></td>
<td>Dependent: Spelling posttest (T1)</td>
<td>Dependent: Spelling posttest (T1)</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Number of times / minutes per week practiced</td>
<td>0.004 (0.002)</td>
<td>0.023 (0.011)</td>
</tr>
<tr>
<td></td>
<td>Spelling pretest</td>
<td>0.789 (0.039)</td>
</tr>
<tr>
<td></td>
<td>0.013 (0.008)</td>
<td>0.013 (0.008)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0.091 (0.069)</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-0.057 (0.070)</td>
</tr>
<tr>
<td></td>
<td>Oldest child</td>
<td>-0.048 (0.069)</td>
</tr>
<tr>
<td></td>
<td>Situation at home</td>
<td>0.035 (0.107)</td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>-0.147 (0.051)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.148 (0.051)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>N = 350</td>
</tr>
<tr>
<td></td>
<td>F(2, 347) = 216.94</td>
<td>R² = 0.55</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>N = 350</td>
</tr>
<tr>
<td></td>
<td>F(26, 323) = 18.63</td>
<td>R² = 0.60</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>N = 350</td>
</tr>
<tr>
<td></td>
<td>F(2, 347) = 212.90</td>
<td>R² = 0.55</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>N = 350</td>
</tr>
<tr>
<td></td>
<td>F(26, 323) = 18.58</td>
<td>R² = 0.60</td>
</tr>
</tbody>
</table>

Note. 2SLS = two-stage least squares.

Controls = religion, country of birth, and teacher Dutch language. Standard errors in parentheses.
Outcome measures are standardized.

7 | CONCLUSION AND DISCUSSION

In this article, we discussed the outcomes of an IT experiment with which a school aims at improving spelling skills among seventh grade students. Instead of creating additional instruction time, the school decided to adopt a noncompulsory approach, offering the IT tool for practice at home, as suggested homework from the regular homework given by teachers for their particular subject. As a consequence, the setup and maintenance costs for the school are relatively low. It buys licenses to the tool, distributes them among the students, and suggests teachers to follow-up, while not attaching grades to practice, nor obliging teachers to support the tool.

Not surprisingly, adoption of the tool by students is uneven. Yet, also among teachers, enthusiasm is mixed, in part because the freedom it offers to students makes it difficult to coordinate home practice learning with a regular, class-based, and fairly homogeneous instruction process.

Nevertheless, the experiment reveals a positive effect of offering the tool, suggesting that uneven student adoption and teacher support did not impede an average benefit to students. Moreover, a further instrumental variable analysis in which we control for selection of usage of the tool shows the potential of the tool. Using the tool contributes substantially to spelling performance, although the analyses reveal that effects are solely present for low ability students. Therefore, getting more students to use the tool or to use it more actively may lead to stronger results than observed in this trial.

These findings are based on whether the students use the tool at least once. However, if we look at the intensity of treatment, we learn that the effect seems to be mostly about using it. With the average amount of times and number of minutes per week that we observe, we see a positive significant effect. However, students that do practice would need to double the amount of minutes per week to achieve a sizeable effect for additional practicing.

With its effect size of 0.16 of a standard deviation, the 16-week experiment seems moderately effective, although the effect size is a little larger than what Cheung and Slavin (2012) found in their recent meta-analysis on reading, as they conclude that the average effect size is around 0.11. On the other hand, spelling is a lot easier to automatize than reading, which could be an explanation for the higher effect size. Furthermore, it should be noted that the experiment operated with a strong control condition. Control students were not deprived of the tool but got access to the tool for training with vocabulary rather than spelling. Hence, the effect we measure is specifically tied to spelling exercises, not to the effect of language practice in general. Moreover, the meta-analysis of Cheung and Slavin (2012) suggested that the supplementary nature of the IT tool might have eroded all effect, which perspective renders this experiment even more successful. A potential explanation for the positive effect observed here is the optional homework aspect, offering students a large degree of freedom regarding the timing and intensity of their effort, which was not the case in previous experiments carried out in school. Also recall that students write on average 15 to 45 min of Dutch homework per week. Therefore, an additional 7 min is an increase of Dutch homework time of 15% to almost 50%. So although the 7 min is only (less than) half of the supposed practice, it is still a considerable increase in language homework time for students.

It is also important to mention that we did not explicitly study the mechanisms behind this effect, which is a limitation of our study. However, previous literature shows that for the mathematics exercises of the same tool, the mechanism was the adaptive nature of the tool (Haelermans & Ghysels, 2017). Furthermore, the instrumental variable approach shows that the usage of the tool is a mediator for the effect of the tool (which makes sense, as it is unlikely that an effect would occur if the tool would be not be used).

Finally, it should be noted that the intervention is much cheaper than a common alternative, an additional hour of regular instruction. For school principals and educational policy designers, the tool thus represents an attractive option when aiming at a central element of basic literacy, namely, spelling skills.
REFERENCES


How to cite this article: Ghysels, J., Haelermans, C. New evidence on the effect of computerized individualized practice and instruction on language skills. *J Comput Assist Learn*. 2018;34:440–449. [https://doi.org/10.1111/jcal.12248](https://doi.org/10.1111/jcal.12248)