Computers & Education 122 (2018) 1–8

Contents lists available at ScienceDirect

Computers & Education

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

The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment

Janke M. Faber[a,*], Adrie J. Visscher[b]

[a] Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioral Science, Management and Social Sciences, University of Twente, 7500 AE Enschede, The Netherlands

[b] ELAN, Institute for Teacher Professionalization and School Development, Faculty of Behavioral, Management and Social Sciences, University of Twente, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Keywords:
Formative assessment
Elementary education
Improving classroom teaching
Teaching/learning strategies

ABSTRACT

In this study, a randomized experimental design was used to examine the effects of a digital formative assessment tool on spelling achievement of third grade students (eight-to nine-years-olds). The sample consisted of 30 experimental schools ($n = 619$) and 39 control schools ($n = 986$). Experimental schools used a digital formative assessment tool, whereas control schools used their regular spelling instruction and materials. Data included standardized achievement pre-posttest data, the number of total assignments completed, and the percentage of adaptive assignments completed by students. Although the results did not show that the use of a digital formative assessment tool affected spelling achievement, the findings point to important issues upon which future research can build.

1. Introduction

Mobile digital tools like tablets can be beneficial for formative assessment practices in education (Sung, Chang, & Liu, 2016). Formative assessment can be defined as "all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (Black & Wiliam, 1998, p.82). Digital formative assessment tools (DFATs) can provide students and teachers with feedback regarding the progress of (individual) students (De Witte, Haelermans, & Rogge, 2015; Koedinger, McLaughlin, & Heffernan, 2010; Sheard, Chambers, & Elliott, 2012). As the use of such tools is increasing, it is important to study whether DFATs improve the quality of teaching and learning processes (Hwang & Tsai, 2011; Sung et al., 2016). In this study, a randomized experiment was conducted to examine the effects of a digital formative assessment tool on spelling achievement of third grade students (eight or nine year olds). With this tool, students complete assignments on their own tablets and teachers follow students' actual progress on their dashboards. Positive effects of this digital formative assessment tool called 'Snappet' on mathematic achievement and mathematic motivation were found in a previous study (Faber, Luyten, and Visscher, 2016). Before we explain Snappet, and the method used, we will provide a brief summary of what can be learned from feedback effectiveness studies in general, as the provision of feedback is an important characteristic of Snappet, and DFATs in general.

* Corresponding author.
E-mail addresses: j.m.faber@utwente.nl (J.M. Faber), a.j.visscher@utwente.nl (A.J. Visscher).

https://doi.org/10.1016/j.compedu.2018.03.008
Received 8 August 2017; Received in revised form 1 March 2018; Accepted 5 March 2018
Available online 07 March 2018
0360-1315/ © 2018 Elsevier Ltd. All rights reserved.

2. Theoretical framework

The effectiveness of feedback has been under research for more than a hundred years already. Meta-analysis findings have revealed that feedback can affect performance positively (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Lysakowski & Walberg, 1982). However, feedback does not unconditionally result in positive performance effects. Feedback content, characteristics of the learning tasks, characteristics of the feedback receiver, the feedback frequency and timing, all influence whether and how feedback will affect performance (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996; Shute, 2008).

Let us first focus on feedback timing and frequency. Whether immediate (feedback right after a response), or delayed feedback (later than right after a response, mostly defined as 'relative to immediate feedback') is more effective, depends on the task characteristics. Shute (2008) found in her review that immediate feedback seems most effective for tasks which, relative to students' capacities, are complex, whereas delayed feedback seems to be more effective for simpler task. Regarding teachers as feedback receivers, teachers connect the feedback they received to the assessments of their students in their follow-up instruction if feedback is received immediately after students finish their assessments (Hellrung & Hartig, 2013; Yeh, 2009). Furthermore, teachers' feedback seems more effective when it is based on frequent assessments, as frequently administered assessments reflect students' actual learning more accurately (Konstantopoulos, Miller, & van der Ploeg, 2013).

The feedback content and the learning tasks on which feedback is given also influence the effectiveness of feedback. Feedback which drives attention to the learning task can be effective, whereas driving attention to the 'self' as a person is not effective (Kluger & DeNisi, 1996). Examples of the latter are praise, discouragement, and normative feedback (i.e., contrasting student performance to that of others). The simplest form of learning task feedback content is referred to as 'knowledge of results feedback' and only informs the recipient whether the answer is (in)correct. More complex learning task feedback is known as 'elaborated feedback'. Elaborated feedback includes, among other things, information on why the response is (in)correct and guides the feedback receiver towards the right direction: how to improve if the answer was wrong and/or information about the misconceptions the student may have. In general, elaborated feedback is more effective for learning than simple feedback (Bangert-Drowns et al., 1991; Shute, 2008). Furthermore, feedback is more effective when combined with setting performance goals (Locke & Latham, 2002). Goals direct learning activities and allow students and teachers to monitor the degree of progress from current learning levels towards the goals set (Hattie & Timperley, 2007). Studies on the effectiveness of computerized feedback for students support the view that elaborated feedback positively influences performance. Smaller effects were found for simple feedback, and for feedback which provides the correct response (Lipnevich & Smith, 2009; Van der Kleij, Feskens, & Eggen, 2015). More complex tasks are those requiring elaborated feedback (Van der Kleij et al., 2015), a finding confirmed by studies on the effectiveness of DFATs on algebra (Bokhove & Drijvers, 2012) and general mathematics achievement (Wang, 2011).

Characteristics of the feedback receiver influence how much is learned from feedback. Findings show that motivated students pay more attention to feedback and spend more time on studying feedback (Timmers, Braber-van den Broek, & van den Berg, 2013; Timmers & Veldkamp, 2011). The use of DFATs and feedback can also positively affect students' engagement, interest, and their attitudes towards the learning tasks on which feedback is provided (Hunsu, Adesope, & Bayly, 2016; Pilli & Aksu, 2013). On the other hand, if students receive mostly negative feedback while their feelings of competence are high, then the use of DFATs might lower their motivation (Muis, Ranellucci, Trevors, & Duffy, 2015). Feedback effects also vary among high-performing and low-performing students. In her review, Shute (2008) concludes that high-performing students benefit more from delayed feedback. High-performing students need facilitative and verification feedback (i.e., fewer hints) as it is assumed that they learn more in environments that are challenging and allow for greater autonomy. Low-performing students need immediate feedback, as they perceive new tasks as more complex and need direct support in their learning. Furthermore, low-performing students benefit more from directive and explicit feedback, as they need guidance and structure when they learn. In most DFATs studies, larger effects are found for low-performing students (Bokhove & Drijvers, 2012; Koedinger et al., 2010; Wang, 2014). Each of these studies provides a different explanation to account for this finding. For instance, the methods effective for novice students may lose their effect for more experienced students (Bokhove & Drijvers, 2012). Alternatively, teachers may be more capable of letting special education students take part in regular classroom instruction with the digital formative assessment tool (Koedinger et al., 2010). Finally, the selection of assignments and text by the digital formative assessment tool may be especially beneficial for low-performing students (Wang, 2014; Molenaar & van Campen, n.p.). Surprisingly, for the DFAT Snappet, in mathematics larger effects were found for high-performing students (Molenaar & van Campen, n.p.; Faber et al., 2016). An explanation might be that with Snappet, high-performing students can easily complete more assignments than normal, and that students are more equally challenged by the adaptive assignments Snappet provides.

3. Snappet

In this section, we describe the features of Snappet as a tool and, more specifically, the characteristics of the feedback Snappet provides. When working with Snappet, students complete assignments on their own tablets. Students and teachers both receive feedback based on the progress of students on these assignments. The learning tasks on which students received feedback in this study are spelling tasks.

Grade three students learn spelling rules (e.g., recognize the distinctions between words with a *sch/schr* sound, *s/z* sound, and words with an *f/v* sound) and apply these rules in the assignments they complete on their tablets. For example, students have to fill in the blanks to complete words in assignments or they have to write or choose words corresponding to images presented on the tablet. Spelling in the Dutch language generally is considered to be not that difficult compared to spelling in other languages; the sounds of

letters (i.e. letters not in a word) are often the same as the sounds of the same letters in a word. Nevertheless, Dutch language has quite a few grammatical rules and exceptions to those rules.

3.1. Feedback to students

On the tablet's start screen, students can select the lesson assigned by the teacher and see an overview of the curriculum assignments belonging to the lesson they had picked. In this overview, green blocks represent correctly completed assignments, red blocks stand for incorrectly completed assignments, orange blocks represent corrected assignments (assignments that were not immediately answered correctly), and blue blocks represent unfinished assignments. Students select the blue blocks to start working on the assignments. They receive feedback immediately after they give a response. The feedback given is 'knowledge of results' feedback, it tells whether the answer is correct or incorrect (a green curl (check mark) or a red cross). Besides the standard curriculum assignments, there is an extra option to work on specific spelling learning goals. With this option, students practice only assignments belonging to the same learning goal. Examples of third grade spelling learning goals are words ending on *d* or *t*, words with double consonants like in the Dutch word 'bru*gg*en' (meaning *bridges*), or in 'pe*tt*en' (*caps*). In another overview, students can see how well they performed on each learning goal relative to their performance on other learning goals; performance is ranked from low to high by zero to four stars. Furthermore, Snappet provides adaptive assignments. When choosing these assignments students can practice with assignments that match their performance level. Snappet uses an item-response theory model to predict students' ability levels based on students' previous responses. These are used for selecting assignments matching students' ability levels. Most teachers require students to complete the curriculum assignments first and then continue with the adaptive or learning goals assignments.

3.2. Feedback to teachers

Teachers can access their own teachers' dashboard through which they can monitor students' progress. During the lesson, teachers could track in real-time the number of assignments students completed during that lesson and whether students' responses were immediately correct, correct after first having given an incorrect response, or incorrect. Teachers can request normative feedback at any time. Normative feedback includes a comparison of a student's performance on a specific learning goal with the performance of other students, or with other classes using Snappet. Snappet uses the following performance categories: students and classes belonging to the 25% highest scoring students, the 50–75% group, the 25–50% group, the 10–25% group, and the students and classes belonging to the 10% lowest scoring students; see Fig. 1 (Faber et al., 2016, p88). Furthermore, teachers could request self-referenced student feedback, i.e., the learning goals on which students perform high/low compared to their own performance on other learning goals.

3.3. Hypotheses

Based on feedback effectiveness studies on DFATs that are comparable to Snappet (Bokhove & Drijvers, 2012; Koedinger et al., 2010; Sheard et al, 2012; Wang, 2011), and based on the Snappet effects found in previous research, we expected that Snappet would have a positive effect on spelling achievement (hypothesis 1). Furthermore, it was expected that the effectiveness of Snappet would be greater, the more students used Snappet (hypothesis 2). Based on previous Snappet research findings, it was also expected that Snappet would be more effective for high-performing students' (hypothesis 3).

4. Method

4.1. Participants

Schools were informed about the study by e-mail, and after a week, school principals were asked whether they were willing to participate. In a previous study, schools were required to participate with a math Snappet package and could additionally choose to participate in this study with a spelling package. Only schools with teachers and students in grade three who did not have any experience with Snappet were allowed to participate. Grade three students are usually eight or nine years old. Seventy-nine primary schools in the East of the Netherlands were recruited and randomly assigned to the experimental group (n schools = 40), the control group (n schools = 50), or to a waiting list (n schools = 7). Randomization was done at the school level as school principals decided which of their grade three classes would participate. Two experimental schools and eleven control schools decided not to participate after they were informed of the randomization outcome; two schools on the waiting list were added to the experimental group. Furthermore, ten experimental schools decided to participate only with math. The initial spelling sample comprised 30 experimental schools (n students = 619) and 39 control schools (n students = 986). In the experimental group, 52.2% of the students were male, in the control group – 46%. In the experimental group, the percentage of disadvantaged students was 5.8% and 4.4% in the control group; a child belongs to the disadvantaged category if none of the parents attained a higher educational qualification than lower vocational education. In Table 1, more descriptive statistics for the students in the experimental and control groups are presented.

4.2. Procedures

Experimental schools used Snappet between January 19 and July 4 (2015), whereas control schools used their regular teaching

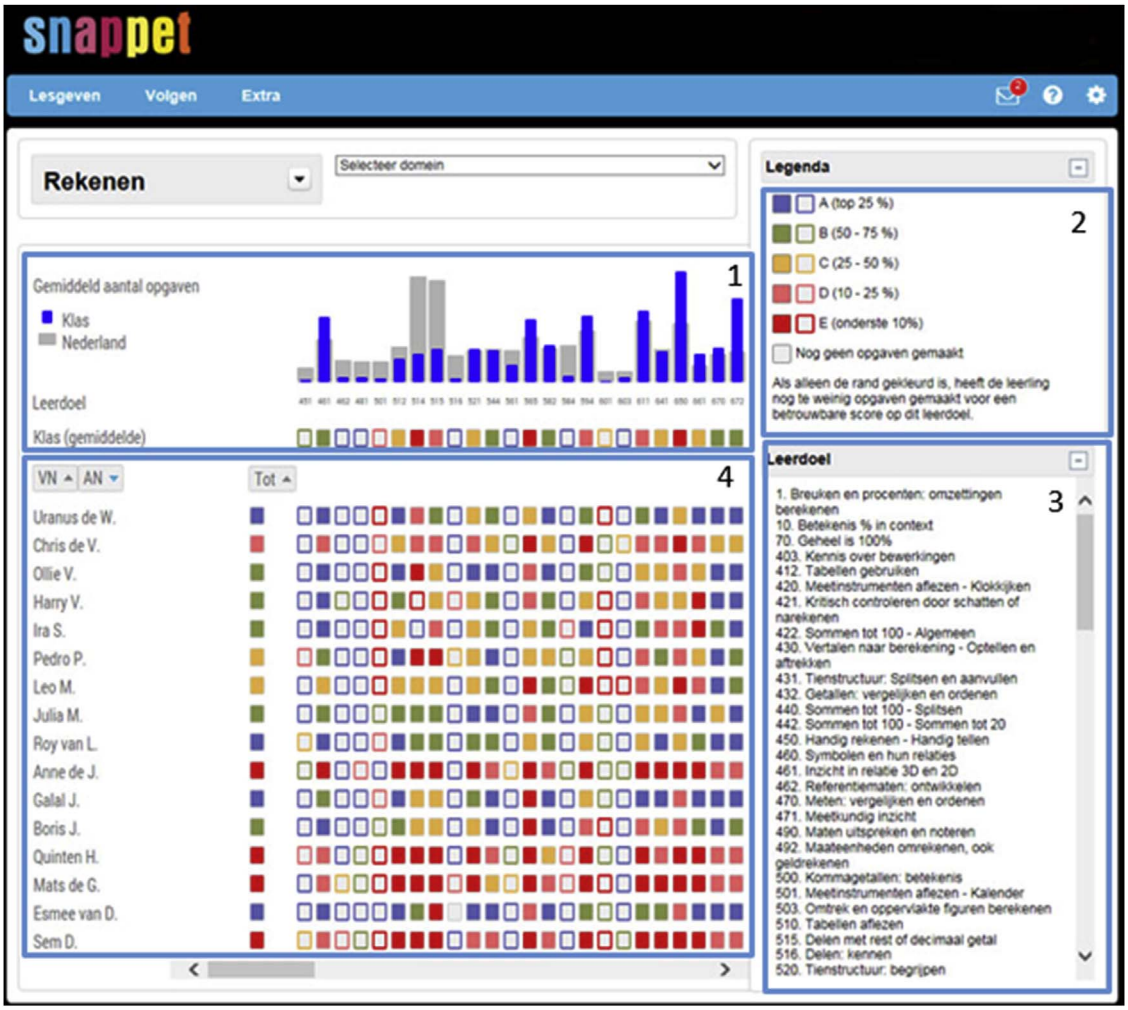

Fig. 1. Teacher dashboard which shows how a class progresses. (1) How this class performs on learning goals compared to national performance. (2) Benchmark legend, e.g., blue represents the 25% highest scoring students/classes. (3) Description of the learning goals in Dutch. (4) The performance of each student compared with other students. Figure reprinted from "The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment," by Faber et al., 2016

Fig. 1 Contains Snappet mathematics package screenshots, the design of teachers' spelling dashboard is the same as the mathematics package design. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Descriptive statistics.

Students	Experimental n = 619		Control n = 986	
	Mean (Sd.)	n	Mean (Sd.)	n
Gender (male)	52.2%	616	46%	957
% disadvantaged students	5.8%	424	4.4%	707
Spelling June 2014	122.7 (6.59)	554	123.1 (6.82)	858
Spelling January/February 2015	128.7 (6.70)	619	128.6 (6.57)	963
Spelling June 2015	132.1 (7.65)	619	131.8 (7.52)	962
Mathematics June 2014	65.02 (14.58)	550	66.02 (14.10)	882
Mathematics January/February 2015	72.93 (14.24)	600	74.72 (13.50)	974
Motivation[a]	3.81 (1.00)	580	3.61 (0.99)	772
Total assignments	2108 (727)	618		
% adaptive assignments	21.72 (11.09)	591		

[a] Scale is 1 (low) – 5 (high).

methods and materials during the same period. With Snappet, the same instructional content is taught as in the regular curriculum. Thus, the experimental group students completed comparable curriculum assignments on their tablets to those used in traditional paper-based settings. In addition, experimental group students could work on assignments belonging to a specific learning goal and on adaptive assignments (see also section 3). In most Snappet lessons, teachers required students to first complete all curriculum assignments and then work on either the adaptive assignments or the learning goals assignments for the remainder of the lesson.

At the start of the intervention, teachers in the experimental condition followed a short introduction program (one afternoon). Teachers learned how to integrate Snappet in their classroom and were given the opportunity to follow an extra training program (one additional afternoon) on the interpretation of the Snappet teacher feedback and how to use that feedback for classroom differentiation. Teachers were free to choose how to integrate Snappet in their lessons. All participating teachers could consult a Snappet coach by telephone in case of questions or difficulties. Twenty-four teachers of the mathematics and spelling teacher sample (n teachers = 44) attended the introduction program and the additional program, and consulted the Snappet coach at least three times. Four teachers followed the introduction program and the additional program. Ten teachers attended only the introduction program and two teachers did not participate in any of the activities (non-response of four teachers).

4.3. Data collection

Prior to data collection, all parents were informed by email/letter about the study and the data collection procedures. We used standardized tests, a student survey, and student log files for data collection (see Table 1 for the descriptive statistics for the students).

Standardized tests. Cito standardized spelling and mathematics tests (Cito is the Dutch national institute for test development) were used to measure student achievement. Most Dutch primary schools use Cito assessments during the entire primary school period, which are administered twice a year in January or February, and in June. Scores from tests in different grades can be expressed on the same ability scale. The spelling tests of June 2015 were used as a posttest. To control for the quality of randomization, the spelling and mathematics tests of June 2014 and January/February 2015 were used.

Student survey. A student survey was used to measure spelling motivation. Items were measured on a five-point Likert scale (1 – strongly disagree, 5 – strongly agree) and scale reliability was calculated using Cronbach's alpha. Experimental and control students responded to the survey at the end of the intervention. For spelling motivation, the following five items were used: “Spelling lessons are boring”, “I like spelling”, “I enjoy doing spelling assignments”, “I think spelling is interesting”, “I think spelling is important” ($\alpha = 0.86$).

Snappet log files. Snappet log files were used to measure the extent to which students used Snappet. Data included in the analyses were the number of total assignments completed and the percentage of these that were adaptive assignments. These data were registered by the Snappet software and provided to us by Snappet for this study.

4.4. Data analysis

As students are nested in classes, a multilevel regression model was used to test our hypotheses (Snijders & Bosker, 1999). We calculated fixed effects for all variables except for the random intercept effect. Results of an independent two-sample t-test showed that the experimental and control group were comparable in the proportion of disadvantaged students, mathematics achievement in June 2014, spelling achievement in June 2014, and for spelling achievement in January/February 2015 (just before the intervention). The percentage of males was significantly higher in the experimental schools ($t = 2.34$, $p < 0.05$). A statistically significant difference was found on the pretest mathematics scores in January/February 2015 ($t = 2.71$, $p < 0.05$), indicating that students in the experimental schools performed lower than students in the control schools. Therefore, gender and pretest mathematics scores were included as covariates in the multilevel analyses. Variables which were included in the analyses are (in order of inclusion to the models): spelling posttest (dependent variable), gender (covariate), mathematics pretest (covariate), spelling motivation (covariate), Snappet (independent variable: yes/no Snappet user), spelling pretest January/February 2015 (interaction with Snappet), and, finally, the total number of assignments and the percentage of adaptive assignments of the experimental students as dependent variables. Scores on all variables were converted to z-scores except for the dichotomous variables student gender (female = 0) and the Snappet condition (control group = 0). All analyses were performed using IBM SPSS Statistics version 22 (IBM Corp Released, 2013).

5. Results

The results of the multilevel analysis predicting spelling achievement are shown in Table 2. Results of the null model showed that a large proportion of the variance in student achievement was situated at the student level (0.94), and the proportion of group level variance of the total variance was 0.07. Covariates were added in the first model. There was a significant negative effect of gender ($\beta = -0.25$, $p < 0.05$) on the spelling posttest, indicating higher performance among females. However, gender was no longer significant in model three in which spelling pretest scores were included. The significant positive effect of mathematics ($\beta = 0.46$, $p < 0.05$) indicated that students with higher mathematics results also performed better on spelling. This effect of mathematics decreased in model three ($\beta = 0.11$, $p < 0.05$) after including the spelling pretest scores. Furthermore, the effect of spelling motivation on the spelling posttest was significant ($\beta = 0.19$, $p < 0.05$), but also decreased after the spelling pretests scores were included ($\beta = 0.08$, $p < 0.05$).

The results of the second model showed that there was no significant effect of Snappet (hypothesis 1 not confirmed). Spelling

Table 2
Multilevel model predicting student spelling achievement.

Model	0		1		2		3		4[a]	
Predictors	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Fixed										
Intercept	−0.01	0.04	0.13*	0.05	0.09	0.06	0.00	0.05	0.04	0.06
Gender (female = 0)			−0.25*	0.05	−0.25*	0.05	−0.05	0.03	−0.04	0.05
Mathematics pretest			0.46*	0.02	0.46*	0.02	0.11*	0.02	0.08*	0.02
Spelling motivation			0.19*	0.03	0.18*	0.03	0.08*	0.02	0.06*	0.02
Snappet					0.09	0.08	0.05	0.06		
Spelling pretest							0.77*	0.02	0.71*	0.03
Snappet x pretest							−0.00	0.03		
Total assignments									0.09*	0.04
% adaptive									0.11*	0.04
Random										
Variance student level	0.94*	0.03	0.72*	0.03	0.72*	0.03	0.28*	0.01	0.28*	0.02
Variance group level	0.07*	0.02	0.06*	0.02	0.06*	0.02	0.05*	0.01	0.08*	0.03

*p < 0.05 one-sided testing.
[a] Model with experimental student data only.

achievements in the experimental group were no greater than the achievements of students in the control group at the end of the intervention period. The results of model three indicate that there was no interaction effect between the spelling pretest scores and Snappet (hypothesis 3 not confirmed), meaning that the use of Snappet was not more effective for high-performing students in this study.

Model four only included the data of the experimental group. The results of this model show that there was a small significant positive effect of the total number of assignments completed ($\beta = 0.09, p < 0.05$), and of the percentage of adaptive assignments ($\beta = 0.11, p < 0.05$) on the spelling posttest. These findings support hypothesis 2: The effectiveness of Snappet is greater for students who used Snappet more.

6. Discussion and conclusion

The use of digital formative assessment tools (DFATs) in education is increasing rapidly. For example, within a few years, 2000 Dutch primary schools (out of 6500 schools in total) have started to use the Snappet DFAT and their number is growing daily. It is therefore important to study whether DFATs improve the quality of teaching and learning. In this study, a randomized experimental design was used to examine the effects of a digital formative assessment tool on spelling achievement of third grade students (eight-to nine-year-olds). The findings of the study showed that there was no effect on spelling achievement. However, Snappet log files data reveal that the achievements of students who used the digital formative assessment tool to a larger extent had higher achievements. We do not know whether students who completed more Snappet assignments performed better on the spelling posttests because they completed more assignments, or that higher performing students simply completed more assignments.

One explanation for the absence of a spelling achievement effect might be that the spelling software was less attractive to students (the lay-out and the type of assignments) compared with the mathematics software. Indeed, several teachers in the experimental condition mentioned this. Students may have spent more time on adaptive mathematics assignments than on adaptive spelling or learning goal assignments. Additional analyses show that compared with our mathematics study (Faber et al., 2016), students completed fewer spelling assignments on their tablet (Table 3). Table 3 shows that high-performing students completed more

Table 3
Spelling and mathematic assignments compared.

Students	Spelling total assignments		Math total assignments		Spelling % adaptive[a]		Math % adaptive[a]	
	Mean (Sd.)	N students	Mean (Sd.)	N students	Mean (Sd.)	N students	Mean (Sd.)	N students
20% lowest[b]	1730 (618)	79	2512 (920)	139	19.7 (12.4)	85	23.1 (11.2)	144
20–40%	1898 (611)	101	2761 (920)	161	20.4 (11.6)	105	23.1 (10.8)	169
40–60%	2138 (634)	84	2976 (1004)	142	20.3 (10.0)	90	23.7 (10.8)	147
60–80%	2092 (686)	157	3218 (1104)	181	21.1 (10.2)	159	25.7 (12.3)	187
20% highest	2410 (795)	170	3350 (1207)	159	24.7 (10.8)	179	26.8 (12.5)	163
Total	2108 (727)	591	2948 (1117)	792	21.7 (11.1)	618	25.0 (12.8)	820

[a] Percentage of adaptive assignments of all completed curriculum assignments and adaptive assignments.
[b] Groups were computed using Cito national student reference groups. Cito composes these groups using a representative sample of Dutch students. Students fall within one of the five groups, ranging from the 20 percent highest scoring students to the 20 percent lowest scoring students, based on their Cito results.

assignments with Snappet than lower performing students, and the percentage of adaptive assignments is higher for high-performing students than for lower performing students. Furthermore, the percentage of adaptive assignments for all students is somewhat higher for mathematics than for spelling. This might explain why the mathematics effect was mainly driven by high-performing students, as those students generally finish the curriculum assignments faster and thus have more time for additional tasks. If the number of assignments completed by students is the reason for the differences found between mathematics and spelling effects, this might imply that the mathematics achievement effect was mainly caused by the fact that students complete more assignments, and to a lesser extent by the student and the teacher feedback. Since we have no data on the number of completed assignments in the control group students, we cannot test this potential explanation.

Our most important finding that students using Snappet do not perform better in spelling contrasts previous DFATs studies showing positive effects on mathematics and spelling achievement (De Witte et al., 2015; Faber et al., 2016; Koedinger et al., 2010; Molenaar & van Camplen, n.p.; Sheard et al., 2012). A meta-analysis of the impact of formative assessment also revealed that formative assessment is effective for both complex subjects (i.e., mathematics and science) and less complex subjects (i.e., English language arts), and that larger effect sizes were found for less complex subjects (Kingston & Nash, 2011). Sung et al. (2016) did not find significant differences in achievement effects between different subjects in their meta-analysis on the use of mobile devices. Slavin (2013) did not find meaningful differences between mathematics and reading as a result of technological innovations in education (Slavin, 2013). However, in their review on the effects of (external) feedback reports to teachers, Hellrung and Hartig (2013) concluded that the effects for mathematics achievements were stronger than for reading. The authors stated that it might be easier for teachers to recognize changes in student progress in the case of clearly structured subjects as mathematics relative to more general subjects (Hellrung & Hartig, 2013). Furthermore, reading and writing on a screen might lead to poorer reading comprehension and writing skills than if this is done on paper (Mangen, Walgermo, & Brønnick, 2013). However, Wollscheid, Sjaastad, and Tømte (2016) did not find evidence in favor of the use of pen and paper on writing skills compared to digital devices in their review of ten studies. Another explanation for our findings might be that hearing word sounds, or phonology, was not included in the digital formative assessment tool. Crucially, phonological awareness correlates well with spelling skills and findings indicate that beginner spellers rely more heavily on phonology (Bosman & Van Orden, 1997). Overall, as far as we know, our finding contradicts most other DFATs studies. Only in one review study on the effects of external feedback reports to teachers, the effects for mathematics were stronger than for reading. Perhaps teachers were better at adjusting their teaching to students' needs in light of the Snappet mathematics teacher feedback, which provides progress information on more clearly defined mathematics rules relative to the Snappet spelling feedback to teachers.

This study contributes to our knowledge base on DFATs, two limitations need to be considered when interpreting our findings. First, randomization could only be done at the school but not at the group level. Some matters beyond our control affected the characteristics of our sample. For instance, some school principals in the experimental group decided not to participate with all their grade three classes and school principals decided that their classes would only participate for mathematics or for mathematics and spelling both. Also, the withdrawal of eleven control schools adversely affected the quality of our randomization. A second limitation of the study is that the data to measure the extent to which students used Snappet were only measured in one experimental group, making it difficult to draw strong conclusions on whether DFATs are more effective if students use the tool to a greater extent.

Although the results of this study did not confirm that the DFAT studied positively affected spelling achievement, the results point to important issues upon which future research can build. The results indicate that further research into the effectiveness of DFATs on learning and teaching is needed as the use of DFATs is growing rapidly, making it important to understand in which conditions these systems are (in)effective. In future work, it will be important to study the effects and the use of DFATs in different subject areas as our results clearly showed effect differences between spelling and mathematics. Especially, more in-depth qualitative research is needed into how teachers and students use the feedback and which elements of the information teachers and students receive form the Snappet tool information is especially helpful for them in improving their teaching and learning (maybe this varies across subject areas or between boys and girls). In future studies, it would also be interesting to investigate whether the use of adaptive assignments is more effective than static assignments, and whether DFATs are more effective in combination with feedback to teachers than without feedback to teachers (by using an additional experimental group). Further research on student feedback is also required as our findings show that knowledge of result feedback was not effective for relatively simple learning tasks such as grade three spelling tasks. The intended effects might have been accomplished by providing students with more elaborated feedback. Finally, it will be worth investigating how teachers can be equipped for using the feedback they receive on student progress for differentiating their teaching. It is generally known that differentiating instruction in line with differences between students is a very complex task for which teachers are only trained to a limited extent during their pre-service and in-service training (Maulana, Helms-Lorenz, & Van de Grift, 2016). We need to investigate what effective teaching with DFATs entails and how the skills needed for such forms of teaching can be trained effectively.

References

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. http://dx.doi.org/10.1080/0969595980050102.

Bokhove, C., & Drijvers, P. (2012). Effects of a digital intervention on the development of algebraic expertise. *Computers & Education, 58*(1), 197–208. http://dx.doi.org/10.1016/j.compedu.2011.08.010.

Bosman, A. M. T., & Van Orden, G. C. (1997). Why spelling is more difficult than reading. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.). *Learning to spell: Research,*

theory, and practice across languages (pp. 173–194). Hillsdale, NJ: Lawrence Erlbaum Associates.
De Witte, K., Haelermans, C., & Rogge, N. (2015). The effectiveness of a computer-assisted math learning program. *Journal of Computer Assisted Learning, 31*(4), 314–329. http://dx.doi.org/10.1111/jcal.12090.
Faber, J. M., Luyten, H., & Visscher, A. J. (2016). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education, 106*, 83–96. https://doi.org/10.1016/j.compedu.2016.12.001.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. http://dx.doi.org/10.3102/003465430298487.
Hellrung, K., & Hartig, J. (2013). Understanding and using feedback – a review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review, 9*, 174–190. http://dx.doi.org/10.1016/j.edurev.2012.09.001.
Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education, 94*, 102–119. http://dx.doi.org/10.1016/j.compedu.2015.11.013.
Hwang, G.-J., & Tsai, C.-C. (2011). Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 42*(4), E65–E70. http://dx.doi.org/10.1111/j.1467-8535.2011.01183.x.
IBM Corp Released (2013). *IBM SPSS statistics for windows.* Armonk, NY: IBM CorpVersion 22.0.
Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. http://dx.doi.org/10.1111/j.1745-3992.2011.00220.x.
Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254.
Koedinger, K. R., McLaughlin, E. a, & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research, 43*(4), 489–510. http://dx.doi.org/10.2190/EC.43.4.d.
Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis, 35*(4), 481–499. http://dx.doi.org/10.3102/0162373713498930.
Lipnevich, A. a, & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*(4), 319–333. http://dx.doi.org/10.1037/a0017841.
Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705–717. http://dx.doi.org/10.1037//0003-066X.57.9.705.
Lysakowski, R. S., & Walberg, H. J. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal, 19*(4), 559–578.
Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61–68. http://dx.doi.org/10.1016/j.ijer.2012.12.002.
Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2016). Validating a model of effective teaching behaviour of pre-service teachers. *Teachers and Teaching*. http://dx.doi.org/10.1080/13540602.2016.1211102.
Molenaar, I., & van Campen, C. (n.d.). Learning analytics in practice the effects of adaptive educational technology Snappet on students ' arithmetic skills. Nijmegen. (Unpublished results).
Muis, K. R., Ranellucci, J., Trevors, G., & Duffy, M. C. (2015). The effects of technology-mediated immediate feedback on kindergarten students' attitudes, emotions, engagement and learning outcomes during literacy skills development. *Learning and Instruction, 38*, 1–13. http://dx.doi.org/10.1016/j.learninstruc.2015.02.001.
Pilli, O., & Aksu, M. (2013). The effects of computer-assisted instruction on the achievement, attitudes and retention of fourth grade mathematics students in North Cyprus. *Computers & Education, 62*, 62–71. http://dx.doi.org/10.1016/j.compedu.2012.10.010.
Sheard, M., Chambers, B., & Elliott, L. (2012). Effects of technology-enhanced formative assessment on achievement in primary grammar. Retrieved from the University of York website https://www.york.ac.uk/media/iee/documents/QfLGrammarReport_Sept2012.pdf.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. http://dx.doi.org/10.3102/0034654307313795.
Slavin, R. E. (2013). Effective programmes in reading and mathematics: Lessons from the best evidence encyclopaedia 1. *School Effectiveness and School Improvement, 24*(4), 383–391. http://dx.doi.org/10.1080/09243453.2013.797913.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.
Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education, 94*, 252–275. http://dx.doi.org/10.1016/j.compedu.2015.11.008.
Timmers, C. F., Braber-van den Broek, J., & van den Berg, S. M. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & Education, 60*(1), 25–31. http://dx.doi.org/10.1016/j.compedu.2012.07.007.
Timmers, C., & Veldkamp, B. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education, 56*(3), 923–930. http://dx.doi.org/10.1016/j.compedu.2010.11.007.
Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. http://dx.doi.org/10.3102/0034654314564881.
Wang, T.-H. (2011). Implementation of Web-based dynamic assessment in facilitating junior high school students to learn mathematics. *Computers & Education, 56*(4), 1062–1071. http://dx.doi.org/10.1016/j.compedu.2010.09.014.
Wang, T.-H. (2014). Developing an assessment-centered e-Learning system for improving student learning effectiveness. *Computers & Education, 73*, 189–203. http://dx.doi.org/10.1016/j.compedu.2013.12.002.
Wollscheid, S., Sjaastad, J., & Tømte, C. (2016). The impact of digital devices vs. pen(cil) and paper on primary school students' writing skills – a research review. *Computers & Education, 95*, 19–35. http://dx.doi.org/10.1016/j.compedu.2015.12.001.
Yeh, S. S. (2009). The cost-effectiveness of raising teacher quality. *Educational Research Review, 4*(3), 220–232. http://dx.doi.org/10.1016/j.edurev.2008.06.002.