The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment

Janke M. Fabera,*, Hans Luytena, Adrie J. Visscherc

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

Abstract

In this study a randomized experimental design was used to examine the effects of a digital formative assessment tool on mathematics achievement and motivation in grade three primary education (n schools = 79, n students = 1808). Experimental schools used a digital formative assessment tool whereas control schools used their regular teaching methods and materials. The tool provides student feedback, feedback to teachers, and adaptive assignments. Data included standardized achievement pre-posttest data, student motivation survey data, classroom observation data, and student log files. Multilevel analysis revealed positive effects on student achievement and motivation. Students’ intensity of use measurements support the effects found on student achievement and motivation. Furthermore, achievement effects were higher for high-performing students.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The use of digital learning tools in education has increased in recent years (Hwang & Tsai, 2011; OECD, 2015; Sung, Chang, & Liu, 2016), and so has the impact of such tools on teaching and learning processes in classrooms. It is therefore important to study the effects of such tools, and how they can be implemented best (Cheung & Slavin, 2013; De Witte, Haelermans, & Rogge, 2015). Digital learning tools offer various advantages compared with traditional learning tools, and meta-analysis findings already revealed positive effects on achievement (Cheung & Slavin, 2013). In particular the use of these tools for formative assessment can be beneficial (Haelermans & Ghysels, 2015; Sung et al., 2016). With digital learning tools like mobile devices, or individualized online learning environments students can practice with assignments adapted to their learning needs. Moreover, students can receive feedback immediately after assignments or assessments are completed (Bokhove & Drijvers, 2012b; Pilli & Aksu, 2013; Van der Kleij, Feskens, & Eggen, 2015; Wang, 2014). Teachers can receive feedback on the progress of their students, at the level of the individual student (e.g. feedback on a specific learning strategy used by a student), and at the classroom level (e.g. a comparison between class performance and national benchmarks) (Koedinger, McLaughlin, & Heffernan, 2010; Pape et al., 2012).

* Corresponding author.

E-mail addresses: j.m.faber@utwente.nl (J.M. Faber), j.w.luyten@utwente.nl (H. Luyten), a.j.visscher@utwente.nl (A.J. Visscher).
Research is necessary to clarify whether, and how digital formative assessment tools (DFATs) contribute to the quality of teaching and learning. The effectiveness of the use of these tools among other things will depend on the level of implementation in the classroom (De Witte et al., 2015). Teachers can have reservations concerning new technologies (Haelermans & Ghysels, 2015). Furthermore, Voogt et al. (2013) argue that most teachers use digital tools to improve student learning, and to a lesser extent for the evaluation and improvement of their own instructional activities. The effectiveness of student feedback in DFATs has been studied extensively and in most studies positive effects on educational outcomes are found (Bokhove & Drijvers, 2012b; Koedinger et al., 2010; Pape et al., 2012; Sheard, Chambers, & Elliott, 2012; Wang, 2011). However, further research remains important as the studies are often based on small samples in combination with a short study duration (Muis, Ranellucci, Trevors, & Duffy, 2015; Narciss et al., 2014; Sung et al., 2016; Wang, 2011, 2014). Especially for investigating motivation effects this can be problematic, as such effects can be caused by novelty effects which do not sustain (Sung et al., 2016). Furthermore, the formative assessment tools and the regular curriculum often are not integrated. As a result, strong conclusions cannot be drawn about the contribution of the tools to teaching and learning in real classrooms settings. Finally, the lack of randomized experiments (Bokhove & Drijvers, 2012a; Hunsu, Adesope, & Bayly, 2016; Koedinger et al., 2010), makes it hard to attribute effects to DFATs, especially when no data on the intensity of the use of the tool is collected (Haelermans & Ghysels, 2015; Hunsu et al., 2016).

This study examines the effects of a digital formative assessment tool called Snappet on student achievement and student motivation (1), it examines to what extent the intensity of use of DFATs affects student achievement and motivation (2), and whether student achievement effects vary between low-performing, average and high-performing students (3). A randomized experiment was conducted to examine the effects on mathematics achievement and motivation in grade three of primary education (most students are eight or nine years old). First, we now will summarize the research on DFATs comparable with Snappet. We start our review with a brief description of Hattie and Timperley’s (2007) model of feedback.

2. Review of the research literature

Digital learning tools might improve the effectiveness of formative assessment activities in classrooms (Cheung & Slavin, 2013; Sung et al., 2016). Black and Wiliam (1998) define formative assessment 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 (p.82)”. Feedback is an important part of formative assessment, and the effectiveness of DFATs will therefore depend to a great extent on the nature of the feedback given by such tools. The impact of feedback on learning outcomes has been under research for several decades (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Lysakowski & Walberg, 1982). Meta-analysis findings have revealed that the effectiveness of feedback depends on the content of the feedback (e.g. information about the correctness of an answer, or norm-referenced feedback), the timing of the feedback (i.e. direct, or delayed feedback), and the context in which the feedback is given (e.g. the learning domain, the type of assessment) (Bangert-Drowns et al., 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Ideally, the feedback content and the timing of the feedback are adapted to the context in which the feedback is given.

Following Kluger and DeNisi’s (1996) meta-analysis, Hattie and Timperley (2007) developed a model to explain how feedback can be effective. First, this model implies that students and teachers need an answer to three feedback questions to improve teaching and learning processes:

1. Where am I going (feed up)? Feedback which answers this question helps students and teachers to construct attainable and challenging learning goals. Learning goals are necessary to direct learning activities and to monitor progress towards those goals (Locke & Latham, 2002).
2. How am I going (feedback)? Feedback answering this question gives students and teachers information about the progress between their current learning levels and the intended learning goals.
3. Where to next (feed forward)? Students and teachers need feedback about learning activities which support, or hinder their progress towards learning goals (Hattie & Timperley, 2007).

Second, Hattie and Timperley (2007) distinguish four feedback levels in their model: the task, the processing of the task, self-regulation, and the self-level. Feedback directed at the task level gives information about a specific task. Task level feedback emphasizes often one solution, or learning strategy, and can hinder the transition to more complex tasks and the development of new learning strategies which can be beneficial for other tasks (Kluger & DeNisi, 1996). Feedback directed at the level of the processing of a task can be supportive for the development of new learning strategies. Feedback at the process level is often more effective for complex tasks than for simple tasks, while feedback at the task level seems to be more effective for simple tasks. The third feedback level is the level of self-regulation. Feedback directed at this level explains how effective learning tasks are executed by the student, or the teacher. Feedback directed at the self-level has no relation with learning tasks but refers to the self as a person. Feedback directed at the self-level often is not effective for learning, while feedback at the other three levels can improve learning processes.

The studied digital formative assessment tool (i.e. Snappet) includes three components: feedback to students, feedback to teachers, and adaptive assignments. The findings of research on those three components in computer-based environments are presented below.
2.1. Student feedback

Feedback effectiveness studies reveal the extent to which Hattie and Timperley’s (2007) feedback model is supported by research findings. Shute (2008) reviewed studies on the effects of student feedback. According to Shute (2008), a stepwise presentation of feedback is often more effective than presenting all feedback together, as feedback overload can be overwhelming and the feedback can be rejected. Feedback is more effective if students have time to independently work on a learning task. So, delayed feedback is often more effective than direct feedback, while low-performing students seem to gain more from direct feedback. Feedback proves to be less effective if it includes praise, or a comparison with the performance of other students (Kluger & DeNisi, 1996; Shute, 2008). This finding supports Hattie and Timperley’s model (2007), who argued that self-level feedback is ineffective as it focuses on the self instead of on learning tasks.

In other studies the effectiveness of student feedback was examined in a computer-based environment. Feedback provided by computers in many cases is delivered faster to students than feedback given by teachers, and different students can receive feedback at the same moment. Findings in a study with college students showed that there was no difference in effects on writing abilities between feedback perceived by students as computer-generated, and feedback perceived by them as teacher-generated (Lipnevich & Smith, 2009), indicating that the content of feedback matters most, and the feedback modality (teacher feedback versus digital feedback) to a lesser extent. Findings indicate small positive effects of ‘knowledge of results’ feedback (i.e. feedback which provides the correct answer), and feedback which informs students whether their answer was wrong, or correct. However, larger positive effects were found for elaborated feedback (Lipnevich & Smith, 2009; Van der Kleij et al., 2015). Examples of elaborated feedback are feedback which informs students on which mistakes were made, and feedback which gives an explanation of the correct answer. Elaborated feedback, especially is more effective than simple feedback (knowledge of results), if students work on complex assignments (Van der Kleij et al., 2015). Studies into the effectiveness of a digital formative assessment algebra (Bokhove & Drijvers, 2012a) and mathematics tool (Wang, 2011) support this notion. However, achievement effects were lower for high-performing students (Bokhove & Drijvers, 2012a).

Overall, these findings indicate that student feedback effects are influenced by student characteristics (e.g. low-performing or high-performing) and the complexity of assignments (Narciss et al., 2014; Shute, 2008), so, ideally DFATs adapt the feedback they provide based on such characteristics and assignments.

2.2. Adaptive assignments

DFATs often provide assignments adapted to individual students’ learning needs. Haelermans and Ghysels (2015) found positive effects on numeracy achievement of seventh grade students and they argue that the effects were caused by the personalized nature of the formative assessment tool. Wang (2014) also found positive effects of adaptive learning materials and assignments. Those effects were, however, only positive for low-performing students. The author states that without adaptive assignments and materials, low-performing students tend to be disorientated in a digital environment, while high-performing students have less difficulty with the selection of appropriate materials and assignments. In contrast, Cornelisz, Van Klaveren, and Vonk (2015) reported no effects of adaptive assignments and even found a negative effect on the achievement of high-performing students. They argue that the non-adaptive assignments prepared students better for the summative test than the adaptive assignments in their tool did.

2.3. Feedback to teachers

DFATs give teachers feedback about the performance and progress of individual students as well as feedback aggregated to the class level. The latter informs teachers about the performance and progress of the whole class. Teacher feedback effectiveness studies report inconsistent findings; some found positive effects on student achievement (Nunney, Ross, & McDonald, 2006; Pape et al., 2012); some only found positive effects for specific student age groups (Konstantopoulos, Miller, & van der Ploeg, 2013); and in other studies positive effects were found after a correction for the intensity of feedback use by teachers (Ysseldyke & Bolt, 2007). An explanation for the aforementioned effect differences of feedback to teachers may be that teachers do not always necessarily use the feedback they receive. Effects will be stronger if teachers use the feedback to improve instruction, and if their instruction is more aligned with the learning needs of students (Ysseldyke & Bolt, 2007). The frequency of teacher feedback might also explain differences in the effects found. Not all assessments are frequently administered, and as a result the feedback may not reflect students’ actual learning needs (Konstantopoulos et al., 2013). Besides the frequency of the feedback other DFATs characteristics also influence the effectiveness of teacher feedback. The timing of the feedback for example is important: teachers who receive feedback immediately after assignments have been completed by their students may make a better connection between instruction and student achievement, than teachers who receive delayed feedback (Hellung & Hartig, 2013; Yeh, 2009). The accessibility of the feedback also influences its effectiveness (Vischer, 2015). Clear graphical representations of class and/or student progress, for example, make feedback more accessible for teachers. Another important characteristic regarding the effectiveness of DFATs is the content of the teacher feedback: feedback can be criterion-, norm-, or individual-referenced (Hellung & Hartig, 2013), it can be specific and inform teachers about those elements of a learning domain students (do not) master (Bangert-Drowns et al., 1991), but it can also be more broad feedback, and only indicate the number of correct student answers.
Audience response programs have small positive effects on student achievement (Hunsu et al., 2016). Such programs aggregate students' responses on questions formulated by teachers. These programs are also used to foster student engagement and classroom discussion (Pape et al., 2012; Shirley, Irving, Sanalan, Pape, & Owens, 2011; Shirley & Irving, 2015). Teacher interviews have revealed that the implementation of audience response programs in classrooms is stimulated by collegial support and technological support. Curriculum pressure (i.e. the pressure of summative assessments which only include items about the subject matter content included in the curricula) (Shirley et al., 2011), and time pressure proved to hinder the implementation of these programs. Furthermore, teachers experienced and valued the improved monitoring and overview of student progress. The teacher feedback informed them on all students' learning needs rather than only on the needs of those students who quickly respond, or to whom teacher questions usually are directed (Shirley et al., 2011).

Some DFATs combine student and teacher feedback: ASSISTments, Questions for Learning and Gotit?! are examples of such tools. Positive effects of ASSISTments and Gotit?! were found for mathematics (De Witte et al., 2015; Koedinger et al., 2010). Questions for Learning had a positive effect on student achievement for grammar (Sheard et al., 2012). Low-performing students, and students who used these tools more intensively seemed to profit to a greater extent (De Witte et al., 2015; Koedinger et al., 2010; Sheard et al., 2010). Furthermore, achievement effects of student feedback were stronger, than effects of teacher feedback (Koedinger et al., 2010). ASSISTment was more effective for students who did not use the tool intensively but whose teacher often read the teacher feedback, than such students with teachers who did not read the teacher feedback regularly (Koedinger et al., 2010). Furthermore, teachers used the feedback more to plan their next day instruction, than to adapt their instruction during the current lesson (Sheard et al., 2012).

2.4. Motivation effects

In just a few studies on the effectiveness of DFATs motivation effects are examined, while DFATs can enhance student motivation. Examples of motivation supportive conditions are (positive) feedback, and the personalized character of adaptive assignments. Both might foster feelings of competence. However, negative feedback and a highly structured digital environment might also reduce students of autonomy, and, as a consequence, motivation (Ryan & Deci, 2000). According to Ryan & Deci’s, (2000) Self-determination Theory (SDT), conditions which foster competence and autonomy feelings facilitate intrinsic motivation. More specifically, feelings of competence influence motivation only if accompanied by a sense of autonomy. Meta-analysis findings reveal that interventions based on the SDT produce positive effects at the student level in terms of self-reported interest, achievement, and behavioral outcomes (Lazowski & Hulleman, 2015). Hunsu et al. (2016) reported positive effects of audience response programs on non-cognitive outcomes like student engagement and interest. Also, Pilli and Aksu (2013) found a positive effect of a digital formative assessment tool on students' attitudes towards mathematics. Furthermore, motivation effects might improve students’ feedback use, as research findings indicate that motivation is related to positive feedback behavior like feedback seeking, feedback study time and attention paid to feedback (Timmers, Braber Van den Broek, & Van den Berg, 2013; Timmers & Veldkamp, 2011). These positive findings contrast with the findings of Muis et al. (2015) who found negative effects on the motivation of five year old children. One possible explanation for these findings may be that effects on motivation are negative if the feedback is negative, and if feelings of competence are high. Effects are positive if the feedback is positive, and if feelings of competence are low, whereas effects are neutral if the feedback confirms competence feelings (Carver & Scheier, 1990).

2.5. Hypotheses

Overall, based on the literature, it was expected that the digital formative assessment tool studied (i.e. Snappet) would have a positive effect on mathematics achievement (hypothesis 1). Furthermore, it was expected that Snappet would have a positive effect on student motivation (hypothesis 2). Both student and teacher feedback are expected to enhance student effects. Therefore, it is expected that Snappet is more effective if students (hypothesis 3), and teachers (hypothesis 4) use Snappet to a greater extent. Previous research findings indicate that student feedback effects are influenced by student characteristics. Thus, it is expected that the effect of Snappet differs between low-performing, average, and high-performing students (hypothesis 5).

3. Method

3.1. Participants

Ninety-seven primary schools were recruited for this study. Schools in the Dutch province Overijssel (East of the Netherlands) were informed about the project by e-mail and after a week school principals were asked whether they were willing to participate, or not. Schools could only participate if their teachers and students of grade three did not have any experience with Snappet. Grade three students are eight or nine years old. Participating schools were randomly assigned to the experimental condition (n = 40), the control condition (n = 50), or to a waiting list (n = 7). Two experimental and eleven control schools decided not to participate after the randomization, and as a consequence, two schools from the waiting list were added to the experimental group. The randomization was at the school level, not at the classroom level, as school principals decided which of their grade three classes would participate in the study. The initial sample comprised seventy-
nine schools: forty experimental schools with 822 students, and thirty-nine control schools with 986 students. In the experimental condition, 53% of the students were male. In the control condition this percentage was somewhat lower (47%). In the experimental condition, the percentage of disadvantaged students was nine, in the control group it was six (in the Netherlands, primary schools receive extra funding for such students: a child belongs to this category if neither of both parents attained a higher educational qualification than lower vocational education). Experimental schools used Snappet for five months (between January the 19th, and July the 4th), during the same period control schools used their regular teaching methods and materials (‘business as usual’).

3.2. The digital formative assessment tool: Snappet

With Snappet students complete assignments on their own Snappet tablets. These assignments are comparable to the assignments used in traditional paper-based settings, and with Snappet the same instructional content is taught as in the regular curriculum. Fig. 1 respectively shows students’ start screen, an assignment example, and the feedback a student receives after (s)he has completed an assignment. In the start screen the assignments belonging to one lesson are presented: green blocks represent correct answers, red blocks incorrect answers, whereas orange blocks represent corrected assignments, and blue blocks new assignments. Students open assignments from the start screen and receive simple feedback immediately after they have completed an assignment. The feedback informs students whether their answer was correct (green curl) or incorrect (red cross). By ticking the plus sign in the start screen students go to their adaptive assignments. Predicted student ability levels are the basis for selecting assignments of matching ability levels. An Item Response Theory model is used to predict student ability levels on the basis of previous responses. Ability levels are recalculated after a new assignment has been completed by the student. Teachers decide which assignments students need to work on: curriculum assignments belonging to one lesson, the adaptive assignments, or both. In most lessons students first complete all curriculum assignments, and then work on adaptive assignments during the remaining time of the lesson.

Teachers follow students’ progress on their dashboard. This dashboard includes three options: teaching, progress monitoring, and extras. With the teaching option, teachers preview assignments belonging to a lesson, and they select assignments for their students. With the progress monitoring option, teachers follow the progress of a lesson, of an individual student, or of the entire class. The progress of a lesson option shows to teachers how many assignments each student completes during a lesson, and whether student answers are correct, or incorrect. Information on the progress of individual students informs teachers about how a student performs on a specific learning goal (e.g., add numbers till 100) compared with that student’s performance on other learning goals. An example of the class progress monitoring option is presented in Fig. 2. Teachers see how their class performs on learning goals compared with the performance of other Dutch classes who also use Snappet, also teachers see how each student performs on specific learning goals compared with the students’ performance on other learning goals. With the option ‘extra’, teachers can select a quiz function, a timer function, and they can also use instruction videos.

![Fig. 1. The Snappet environment for students. Image left: assignments overview. The first column on the left presents various lessons (lessons 1–10). The blocks after each lesson represent the curriculum assignments belonging to a lesson. If students click on the plus sign with the red circle an adaptive assignment is presented on the screen. Middle image: assignment example in Dutch: “add the following numbers: 3 + 5 + 7 + 5”. Image on the right: student feedback. Positive (green curl), and negative feedback (red line) in response to the answer given (correct/false). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)](image-url)
Teachers in the experimental condition followed an introduction lesson (one afternoon) in which they received information about how to integrate Snappet in their classroom lessons. Furthermore, teachers could follow an additional lesson about how to interpret and use Snappet feedback for differentiating instruction. All teachers were supervised by a coach who they could consult in case of difficulties with the use of Snappet. Twenty-four teachers attended the introduction lesson, the additional lesson, and were supervised at least three times by their coach. Four teachers followed the introduction and the additional lesson, ten teachers only attended the introduction lesson, and two teachers did not participate in these activities at all (non-response of four teachers).

3.3. Instruments and data collection

Data were collected using four instruments: standardized assessments, a student motivation survey, classroom observations, and student log files. Descriptive statistics for these instruments are presented in Table 1. We informed all parents by email/letter about the study and the data collection procedures, and no data were collected from children whose parents did not provide permission for the participation of their children.

Standardized assessments. Cito standardized mathematics assessments (Cito is the Dutch Institute for Test Development) were used to measure student achievement. Cito spelling assessments were used to control for comparability between experimental and control groups. Most Dutch primary schools use these assessments during the entire primary school period. Scores from different grades can be expressed on the same ability scale. Mathematics assessments in grade three measure two domains: arithmetic and geometry, time, and money calculations. Cito assessments are administered twice a year in January, or February, and in June. We used two mathematics student achievement variables in the analyses: posttest (the June 2015 assessment), and pretest (the January/February 2015 assessment).

Student motivation survey. A student survey was used to measure students’ mathematics motivation. The following five items were included in the mathematics motivation scale: Mathematics lessons are boring (1), I like mathematics (2), I enjoy
Table 1
Descriptive statistics.

<table>
<thead>
<tr>
<th></th>
<th>Experimental n = 822</th>
<th>Control n = 986</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male)</td>
<td>53% (12.8) 819</td>
<td>47% (12.8) 957</td>
</tr>
<tr>
<td>% disadvantaged</td>
<td>9% (1.3) 533</td>
<td>6% (1.3) 707</td>
</tr>
<tr>
<td>Posttest mathematics</td>
<td>83.0 (13.2) 803</td>
<td>81.7 (12.2) 971</td>
</tr>
<tr>
<td>Pretest mathematics</td>
<td>73.0 (13.7) 810</td>
<td>74.7 (13.5) 974</td>
</tr>
<tr>
<td>Mathematics June</td>
<td>65.5 (14.1) 750</td>
<td>66.0 (14.1) 882</td>
</tr>
<tr>
<td>Spelling January/</td>
<td>128.7 (6.7) 619</td>
<td>128.6 (6.6) 963</td>
</tr>
<tr>
<td>February 2015</td>
<td>122.7 (6.6) 555</td>
<td>123.1 (6.8) 858</td>
</tr>
<tr>
<td>Motivation*</td>
<td>3.74 (0.90) 778</td>
<td>3.60 (0.97) 772</td>
</tr>
<tr>
<td>Total assignments</td>
<td>2947.6 (1116.8) 792</td>
<td>2547.6 (1216.8)</td>
</tr>
<tr>
<td>% adaptive assignments</td>
<td>24.9 (12.8) 820</td>
<td></td>
</tr>
</tbody>
</table>

* scale is 1 (low) − 5 (high).

doing mathematics assignments (3), I think mathematics is interesting (4), I think mathematics is important (5) (α = 0.84). Items were based on two Dutch studies, the COOL 5–18 cohort study Driessen, Mulder & Roeleveld (2012), and the Intrinsic Motivation Inventory (Meijer, Eck, & Felix, 2008). Items were measured on a five-point Likert scale (1 = strong disagree, 5 = strong agree), and the reliability of this scale was calculated using Cronbach’s alpha. Experimental and control students responded to the survey at the end of the intervention. Students from all experimental schools, and from thirty-three control schools filled out the questionnaires.

Classroom observations. Classroom observations were conducted in the experimental schools only. They were used to measure the degree to which teachers, based on Snappet feedback, differentiated their teaching in line with the instructional needs of individual students, and/or groups of students. In general three mathematics lessons of each teacher were observed, and the first observation took place at the end of April, the second at the end of May, and the last observations were conducted at the end of June. Due to scheduling problems and maternity leave, of five teachers only two lessons could be observed. In each observation an entire lesson was observed, including the introduction to new learning topics, the completion of assignments by students, and a lesson evaluation (the average observation time was 54 min). The observation instrument was developed by the researchers based on the differentiation literature (Roy, Guay, & Valois, 2013; Tomlinson et al., 2003), and in line with the differentiation support provided by the Snappet feedback. In a pilot study in fifteen additional classrooms the instrument was tested and optimized. The initial instrument included eight items which were scored on a five-point scale (low - high). The Cronbach’s alpha coefficient for the reliability of the instrument was α = 0.69 (see Appendix A for the final instrument).

Lessons were rated by three raters after they had been trained by means of recorded classroom observations. During the training the recorded lessons were rated and discussed afterwards (to maximize rater reliability). The training lasted approximately seven hours (two afternoons). Raters’ scores at the end of the training were used to calculate the intraclass correlation coefficient (ICC) as an indicator of the agreement between raters. In the actual study each lesson was observed by one rater. The ICC was 0.50 (reliability interval: 0.05−0.77, Model Two way random, Consistency, IBM SPSS Statistics version 22), indicating a moderate agreement between raters.

Snappet log files. Snappet log files were used to measure students’ intensity of Snappet use. Measures that were included in the analyses were the number of total assignments completed, and the percentage of adaptive assignments out of the total number of assignments completed. These data were registered by the Snappet software and provided by Snappet for the present study.

3.4. Data analysis

The quality of the randomization was examined using an independent two sample t-test. We used students’ gender, the percentage of disadvantage students, Cito mathematics and spelling results of June 2014, and January/February 2015, to examine the comparability of experimental and control groups before the intervention.

In order to test the five hypotheses a multilevel regression model was used, as the students are nested within classes (Snijders & Bosker, 1999). Fixed effects were calculated for all variables except for the random intercept effect. The model was built up from an empty model with only student mathematics posttest scores, or student motivation as dependent variables, to the final model that included the interaction effect between student pretests scores and the Snappet condition. This
interaction effect was examined in order to answer hypothesis 5 (the expectation that the effect of Snappet differs between low-performing, average and high-performing students). The scores on all variables included in the multilevel analysis were converted to z-scores except for the dichotomous variables student gender (0 = girl), and the Snappet condition (0 = control group). This facilitates the interpretation of the multilevel findings. Effects on student achievement and student motivation were examined using two separate analyses. All analyses were performed using IBM SPSS Statistics version 22.

4. Results

The results of the independent t-test are reported first, thereafter the effects on student achievement and motivation are presented. No significant difference was found between the experimental and control condition with respect to the percentage of disadvantaged students (t = 1.00, p = 0.32), mathematics achievement in June 2014 (t = 0.68, p = 0.50), spelling achievement in June 2014 (t = 1.08, p = 0.28), and spelling achievement in January/February 2015 (t = 0.41, p = 0.68). The percentage of boys was significantly higher in the experimental schools (t = 2.34, p < 0.05). Furthermore, a 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 condition performed lower than students in the control condition at the start of the intervention. Gender, and pretest mathematics scores of January/February 2015 were therefore both included as covariates in the multilevel analyses.

The results of the multilevel analysis predicting student achievement are shown in Table 2. The null model included only the intercept, the results show that a large proportion of the variance in student achievement is situated at the student level. The proportion of group level variance of the total variance was 0.05. Covariates were included in the first model. Pretest scores (β = 0.82, p < 0.01) and gender (β = 0.06, p < 0.05) had a significant effect on student achievement, indicating that students with high pretest scores performed higher on the posttest, and that boys performed slightly better than girls on the posttest. There was a significant and positive effect of Snappet (β = 0.21, p < 0.01) (hypothesis 1 confirmed). Students’ achievement in the experimental condition was significantly higher than the achievements of students in the control condition at the end of the intervention period. Their advantage amounts to one fifth of a standard deviation over the five month period. Interestingly, the interaction reveals that experimental students with high pretest scores performed significantly better than students with high pretest scores in the control condition (β = 0.08, p < 0.01) (hypothesis 5 confirmed), indicating that high-performing students profit more from Snappet than lower performing students. This effect is shown in more detail in Table 3. The average learning gains are broken down by ability level. The highest ability level include the students in the top 20% on the mathematics pretest. The second ability level include the next 20%, and so on. All ability levels show more learning gain in the experimental condition, but the difference in mean achievement growth is highest for the 20% best performing students (mean achievement growth of 5.69 on mathematics assessments).

Model four and five include data of the experimental condition only. There was a significant positive effect of total assignments completed with Snappet (β = 0.16, p < 0.01) on achievement. No significant effect was found for the percentage of adaptive assignments (β = 0.05, n.s.). The correlation between total assignments and the percentage of adaptive assignments was r = 0.38, p < 0.01, indicating that these variables overlap to some extent. Together these results partly support hypothesis 3: Snappet is more effective if students use Snappet to a larger extent. A small positive effect was found for teachers classroom observations scores (β = 0.06, p < 0.05), indicating that student achievement was higher for students whose teachers scored relatively high on classroom observations (hypothesis 4 confirmed).

The results of the multilevel analysis predicting student motivation are shown in Table 4. The results in the null model show that a large proportion of the variance in student motivation scores is situated at the student level, the proportion of group level variance of the total variance was 0.08. There was a significant positive effect of the covariate pretest (β = 0.21,

### Table 2

<table>
<thead>
<tr>
<th>Predictors</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>Coeff.</td>
<td>SE</td>
<td>Coeff.</td>
<td>SE</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−0.00</td>
<td>0.04</td>
<td>−0.03</td>
<td>0.03</td>
<td>−0.13***</td>
<td>0.03</td>
</tr>
<tr>
<td>Gender (0 = girl)</td>
<td>0.06*</td>
<td>0.03</td>
<td>0.06*</td>
<td>0.03</td>
<td>0.06*</td>
<td>0.03</td>
</tr>
<tr>
<td>Pretest</td>
<td>0.82**</td>
<td>0.01</td>
<td>0.82**</td>
<td>0.01</td>
<td>0.78**</td>
<td>0.02</td>
</tr>
<tr>
<td>Snappet</td>
<td>0.21**</td>
<td>0.04</td>
<td>0.21**</td>
<td>0.04</td>
<td>0.21**</td>
<td>0.04</td>
</tr>
<tr>
<td>% adaptive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance student level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance group level</td>
<td>0.05**</td>
<td>0.02</td>
<td>0.03**</td>
<td>0.01</td>
<td>0.01**</td>
<td>0.00</td>
</tr>
<tr>
<td>N students</td>
<td>1774</td>
<td>1738</td>
<td>1738</td>
<td>1738</td>
<td>772</td>
<td>772</td>
</tr>
</tbody>
</table>

*p < 0.05 one-sided testing **p < 0.01 one-sided testing.

* Models with experimental student data only.
Further research on the effectiveness of adaptive assignments is therefore important, to specify when, and how adaptive assignments are more targeted on student learning needs than regular curriculum assignments. However, the positive effect of Snappet on student motivation was found ($\beta = 0.16$, $p < 0.05$), these findings support hypothesis 2: it was expected that Snappet would have a positive effect on student motivation. Again, only significant positive effects were found for total assignments ($\beta = 0.19$, $p < 0.01$). This indicates that students who made more assignments were also more motivated. Furthermore, no significant positive effect for teachers’ observation scores on student motivation was found ($\beta = 0.08$, n.s.).

5. Discussion and conclusions

It is important to study whether and how digital learning tools contribute to the improvement of teaching and learning processes since the use of digital learning tools in education is growing rapidly. The purpose of the present study was to examine the effects of a digital formative assessment tool (i.e. Snappet) on student achievement and motivation. A randomized experiment was conducted to examine these effects in grade three primary education. Furthermore, teachers’ and students’ intensity of DFATs-use measurements were included to gain more understanding of the use, and effects of DFATs in regular classrooms settings.

The findings of this study indicate that a digital formative assessment tool can have a positive impact on mathematics achievement of students in grade three. This finding is in line with previous research (Bokhove & Drijvers, 2012a; De Witte et al., 2015; Haelermans & Ghysels, 2015; Koedinger et al., 2010; Sheard et al., 2012; Wang, 2014). Students’ intensity of use measurements show that experimental students who used the tool to a greater extent (i.e. the total number of assignments completed) performed better than experimental students who used the tool to a lesser extent. However, we cannot state that students’ greater use of Snappet is causing the effect, since it could be that (for example) in the experimental group the more motivated students used Snappet more and that this caused the effects. Still, the intensity of use measurements support the student achievement differences found between the students (randomly) assigned to the experimental group, and the students assigned to the control group.

Student feedback and adaptive assignments are important characteristics of Snappet. Since the assignments for experimental and control students were comparable, and control students did not receive knowledge of result feedback, we could conclude that the feedback element contributed to the effect on achievement. On the other hand, the opportunity for experimental students to make more assignments and spend more time on assignments might also explain the effectiveness. Since no data on the total number of assignments completed by students in the control group was available, this explanation was not tested. It was also expected that the number of adaptive assignments would positively affect student achievement as adaptive assignments are more targeted on student learning needs than regular curriculum assignments. However, the analyses do not show a significant effect of the percentage of adaptive assignments of the total number of assignments on achievement. One explanation might be that students who complete many assignments also have a high percentage of adaptive assignments as students often completed adaptive assignments after they had finished the regular curriculum assignments. Effects of adaptive assignments on student achievement prove to differ (Cornelisz et al., 2015; Haelermans & Ghysels, 2015; Wang, 2014). Effects depend on the domain of learning, the type of assignment or assessment, and on how techniques are used for selecting adaptive assignments for individual students (Huang, Lin, & Cheng, 2009; Merrell & Tymms, 2007). Further research on the effectiveness of adaptive assignments is therefore important, to specify when, and how adaptive assignments affect student achievement in the intended ways.

Effects of feedback to teachers on student achievement were lower than effects of feedback to students. This finding is in line with the findings of Koedinger et al. (2010) who also found stronger effects of student feedback. The relationship between teachers’ observation scores (indicating their differentiation-practices with Snappet) and student achievement suggests that the effectiveness of DFATs also depends on if, and how intensive DFATs are used by teachers in the classroom. An important advantage of Snappet is that teachers received frequently feedback based on students’ daily assignments. Teachers in the present study reported a better overview of the progress of lessons, and of students’ understanding of the assignments. Furthermore, teachers reported that they were more able to respond quickly and purposefully to students’ questions and
learning needs. Since in the present study the intervention for teachers was limited and teachers’ observation scores showed that teachers on average did not use the teacher feedback to a great extent, larger effects of feedback to teachers may be feasible. Furthermore, DFATs might have a stronger impact on student achievement if the whole school team is trained, as teachers’ colleagues can have an important supportive role during the implementation of new tools (Shirley et al., 2011).

The present study shows that Snappet was most effective for high-performing students. This finding matches with the findings of another Snappet research project executed in other grades (Molenaar & Van Campen, 2015), indicating that these findings on the achievements of high-performing students are quite persistent across studies. The finding in a way is remarkable as in most studies larger effects of DFATs were found for low-performing students (Bokhove & Drijvers, 2012a; Koedinger et al., 2010; Sheard et al., 2012). A possible explanation for our finding might be that high-performing students complete more mathematics assignments by using Snappet than high-performing students in business as usual settings. Teachers decide what students will do after the curriculum assignments have been completed, and Snappet offers teachers opportunities to let students work on additional mathematics assignments. Teachers do not have these opportunities, or only to a lesser extent in business as usual settings. Moreover, students of different ability levels are more equally challenged by the adaptive assignments. As high-performing students are often less challenged by regular assignments than low-performing, and average-performing students, this in particular is in their advantage. Another possible explanation for the effect on high-performing students may be found in the Dutch educational context. International comparative research findings reveal that Dutch high-performing students perform low compared with high-performing students in other countries (Meelissen et al., 2010; Sheard et al., 2012).

The present study shows that Snappet was most effective for high-performing students. This finding matches with the findings of another Snappet research project executed in other grades (Molenaar & Van Campen, 2015), indicating that these findings on the achievements of high-performing students are quite persistent across studies. The finding in a way is remarkable as in most studies larger effects of DFATs were found for low-performing students (Bokhove & Drijvers, 2012a; Koedinger et al., 2010; Sheard et al., 2012). A possible explanation for our finding might be that high-performing students complete more mathematics assignments by using Snappet than high-performing students in business as usual settings.

Motivation effects were also examined in this study. The expected positive effects on student motivation were confirmed by the findings. Such positive effects were expected since students would receive (mostly) positive feedback during the intervention, and positive performance feedback can enhance feelings of competence (Ryan & Deci, 2000). Because of the adaptive assignments, students with Snappet can complete more assignments suited to their own learning needs, and thus have to complete fewer assignments that are too demanding. Motivation effects of Snappet might be stronger if students will have more opportunities for self-regulated learning, as Snappet is well-structured and teachers mostly decide which assignments their students will complete. Additionally, in the experimental groups high motivation correlates with high Snappet usage by students. However, it remains unclear if students in the experimental group who used Snappet to a larger extent became more motivated, or if more motivated students make more assignments than students who are less motivated.

Besides the contributions of this study there are also some limitations. Snappet included student feedback, feedback to teachers, and adaptive assignments, which all may have contributed to the effects found. The intensity of use measurements suggest which of these characteristics contributed most to the effect. However, the intensity of Snappet-use measurements were only measured in the experimental condition, making it difficult to draw strong conclusions about the contribution of the different DFATs elements on student achievement. Therefore, in further work an experimental design with three conditions (e.g., business as usual, an experimental condition with adaptive assignments, and an experimental condition without adaptive assignments) should be conducted. Second, no validated observation instrument was used since we did not find a classroom observation instrument which met our content criteria. Reliability analysis on the scores of our newly developed observation instrument revealed a moderate consistency between raters’ observation scores. Therefore, the observation scores of different rater pairs (i.e., rater 1 and 2, rater 1 and 3, rater 2 and 3) were included separately in the analysis. No differences in results were revealed, indicating that the moderate rater consistency had no impact on the conclusions drawn from teachers’ use measurements. A third study limitation is our randomization at the school level instead of at the classroom level. School principals decided which of their grade three classes would participate in the study after the randomization

| Table 4 Multilevel models predicting student motivation. |
|---|---|---|---|---|
| Model | 0 | 1 | 2 | 3* |
| Predictors | Fixed Intercept | Gender (0 = girl) | Pretest | Snappet |
| **Coeff.** | 0.00 | -0.09 | 0.21** | 0.16* |
| **SE** | 0.04 | 0.05 | 0.03 | 0.08 |
| **Coeff.** | 0.07 | -0.10 | 0.21** | 0.16** |
| **SE** | 0.05 | 0.05 | 0.03 | 0.04 |
| **Coeff.** | -0.02 | -0.18* | 0.16** | 0.16** |
| **SE** | 0.06 | 0.07 | 0.04 | 0.04 |
| **Coeff.** | 0.06 | 0.07 | 0.04 | 0.04 |
| **SE** | 0.06 | 0.06 | 0.06 | 0.06 |
| **Coeff.** | 0.08 | 0.08 | 0.08 | 0.08 |
| **SE** | 0.05 | 0.05 | 0.05 | 0.05 |

*p < 0.05 one-sided testing **p < 0.01 one-sided testing.

a Models with experimental student data only.
procedure. Principals with more than one grade three classes in their school often decided to participate with one, or two classes instead of with all grade three classes, while schools in the control condition participated with all grade three classes. This explains why in the control condition more students participated than in the experimental condition. More information on the randomization procedure can be requested from the corresponding author. Furthermore, in further work it would be interesting to study whether our findings still holds for older age groups who work on more complex assignments and where, presumably, more elaborated feedback is needed (Hattie & Timperley, 2007; Van der Kleij et al., 2015). Positive effects on student achievement are not always found in elaborated feedback studies (Van der Kleij, Eggen, Timmers, & Veldkamp, 2012), or there is no effect difference between elaborated and simple feedback (Maier, Wolf, & Randler, 2016). Therefore, further research is important to specify how (elaborated) feedback can be most beneficial, and how effects are moderated by characteristics of students and learning tasks (Maier et al., 2016; Timmers & Veldkamp, 2011; Van der Kleij et al., 2012).

Overall, the present study findings confirm that DFATs can have a positive effect on student achievement and motivation. An important contribution of the study is that the findings reveal that DFATs can also contribute to the performance of high-performing students, besides the contribution to the performance of low-performing students, which was frequently found in other studies. Furthermore, the findings indicate that the more teachers were able to differentiate their instruction, based on the Snappet feedback, the more this affected student achievement. The same applies to the extent to which students used the DFAT: more intensive use went together with higher levels of mathematics achievement, and mathematics motivation. We also found that there is room for more intensive DFAT use. It may well be that if we will be able to promote teacher DFAT use (e.g. by means of teacher professionalization), then the effects on student achievement also will be stronger.

Acknowledgements

This study was funded by Kennisnet, a public organization for education and ICT in the Netherlands. We would like to thank Snappet for making the Snappet data log files available for this study. Neither Kennisnet nor Snappet influenced the design of the study, its results, and the content of this article.

Appendix. Classroom observation instrument

The instrument included eight items scored on a five-point scale. A description of observable teacher actions is given per item. Teachers received few points for the first actions mentioned for an item, and higher scores for the actions mentioned later on for an item. Raters judged which teacher actions were (most) observed during the observations.

Item 1: Teacher’s use of the dashboard. Teacher actions:
1. The teacher has not opened the dashboard during the entire lesson (1 point).
2. The teacher has opened the dashboard but no actions, based on the feedback, were observed (2 points).
3. The teacher used the feedback during the lesson: to check students who do not progress on their assignments, to give students feedback regarding their progress, to help students who made several mistakes (3 points).
4. The teacher used Snappet feedback at the start of the lesson, to decide which assignments each student needs to complete, or to distribute students over different instruction groups (4 points).
5. During the lesson teacher actions as described under point 3 and 4 were observed (5 points).

Item 2: Teacher’s use of Snappet to improve the instruction. Teacher actions:
1. Snappet feedback was not used to improve the instruction (1 point).
2. Example assignments, or/and instruction videos were used for instruction (2 points).
3. Snappet feedback on the progress of students was used, to compliment students on their progress, or stimulate students to improve (3 points).
4. Snappet feedback on the progress of students was used to give extra attention to the instructional content and assignments most students faced difficulty with (4 points).
   a. AND: Snappet feedback on the progress of students was used to give extra instruction to specific students who faced difficulty with instructional content and assignments (5 points).

Item 3: Instruction to low-performing students. Teacher actions:
1. During the observation it was not clear which students performed low in comparison with the average performance of the class (1 point).
2. During the observation low-performing students were grouped based on their achievements scores on standardized tests. Low performing-students received no additional instruction (2 points).
3. During the observation low-performing students were grouped based on Snappet feedback. Low-performing students received no additional instruction (3 points).
4. During the observation low-performing students were grouped according to Snappet feedback. Low-performing students received additional instruction (i.e. repeated instruction) (4 points).
   a. Low-performing students received extra instruction (e.g. with other learning materials, other examples) instead of repeated instruction only (5 points).

Item 4: Instruction to high-performing students. Teacher actions:
1. During the observation it was not clear which students performed high in comparison with the average performance of the class (1 point).
2. High-performing students followed a shortened instruction (2 points).
3. High-performing students followed a shortened instruction, and were selected by the teacher based on Snappet feedback (3 points).
   a AND: the teacher pointed high-performing students to additional learning strategies during regular classroom instruction (4 points).
   b AND: the teacher provided additional instruction about more complex mathematics topics to high performing students (5 points).

Item 5: Adaptation of assignments to learning needs.
Teacher actions:
1. All students completed the same assignments (1 point).
2. Some students completed more, or fewer assignments (2 points).
3. Some students completed more, or fewer assignments. Students who completed their regular assignments thereafter completed adaptive assignments (3 points).
4. The teacher varied the number of assignments between students and ensured that all students had enough time to complete the adaptive assignments (4 points).
   a AND: ensured that students completed the adaptive assignments which belonged to the learning goals students faced difficulty with (5 points).

Item 6: Flexible classroom management.
Teacher actions:
- The teacher composed flexible ability groups based on Snappet feedback (i.e. no ability group composition based only on the standardized tests assigned twice a year).
- The teacher used assignments to support student collaboration forms.
- The teacher deviated from the planned instruction based on Snappet feedback.
- The teacher used alternative learning materials based on Snappet feedback.
- The teacher asked various (open) questions to assess students’ learning needs.
1 point: none of the teacher actions mentioned was observed.
2 points: one teacher action was observed.
3 points: two of the teacher actions were observed.
4 points: three of the teacher actions were observed.
5 points: all teacher actions were observed.

Item 7: Adaptation of learning goals to learning needs.
Teacher actions:
1. Variation in learning goals, or in teacher expectations towards students were not observed (1 point).
2. The teacher gives additional attention to students who do not understand the instructional content (2 points).
3. There is variation in learning goals, and/or teacher expectations of student performance (e.g., the teacher expects that some students complete more assignments than other students) (3 points).
4. There is variation in the number, and/or complexity of assignments, or there is variation in the instructional content between students (4 points).
5. During the lesson teacher actions as described under point 3 and 4 were observed (5 points).

Item 8: Examining of students’ understanding.
Teacher actions:
1. The teacher does not examine why students make mistakes (1 point).
2. The teacher asked with which assignments students experienced difficulties (2 points).
3. The teacher explained an assignment to all students based on Snappet feedback (i.e., an assignment which most students answered incorrectly) (3 points).
   a AND: examined why most students made mistakes in specific assignments (e.g., he/she asked several students to explain how they solved the assignment) (4 points).
4. In addition to 3 and 3a: these teacher actions were not only directed towards the whole class but also towards specific groups of students (5 points).

References


