frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 16 March 2023
DOI 10.3389/feduc.2023.1101192

OPEN ACCESS

EDITED BY
Yong Luo,
NWEA, United States

REVIEWED BY
Xiaoliang Zhou,
Australian Council for Educational Research, Australia
Ting Wang,
American Board of Family Medicine, United States
Fei Zhao,
NWEA, United States

*CORRESPONDENCE
Erika Boström
✉ erika.bostrom@umu.se
Torulf Palm
✉ torulf.palm@umu.se

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 17 November 2022
ACCEPTED 23 February 2023
PUBLISHED 16 March 2023

CITATION
Boström E and Palm T (2023) The effect of a formative assessment practice on student achievement in mathematics. *Front. Educ.* 8:1101192. doi: 10.3389/feduc.2023.1101192

The effect of a formative assessment practice on student achievement in mathematics

Erika Boström[1,2]* and Torulf Palm[1,2]*

[1]Department of Science and Mathematics Education, Umeå University, Umeå, Sweden, [2]Umeå Mathematics Education Research Centre (UMERC), Umeå University, Umeå, Sweden

Research has shown that formative assessment can enhance student learning. However, it is conceptualised and implemented in different ways, and its effects on student achievement vary. A need has been identified for experimental studies to carefully describe both the characteristics of implemented formative assessment practices and their impact on student achievement. We examined the effects on student achievement of changes in formative assessment of a random sample of 14 secondary school mathematics teachers after a professional development programme. This study describes practices implemented and students' achievement as measured by pre-tests and post-tests. We found no significant differences in achievement on the post-test, after controlling for pre-test scores, between the intervention group and control group, and no significant correlation between the number of formative assessment activities implemented and the post-test scores (controlled for the pre-test scores). We discuss characteristics of formative assessment implementations that may be critical for enhancing student achievement.

KEYWORDS

formative assessment, assessment for learning, student achievement, effect, mathematics

1. Introduction

In their seminal review of the effects of formative assessment Black and Wiliam (1998) concluded that it can significantly improve student achievement. Since then there has been a large increase in the number of research publications on formative assessment[1] (Baird et al., 2014; Hirsh and Lindberg, 2015). However, this literature shows a large variation in the size of the effects of formative assessment, and some interventions with formative assessment produced no effects at all on student achievement (Bennett, 2011; Kingston and Nash, 2011; Briggs et al., 2012).

1 We use the terms *formative assessment* and *assessment for learning* synonymously. This use of terminology is in accordance with some scholars (e.g., Black and Wiliam, 2009; Bennett, 2011; Baird et al., 2014), whilst others use the terms with somewhat different connotations (e.g., Swaffield, 2011).

1.1. Formative assessment

Differences in effects of formative assessment on student achievement may depend on the conceptualisation of formative assessment. There has been a significant amount of variation in how formative assessment is conceptualised and implemented, and some of the common conceptualisations to formative assessment were described by Baird et al. (2014).

The following well-cited definition by Black and Wiliam (2009) includes several of these conceptualisations to formative assessment:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p. 9)

The heart of this conceptualisation is the use of assessment information to adapt teaching and learning to the learning needs identified through the assessment. Thus, the 'big idea' should permeate all work with formative assessment. This idea clarifies that evidence about student learning needs to be collected, interpreted and used by teachers and learners to decide on the next steps in instruction. The term 'instruction' is used in the sense described by Black and Wiliam (2009) to include both teaching and learning, including "any activity that is intended to create learning" (Black and Wiliam, 2009, p. 10). Thus, formative assessment "is concerned with the creation of, and capitalization upon, 'moments of contingency' in instruction for the purpose of the regulation of learning processes" (Black and Wiliam, 2009, p. 10). Such moments may be created by teachers or students through planned assessment activities, and can result from using any kind of assessment procedure and artefact (e.g., tests, informal observations or dialogue) to reveal student knowledge and skills. These moments may also be noticed and acknowledged during learning activities in a lesson.

This definition affords several different approaches to formative assessment. In one approach, formative assessment is viewed as a process in which teachers assess students' learning to provide feedback to students or modify instructional activities to better meet students' learning needs. Two other approaches emphasise the importance of the students' proactive participation in the formative assessment processes. In one of these approaches the students act as support for each other's learning through peer assessment and peer-feedback, and suggest to their peers ways to reach their learning goals. In the other approach focus is on the students as self-regulated learners who use self-assessment and actions based on that self-assessment to reach their own learning goals. All three of these approaches may also include a focus on helping students to understand the learning goals. It is also possible to combine these three approaches and integrate them into a unity, which may be seen as a fourth approach to formative assessment. The emphasis on planned assessment processes, or on information gathered about student learning from informal day-to-day activities such as observations or dialogue may also vary between and within approaches.

The definition by Black and Wiliam (2009) was operationalised by Wiliam and Thompson (2008) in a form that facilitates the learning, and practical use, of formative assessment in the classroom, and this framework was used in the analysis of teacher practices in the present study. Wiliam and Thompson described formative assessment as a practice based on adherence to the 'big idea' of using evidence about student learning to adjust instruction to better meet the students' learning needs, and the use of the following five key strategies (KS) (Wiliam and Thompson, 2008):

> *KS 1.* Clarifying, sharing and understanding learning intentions and criteria for success.
>
> *KS 2.* Engineering effective classroom discussions, questions and tasks that elicit evidence of learning.
>
> *KS 3.* Providing feedback that moves learners forward.
>
> *KS 4.* Activating students as instructional resources for one another.
>
> *KS 5.* Activating students as the owners of their own learning.

The first key strategy emphasises the importance of teachers and students sharing an understanding of the learning goals. The second stresses the teacher's role in collecting evidence about student learning that can both form the basis of feedback to help meet students' learning needs and also to make better-informed decisions about how to continue with and adapt instruction. The third key strategy is providing such feedback and instructional activities that improve students' learning. The fourth and fifth strategies recognise and emphasise the roles of both teachers and students as active agents in carrying out the processes involved in the 'big idea'. For example, students may assess their peers' work and provide feedback (KS 4) and also, as self-regulated learners (KS 5) assess their own performances and decide how to take the next steps in their learning. The role of the teacher is to support the students' development of these skills and to motivate the use of these skills in practice.

The framework does not posit that any particular activities are required to carry out the key strategies, but some classroom activities may contribute more than others to the purposes of the big idea and to each key strategy. When all the strategies are used together as an inherent part of a unified classroom practice, they can support each other in facilitating student engagement and learning. For example, the students' involvement as proactive agents in the formative assessment processes as peer-assessors and self-regulated learners is facilitated by both teacher's and students' engagement in attaining a common interpretation of learning goals and success criteria. Frequent assessment of students' knowledge and skills, and valid interpretations of their responses, would facilitate the possibilities of providing appropriate feedback and instruction that often meet the students' learning needs.

Differences in formative assessment practices may also be due to not only differences in conceptualisation, but also to differences in implementation of each approach, and these differences may be quantitative or qualitative. Quantitatively, teachers may, for example, gather information about student learning and adapt their instruction or provide individual feedback to students several times each lesson, each month or each term. Qualitatively, some questions are better than others for capturing relevant student knowledge; some interpretations and inferences based on student responses may be more valid than others; some feedback and instructional

modifications may be better adapted than others to the learning needs identified in the assessment. The formative assessment practices may also be carried out representing different foundational principles. For example, some scholars, like Marshall and Drummond (2006) have described the foundational principle of formative assessment (or assessment for learning) as promotion of student autonomy. These scholars made a distinction between practices that capture the essence of this principle (i.e., practices that follow the 'spirit' of formative assessment), and procedures that do not embody this principle (procedures that adhere to the 'letter' of formative assessment). Several studies have found that many teachers focus on teacher-centred practices in which the teacher is the proactive agent in the formative assessment processes, at the expense of promoting student autonomy, even though such a focus was not in accordance with the conceptualisation of the formative assessment meant to be implemented (Jönsson et al., 2015; Wylie and Lyon, 2015).

1.2. Effects of formative assessment on student achievement

Several research reviews that include many different studies of each of the first three approaches have shown that all of these approaches to formative assessment can improve student achievement, and that the size of the effects vary substantially between individual studies that take a given approach. Starting with the first approach, several research reviews have looked at studies investigating the effects of teachers' feedback to students. Hattie and Timperley (2007) reported a mean average effect size of $d = 0.8$ with effects sizes in individual studies varying between 0 and 1.2. Wisniewski et al. (2020) found an average effect size of 0.5, and they too reported a notable variability of the effects between studies included in the review. Koenka et al. (2021) found an average effect size of 0.25 when comparing feedback as grades with no feedback, and an average effect size of 0.32 when comparing the effects of feedback provided as comments with feedback given as grades. Another review by Shute (2008) analysed earlier reviews on the effects of feedback on student achievement. She concluded that effects of feedback on student achievement varied between negative effects to very large positive effects. Yeh (2009) reviewed four studies on the effectiveness of frequent use of small computer-based tests to provide feedback and give differentiated instruction to students. The effects on student achievement in the analysed studies varied between 0.3 and 0.4.

The second approach to formative assessment comprises processes including peer-assessment and peer-feedback. In the meta-analyses by both Double et al. (2020) and Sanchez et al. (2017), the average effect size of the included studies was 0.3. The Double et al. (2020) study reported very large variation between −1 and 1.5, whilst the effects in Sanchez et al. (2017) varied between 0.2 and 0.6. As a part of their meta-analysis, Sanchez et al. (2017) also investigated the effects of the third approach to formative assessment, which emphasises self-assessment. They found similar effects from self-assessment as from peer-assessment (the average effects size was 0.3 and the effect sizes varied between −0.8 and 1.8). Earlier, Ross (2006) had reported studies showing positive effects and other studies showing negative effects from self-assessment on student achievement, but the most common effect size was at a level of approximately 0.5. Graham et al. (2015) found an average effect size of 0.6, and although Andrade (2019) did not report an average effect size in her review she concluded that all studies in her review showed a positive association between self-assessment and learning.

Studies on the effects on student achievement from the fourth approach to formative assessment, which comprises classroom practices including all of the first three approaches, are much more rare than studies on one of the three approaches individually. At this time, we have not found any reviews of the effects of the fourth approach on student achievement, only a few individual studies. A study by Wiliam et al. (2004) found a positive effect on student achievement with an effect size of 0.3, and a parallel study to the one presented here found a positive effect of 0.7 (measured at the teacher level) (Andersson and Palm, 2017). A quasi-experimental study by Wafubwa and Csíkos (2022) reported a positive effect of 0.38. Two other experimental studies analysing the effects of implementing formative assessment in line with this conceptualisation, but with a focus on peer-assessment and self-assessment, have also found statistically significant effects on student achievement (Chen et al., 2017; Chen and Andrade, 2018). The effect size in Chen et al. was 0.25 and the three effects sizes in Chen and Andrade varied between 0.15 and 0.25, with the latter being statistically significant.

Very few scholars question the potential of formative assessment, but several argue that more research is needed to establish both the size of its effect on achievement in different student populations and in different contexts, and the mechanisms by which it improves learning (Dunn and Mulvenon, 2009; Bennett, 2011; Kingston and Nash, 2011; Briggs et al., 2012; McMillan et al., 2013). Baird et al. (2014) observed that most available studies are case studies involving only a few students, and Flórez and Sammons (2013) and Wafubwa (2020) concluded that most studies of formative assessment have not measured change directly, but only through participants' perceptions. Consequently, calls have been made to complement such investigations with more experimental studies with representative samples, control groups, and pre-and post-tests to measure students' learning gains (Bennett, 2011; Flórez and Sammons, 2013; Baird et al., 2014; Wafubwa, 2020). In the time since those calls were made a number of such experimental studies have been published on the effects of peer-assessment and self-assessment, which made meta-analyses about these approaches possible (Sanchez et al., 2017; Double et al., 2020). However, as described above, experimental studies on the effects of the fourth approach, in which all of the three first approaches are included, are still rare.

1.3. Formative assessment in mathematics

Studies on the effects of formative assessment on student achievement in mathematics indicate that the conclusions drawn more generally also hold for this specific subject. The effects sizes vary substantially between different implementations within each approach, and there is a lack of experimental studies investigating the effects on student achievement from the fourth approach to formative assessment. The National Mathematics Advisory Panel (2008) conducted a meta-analysis on experimental studies within the first approach to formative assessment. They included studies investigating the effects of a regular use of brief tests for formative purposes on student achievement in mathematics, and they found an average effect size of 0.3 (varying between 0 and 0.6). These brief tests may

be paper-and-pencil tests, but there have also been a growing number of studies investigating the effects of interventions using computer programmes to aid the formative assessment processes. These interventions have often showed positive effects of different sizes on student achievement in mathematics (e.g., Yeh, 2009; Burns et al., 2010; Koedinger et al., 2010; Faber et al., 2017; Murphy et al., 2020). The computer programmes generate tests and information about students' learning needs, and depending on the intervention, either the computer programme provides students with feedback, or the intervention includes training sessions for the teachers about how to use the information from the test results to adapt their teaching to the students' identified learning needs. In a review by Rohrbeck et al. (2003), positive effects on student achievement in mathematics were found from peer-assisted learning including peer-assessment and subsequent feedback (the second approach to formative assessment). The mean effect size was 0.3, which was not statistically different from the sizes of the effects found for other subjects such as reading. The review by Palm et al. (2017) reported on studies from the first approach (teacher assessment of student learning and adapted feedback and instructional activities), and third approach to formative assessment (self-assessment with subsequent actions based on the assessment). Although the studies included revealed a large variation in the effects on student achievement in mathematics, positive effects were found from both approaches. Since many studies included in the review did not report effects sizes, no average effect size was calculated. However, the review indicated a need for studies investigating the effects of the fourth approach to formative assessment, encompassing teacher assessment of student learning followed by adapted feedback and instructional activities, as well as peer-assessment and self-assessment.

1.4. The need for experimental studies that include careful descriptions of the implemented practice

In addition to the research needs outlined in the previous sections, it is also important to conduct studies that carefully describe both the particulars of the formative assessment practices that the teachers implement and the impact of that kind of implementation on student achievement. Such studies could deepen our understanding of how formative assessment works to improve learning, which would improve our ability to predict the conditions and population groups in which formative assessment is likely to work in certain ways (Bennett, 2011; Kingston and Nash, 2011; McMillan et al., 2013). Unfortunately, such studies are scarce (Schneider and Randel, 2010). Several researchers argue that the vague description of the formative assessment practices actually implemented by the teachers in many studies (not only what they were supposed to implement) makes it difficult to connect the outcomes in terms of student achievement to the particular characteristics of those practices (Kingston and Nash, 2011; McMillan et al., 2013). The shortage of empirical studies including careful descriptions of the particular formative assessment practice implemented, representative samples, and pre- and post- measurements of actual student achievement, are particularly rare with the fourth approach to formative assessment. In their 2013 review, Flórez and Sammons (2013) identified only one quantitative study examining the effects of such a conceptualisation on student achievement, which was the aforementioned study by Wiliam et al. (2004). In 2017, a parallel study to the one presented here found that after controlling for pre-test scores, students of a random sample of teachers who implemented a more formative classroom practice in school year 4 significantly outperformed students in the control group in a post-intervention test that year (Andersson and Palm, 2017).

Studies of the effects of formative assessment with large and randomised samples and control groups require using professional development programmes (PDP) that support teachers in their development of formative assessment practices. However, high-quality formative assessments, and in particular those that belong to the fourth approach, constitute complex and advanced practices. Whilst some PDPs have been successful in accomplishing such substantially developed formative assessment practices to the extent that increased student achievement was obtained (e.g., Andersson and Palm, 2017), providing sufficient support for teachers to use these programmes to develop classroom practices in accordance with the fourth approach has proven to be difficult with large samples of teachers (e.g., Bell et al., 2008; Randel et al., 2016). There is a general consensus that certain characteristics of PDPs are important for attaining desired teacher and student outcomes (Timperley et al., 2007; Desimone, 2009); these include (1) a focus on teaching and learning subject matter, (2) active learning including hands-on practice, (3) teacher collaboration and discussions about the impact of activities tested in the teachers' classes, (4) coherence between what is being taught in the programme and wider policy trends and research, (5) time spent on the programme, and (6) engagement of school leaders and external expertise. Heitink et al. (2016) also identified similar characteristics important for developing formative assessment practices, whilst DeLuca et al. (2019) additionally suggested that identifying teachers' learning continua would be an important support for their continued development, both in terms of conceptual understanding and enacted formative assessment practices.

2. Research questions

In the present study, a random sample of year-7 mathematics teachers implemented formative assessment practices after having participated in a professional development programme (PDP) focused on the fourth approach to formative assessment as described by Wiliam and Thompson (2008). The main goal of the study was to see whether these formative assessment practices would have an effect on students' achievement. A complementary question is whether implementing more (rather than fewer) formative assessment activities would have a positive effect on the students' achievement gains (for a definition of formative assessment activities, see the next section). This complementary research question is intended to provide some empirical evidence about the issue of the effects of quantity and quality of formative assessment. For example, it might be beneficial for students' learning if teachers gathered information about their students' learning needs in many different ways, and used this information to adapt their teaching, or if teachers provided students with several different opportunities to self-assess. Indeed, these activities are central parts of many conceptualizations of formative assessment. However, another possibility is that the use of a diversity of different assessments activities and adaptation of teaching and learning do not affect student achievement, or that the quality of these practices needs to be high to achieve an effect. We have not found any study investigating this specific issue empirically using randomised samples with control groups and actual measurements of student

achievement. Such a study could be useful for understanding what constitutes both sufficient quality and sufficient quantity for formative assessment practices to have a positive effect on student achievement.

The present study investigates the following two research questions:

1. To what extent do the formative classroom practices implemented by the year-7 teachers who participated in the professional development programme affect student achievement in mathematics?
2. To what extent is there a correlation between the number of formative assessment activities implemented and students' achievement gains in mathematics?

3. Methods

3.1. Procedure

A randomised selection of schoolyear-7 teachers was invited to participate in a professional development programme in formative assessment (described below) held in the spring of 2011, and in the research study. The next school year (autumn 2011 – spring 2012), the teachers returned to full-time teaching at their schools. The rest of the schoolyear-7 teachers in the municipality were the control group. To study the effect of the implemented formative assessment practices on students' achievement in mathematics, a pre-test was administered to all schoolyear-7 students in the municipality in August 2011, and a post-test was administered to the same students near the end of May 2012 (see Figure 1 for a timeline of the data collection and teacher activities). Results for students of teachers who participated in the PDP were compared with those for students of teachers in the control group. The tests were administered by the municipality as one of their evaluations of their schools. The researchers were then allowed access to anonymized spreadsheets with the results from the tests. Since the tests were carried out under municipality regulations, and researchers only used anonymized municipality data, consent from teachers and students was not required. However, all teachers were informed that researchers would be allowed to use anonymized data from the tests.

There were only two types of contacts between the researchers and the teachers during the school year 2011–2012. One was the organisation of three 2 h sessions so the teachers who participated in the PDP could get together and discuss their formative classroom practice if they wished. Only a few did. Teachers mentioned a lack of time as a reason for not participating in these sessions. They also mentioned that they thought they knew how to implement the formative assessment practices they wanted to carry out this schoolyear, so they did not feel a need for those sessions at this time. The other type of contact was the data collection for analysing changes to the teachers' formative classroom practice after the PDP. These data collection occasions included unannounced classroom observations during the school year and both a questionnaire and an interview at the end of the school year. For these parts, consent from the teachers was obtained.

3.2. Participants

The 14 teachers (6 females, 8 males) who participated in the PDP were selected by a stratified random sampling procedure from the population of all 35 teachers who were scheduled to teach mathematics in a schoolyear-7 class (students approximately 13 years old) in the upcoming school year in a mid-sized Swedish municipality. In the selection procedure, secondary education schools were first stratified based on the number of classes in schoolyear 7, and then one to three teachers were randomly selected from each school depending on the number of classes in these schools. There were 14 schools with year-7 classes in the municipality, and each school included between one and 6 year-7 classes. For schools with an even number of year-7 classes, half of these teachers were randomly selected to participate in the PDP. For the schools with 3 year-7 classes half of them were randomly selected to participate with two teachers, and the other schools with one teacher. A similar procedure was used for schools with one or 5 year-7 classes, but they could contribute with one or two, and two or three teachers, respectively. Two of the 20 selected teachers declined to participate in the PDP. Another four had to withdraw from the study for reasons such as moving to another city or not being assigned a year-7 class after the PDP. These six teachers were not included in either the intervention group or the control group. The 14 teachers who participated in the PDP taught a total of 291 students (the intervention group). The students belonging to this group were 148 females and 143 males. The remaining 15 teachers (6 females, 9 males)

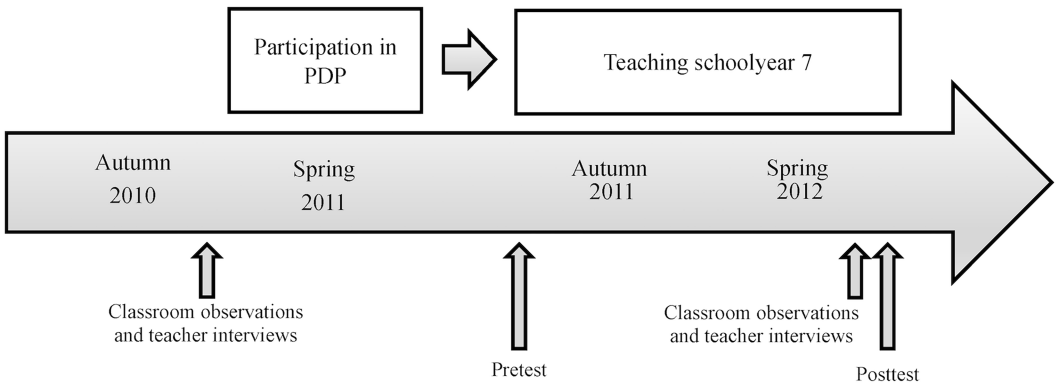

FIGURE 1
Timeline for data collection and teachers' activities.

that taught year-7 mathematics that school year taught a total of 275 students (the control group). Of these students, 141 were females and 134 were males. The teachers in the control group did not receive in-service training, and had no prior training, in formative assessment. To avoid influencing the teachers who had not participated in the professional development programme, the teachers who had participated in the programme were asked not to discuss what they had learnt in the PDP with their colleagues. Both teacher groups included teachers with varying lengths of teaching experience, ranging from teachers with only a year of teaching experience to those close to retirement. The students in both groups were diverse in their socio-economic and cultural backgrounds.

In Sweden, children are obliged to attend school for 9 years, and during these compulsory school-years the students take 16 subjects such as mathematics, history and music. There is a national curriculum that includes the learning goals in all subjects. National tests are administered in schoolyears 3, 6, and 9, and the results should be given special consideration when setting grades. The students are given grades in each subject in school years 6–9. The final grades at the end of school year 9 are used for admission to upper-secondary school programmes (school years 10–12). Almost all students go on to upper-secondary school, but not all are admitted to their first-choice programme. The teachers teaching mathematics usually follow their class through school years 7 to 9, and commonly teach 3–4 subjects (for example mathematics, physics, chemistry and biology).

3.3. Teachers' formative assessment practices

3.3.1. Design of the professional development programme

The research questions in the present study concerns the effects on student achievement from the formative assessment practices the teachers implemented after their participation in the professional development programme (PDP). Thus, the object of study is not the PDP itself or the PDPs' effects on teacher practices or student achievement. However, in order to provide context for the study we here briefly describe the PDP.

The PDP the teachers participated in was designed to have many of the characteristics identified by researchers as important for accomplishing teacher development (Timperley et al., 2007; Desimone, 2009). The programme included 4 h meetings, once a week over one term, for a total of 96 h. The teachers had another five hours per week, or 120 h in total, available for reading literature and planning and reflecting upon formative assessment activities that were new to them ("new" in the sense that they had not used them, or used them to a lesser extent, prior to the PDP). A regular meeting comprised lectures presenting the theory of formative assessment, research supporting the value of formative assessment, suggestions for concrete formative assessment activities to try out in the classroom before the next meeting, group discussions about the content and how to implement it, and discussions about the previous week's classroom try-outs. The programme was process-oriented, had a formative character, and was organised and led by the second author. The framework by Wiliam and Thompson (2008) formed the content of the PDP, and it included the general principles of formative assessment. Meetings during the PDP focussed on how these principles could be applied to the school subject mathematics, and in particular to the mathematics curriculum the teachers and their students were engaged in during the PDP. For example, when Key Strategy 1 was explored during a meeting, then the specific learning goals of the mathematics curriculum being taught during the PDP were used in the discussions of how to obtain a shared understanding of the learning goals between teacher and the students.

3.3.2. Data collection and data analysis

The formative assessment practices implemented by the year-7 teachers after the PDP were identified and described in detail in Boström and Palm (2019), and are summarised in the next section. The analysis and identification of these formative assessment practices described in Boström and Palm (2019) was possible because the teachers' practices before the PDP were described in an earlier study (Andersson et al., 2017). Thus, the teacher practices before and after the PDP could be compared. The analyses in these previous studies, which are described in detail in Andersson et al. (2017) and Boström and Palm (2019), of the teachers' practices were based on data collected through classroom observations and teacher interviews. At least four classroom observations were made for each of the 14 teachers (two before and two after the PDP). The classroom visits were unannounced in order to increase the probability of observing regular lessons. The interviews were conducted before the PDP and after the school-year following the PDP. The classroom practices of the teachers in the control group were not analysed. Since the teachers in the intervention group were randomly selected, and there were no other professional development initiatives going on at the participating schools at the time, the assumption is that both groups of teachers had similar practices before the PDP, and that only the teachers in the intervention group changed their practices after the PDP.

Changes in teachers' formative assessment practices were described in terms of the formative assessment activities they used before and after the PDP, and the framework comprising the 'big idea' and five key strategies (Wiliam and Thompson, 2008) was used in the analysis of the data. *Formative assessment activities* were defined as activities that contribute to the attainment of the goals of each key strategy and the big idea. The description of the formative assessment practices in terms of formative assessment activities does not imply that these activities were carried out in isolation. For an activity to be categorised as a formative assessment activity, it needed to be carried out in conjunction with other activities to form a formative assessment practice. For example, for an activity to be categorised as belonging to Key Strategy 2 (gathering information about students' learning needs), the information about student learning that was identified in that activity had to be used for adjusting feedback or instruction to meet those needs.

For a formative assessment activity to be considered as part of a change in a teacher's formative assessment practice it needed to be an activity that was new in the sense that the teacher had not used it (or had used it to a lesser extent) before the PDP. That activity also needed to be used regularly as a part of the new practice. For the researchers to conclude that an activity constituted a regular part of a new practice, it was required that the teacher in the interview either provided details or examples of the uses and outcomes of the activity, or that the interview data was supplemented with classroom observation data from which it could be concluded that the activity was used regularly (in other words, it was not sufficient that the teachers only said that they used the

activity). Such indications from the classroom observations include students who seemed used to, or who asked for, an activity. Examples of such classroom observations are students providing answers to teacher questions on mini-whiteboards (a technique new to many of the teachers) without asking about the procedure, and students not having to ask how to use suggestions written by the teacher on the big whiteboard about how to help them monitoring and developing their own learning. The teachers were also asked to relate their descriptions of their new practices to the lessons the researchers had observed. Thus, the interviews were the leading source of information about the teachers' practice, and the classroom observations served to validate or reject conclusions drawn from the interviews.

3.3.3. Description of formative assessment practices implemented after the PDP

All teachers implemented new formative assessment activities in their classrooms and began to use them regularly. These activities strengthened classroom practice in line with the big idea of collecting evidence of student learning in order to adjust instruction to better meet students' learning needs. Each teacher implemented from 3 to 19 new activities, with a median of 11.5. About half of the teachers complemented previous instruction with 3 to 7 new activities, whilst others made substantial changes in their classroom practice to include many new activities connected to each key strategy and to the big idea. The largest number of newly implemented activities were connected to Key Strategy 2, gathering information about student learning, and only a few were related to Key Strategy 4, activating students as instructional resources for one another.

The most common change was to more frequently elicit evidence of student learning with the purpose of adjusting instruction (activities pertaining to KS 2). The teachers implemented new small and quick assessment activities that were used regularly. Almost all teachers started to use 'exit-passes' (Wiliam, 2011), question(s) that all students were required to answer in writing at the end of the lesson. About half also started to let their students answer teacher questions during the lessons on mini-whiteboards. These 'all-response' systems provided teachers with more frequent information about all of their students' learning. Consequently, the teachers could make more frequent and well-founded adjustments to their instruction to better fit their students' learning needs. The teachers realised they now had information about all of their students' learning needs earlier than when they had gathered such information mainly from student questions during seat-work and from students who raised their hands to answer teacher questions during whole-class sessions. Using those previous assessment techniques, instruction was often adjusted based on information from only a few students. Using the new assessment techniques, with information from all students, teachers perceived they could now help more students more effectively. The teachers also perceived that the use of mini-whiteboards contributed to increased engagement and thinking among all students during whole-class sessions. Half of the teachers started to use a system of randomly distributing questions to different students, which they also felt improved the students' engagement in learning activities.

With improved assessment of their students' learning, most teachers also began to more frequently provide adjusted instruction for individual students, and about half of the teachers began to more often adjust their instructional activities for the whole class. Thus, not only did the frequency of the adjustment cycle increase, but the adjustments themselves were also better founded since teachers had more information about every student. However, some teachers mostly posed questions targeting 'basic knowledge' to examine whether students had sufficient understanding to be able to solve standard textbook tasks during seat-work, or to follow the fundamentals of an upcoming lecture. In those cases, assessments and subsequent adjustments targeted the learning needs of only a few students, and not those of students who understood the basics but were struggling for higher understanding.

The most common change in relation to Key Strategy 1, which was made by about half of the teachers, was to break the learning goals down into subgoals and present these lesson goals in the beginning of each lesson. Teachers generally also started to talk more about these learning goals with their students, and thus to focus on goals rather than on the number of tasks to be solved. However, they did not involve students in other activities that might help them to understand these goals. For example, teachers did not invite the students to actively discuss and negotiate the meaning of the goals, nor did they provide detailed examples of the goals at different levels and the criteria for attaining them, nor did they give students feedback on their interpretation of the goals.

In relation to Key Strategy 3, most teachers became more conscious of the characteristics of their feedback to students, which this resulted in their feedback being more consciously thought out. About half tried to include in their feedback two things each student had done well and one suggestion for improvement ("two stars and a wish," Wiliam, 2011). Such 'stars' are likely to be perceived as motivators, and detailed suggestions for improvement are useful for learning. However, other than during seat-work, the teachers did not set aside specific time to work with the feedback.

Only minor changes were made concerning Key strategy 4, and half of the teachers did not change their practice with respect to this key strategy at all. A few teachers encouraged their students to collaborate more by giving them group tasks and describing how to seek and provide help. In relation to Key strategy 5 the most common change was that half of the teachers discussed and decided with their students about how they, with the aid of self-assessment, could think and act when they got stuck in solving a task, in order to learn and solve that (and other) task(s). A few teachers also began to support students' self-regulated learning by giving them responsibility for correcting their diagnostic tests (something half of the teachers had already done before the PDP). However, teachers rarely used activities that specifically helped their students to take an active role in the formative assessment processes of peer assessment, peer feedback, and self-regulated learning. For example, they rarely described how students could assess themselves and their peers, what are the important characteristics of peer feedback, and how to adjust their learning. Neither did they set up activities in which students could practise and get feedback on these skills. In addition to having implemented a practice including the regular use of activities pertaining to the big idea and Key Strategies 2 and 3, most of the teachers also began to regularly use activities pertaining to either (or both) Key Strategies 4 or 5. However, much of the responsibility of the formative classroom practice remained with the teachers.

The big idea and all key strategies were focussed on in the PDP. However, the activities the teachers chose to implement after the PDP were most often those they expected they would be able to carry out successfully. Expectations of successful implementations were mainly based on whether they felt they had mastered an activity that

they tried out in their classrooms during the PDP. Another factor determining activity choice was the anticipated value and cost of the implementation. The teachers most often choose activities that they anticipated would not be too time-consuming (including both teacher preparation and teacher and student implementation time) and not too difficult to carry out (for both teacher and students), but still provide increased student engagement and learning. A teacher's beliefs about what a good teacher is and what constitutes good teaching were also considered when choosing activities to implement in their classroom practices (Boström and Palm, 2020).

3.4. Changes in student achievement

3.4.1. Data collection

A pre-test and a post-test were used to investigate possible effects on student achievement from the formative assessment practices the teachers implemented after the PDP. At the end of spring 2011, the teachers were informed that all year-7 classes in the municipality would take the pre-test at the beginning of the next school year (middle of august). At the beginning of August, before the students arrived, the teachers were provided with written information about the test, which consisted of two parts to be administered on different days, one part for which calculators were allowed and one for which they were not. Students were given 40 min to work on each part. The total maximum score on the pre-test was 60. The first part consisted of 33 tasks, including subtasks, with a total maximum score of 33, and the second consisted of 14 tasks, including subtasks, with a total maximum score of 27.

At the end of the school year, on specified days in May, the post-test was administered to all year-7 students in the municipality. At the beginning of the term, teachers were informed about the dates, although the tests and accompanying instructions were distributed to the teachers only a few days before the test was conducted. The post-test also consisted of two parts, one with and one without calculators. Students were given 40 min to work on each part. The total maximum score on the post-test was 58. The first part consisted of 31 tasks, including subtasks, with a total maximum score of 31, and the second part consisted of 17 tasks, including subtasks, with a total maximum score of 27.

The tests were developed for the municipality's evaluation purposes, and to also fit this study, by a group consisting of the authors, experienced secondary school teachers, and national test developers. The tests were designed to assess the mathematics specified in the national curriculum documents, which would provide the teachers in the municipality assistance in interpreting the new national curriculum. The new national curriculum was mostly an update of the former curriculum, although some vaguenesses in the previous formulations were clarified. The pre-test covered content that students were expected to have learnt by school year 6, and the post-test covered content that students were expected to learn in year 7. The general content areas were the same in the two tests, but on a more advanced level in the post-test. In both tests, this content included understanding and use of numbers, algebra, geometry, probability and statistics, and relationships and change. The tasks also required the same process standards in both tests. These process standards were handling of procedures, use of mathematical concepts, reasoning, problem solving and mathematical communication. The tests included tasks with multiple-choice, fill-in the blanks, and short-answer constructed-response formats, and for many tasks, students were required to show how they arrived at their answers. For the research study, the scores on the tests were used to evaluate differences in student achievement on the post-test between the intervention group and control group, controlling for initial differences in mathematical proficiency between the groups on the pre-test.

The pre-test was piloted in 2 year-6 classes (at the end of year 6), and the post-test was piloted in 2 year-7 classes in other municipalities before they were used for this study. These pilot studies were done to gather information about the students' understanding of the tasks, time for solving the tasks, and indications of distributions of scores and possible ceiling effects. A few tasks were removed from the tests as a result of the pilot studies. The piloting teachers and the test development group agreed that the tests were consistent with the national curriculum documents for the relevant school years.

The teachers in the main study received detailed instructions including, for example, how to introduce the tests to students, how much support could be given to students, how to handle missing students, and how to deliver the tests back to the municipality office. The authors and a group of experienced retired mathematics teachers were hired by the municipality to mark the tests. The markers did not know whether students belonged to the intervention group or to the control group. The municipality then compiled the results in a spreadsheet. An anonymized version of this file was made available to the researchers who transferred this data to a statistical software programme. The teachers received a file with the results of their own students so that they could, if they wished, use this data to inform their subsequent instruction. The teachers were also informed that the results of their students would not be reported to anyone else.

3.4.2. Validity and reliability of test score interpretation

An interpretation of test scores may be seen as valid to the extent that a (validity) argument is supported by appropriate (theoretical and/or empirical) evidence (Kane, 2016). In the current case the interpretation being made is that the tests provide a measure of the students' knowledge and skills in the mathematics curriculum for school-year 6 (pretest) and school-year 7 (posttest). In general, validation of such an interpretation of scores from school tests could be done by evaluating how well the content of the tests reflects the curriculum and actual classroom teaching. A common way of making such evaluations is to gather a panel of teachers teaching the content in the relevant school-years (and if available, other experts on the relevant curriculum) to review the tests. This procedure was followed in this study. A panel consisting of mathematics national test developers and mathematics teachers in the relevant school-years compared the tests with the national curriculum documents as well as with the textbooks used in the municipality (textbook writers also interpret the national curriculum documents and Swedish teachers rely heavily on the textbooks in their teaching (Skolverket, 2012). This panel also judged the difficulty of the tasks in relation to the particular student group of year-7 students. Empirical evidence of both the difficulty level and the students' understandings of the tasks were available from the pilot tests. Based on that evidence and their judgements of the contents of the tests in relation to the national curriculum documents and used textbooks, the panel unanimously concluded that the tests would provide an appropriate measure of the students' knowledge and skills in the mathematics curriculum of school-year 6 and 7, respectively.

Reliability considerations may also be seen as a part of the validation of test score interpretation. The Cronbach's alpha coefficient for the pre-test with this sample was 0.88, which suggests high internal consistency reliability (Cohen et al., 2011). For the post-test, the Cronbach's alpha coefficient was calculated as 0.92, which suggests the post-test also had high reliability. A detailed scoring procedure was put in place to secure a high level of agreement between the raters in the scoring of the tests (no measure for the inter-rater reliability was calculated). The maximum score on a subtask varied between one and two points. The raters were provided a marking scheme that had been tested in the pilot studies, and included detailed directions about the student answers required for awarding each scoring point. In addition, examples of student work were also provided to aid the scoring of some of the tasks. All five raters were sitting in the same room. They first started by all scoring the same task solutions from the same students in all tasks, and then discussed their scorings when they differed. They repeated this procedure until scoring task solutions from new students did not result in much difference in the scoring. They then divided the scoring of the students' solutions among themselves, but as soon as they were uncertain about a particular scoring, they discussed these uncertainties with the group and reached consensus.

3.4.3. Data analysis

The ultimate goal of the study was to investigate whether the formative assessment practices implemented by the teachers had any effect on the students' mathematics achievement. Since the students are nested in classes, a multilevel modelling (MLM) technique could have been suitable for investigating these effects, but for such techniques to be appropriate, it is conventional for MLM to use a minimum of 20 cases per Level 2 unit (20 students in each class) and 20 Level 2 units (20 classes) in the data. There were enough students at Level 1 to ensure there were 20 cases per class. However, there were fewer than 20 classes in both the intervention group and the control group. In addition, the intraclass correlation (ICC(2)) were 0.56 for both groups and the whole sample for the pre-test, and varied between 0.67 and 0.70 on the post-test. These values, particularly the ICC(2) values for the pre-test, indicate an insufficient degree of reliability with which class-mean ratings differ between classes for MLM to be used (acceptable value are 0.7 or higher; e.g. Marsh et al., 2012). Thus, any values at Level 2 (classroom level) would be potentially misleading if MLM would be used. However, the ICC(1) values for both the whole sample and subgroups for both tests were between 0.06 and 0.11. Such low ICC(1) values mean that the between-class variation is very small and does not contribute much to the total variation of scores (Lam et al., 2015). Thus, the possible effects of not including Level 2 in the analysis would be small (because of the small ICC(1) values). In accordance with common practice when having such low ICC(1) values, we proceeded with a Level 1 analysis only. However, when interpreting the results and answering the research questions, the possible small effects of not including Level 2 in the analysis is accounted for. It may be noted that such effects would cause negatively biassed standard errors estimated for Level 2 variables (which is the concern of interest when the intervention is delivered at the classroom level, as it is in our case), which may increase the risk of type 1 error but not type 2 error (Maas and Hox, 2004; Huang, 2016). That is, not including a Level 2 analysis would only increase the risk of concluding that a treatment is effective when it is not, and not increase the risk of concluding that a treatment is ineffective when in fact it is effective (Gorard, 2007). Since the Level 1 analysis in the present study did not detect a statistically significant effect from the intervention (see the Results section), an inclusion of Level 2 in the analysis would not produce different results.

To assess whether the formative assessment practices implemented by the teachers had any effect on the students' mathematics achievement, a one-way between group analysis of covariance (ANCOVA) was conducted using SPSS Version 27. An ANCOVA tests for significant differences in post-test scores between the students in the intervention group and the students in the control group, whilst controlling for differences in the pre-test scores by calculating an 'adjusted' mean for the post-test scores. SPSS uses regression procedures to remove the variation in the dependent variable (the post-test scores) that is due to the covariate (the pre-test scores), and then performs a normal analysis of variance on the post-test scores. To examine the ANCOVA assumptions of normality and homogeneity of variances, we used the Shapiro–Wilk test and Levene's test, respectively.

We also conducted a partial correlation analysis to study the relationship between the number of new formative assessment activities implemented by the teachers who had participated in the professional development program and their students' achievement on the post-test when controlled for the scores on the pre-test. Partial correlation is similar to Pearson product–moment correlation, except that it allows control for an additional variable (in this case, the pre-test).

4. Results

In the ANCOVA, the group variable was professional development of the students' teachers, that is, the students taught by teachers who had participated in the formative assessment programme were assigned to one group and the other students were assigned to the other group. The dependent variable was the students' scores on the post-test. The students' scores on the pre-test were used as the covariate in the analysis. Analyses of normality, linearity, and homogeneity of variances show that all prerequisites for ANCOVA were met. The Shapiro–Wilk's test ($p > 0.05$) yielded significance values above 0.8, which indicate that the mean post-test scores of both student groups were approximately normally distributed. This was also supported by visual inspection of the histograms, normal QQ-plots, and box plots of both groups' mean post-test scores. The distribution of scores indicates a linear relationship between the dependent variable and the covariate for both the intervention group and the control group. Thus, the assumption of linearity is not violated. Levene's test yielded a significance value of 0.84, which verified the equality of variances in the samples (homogeneity of variance, $p > 0.05$). In addition to the prerequisites for ANCOVA being met, a significance value of 0.89 showed that there was no statistically significant interaction between the group variable and the covariate. This supports the assumption of homogeneity of regression slopes, which enables a direct comparison of adjusted means between groups.

The results of the ANCOVA show that, after adjusting for the pre-test scores, there was no significant difference in post-test scores between the intervention group and the control group ($F(1,563) = 0.037$, $p = 0.85$, $\eta_p^2 = 0.000$). Indeed, the effect size measured by eta squared was 0.000. Thus, when controlling for achievement on the pre-test, the formative classroom practices implemented by the teachers in the intervention group did not have an effect on student

achievement on the post-test in comparison with the classroom practices of the teachers in the control group. As described in the Data analysis section above, the decision to not include Level 2 in the analysis would increase the risk of type 1 error but not type 2 error (Maas and Hox, 2004; Huang, 2016), and would therefore not increase the risk of the results not showing an effect of the intervention if there actually was one. Thus, the decision to not include Level 2 in the analysis did not affect the results of the study.

To find indications of whether students at different achievement levels were affected by the formative assessment practices, we conducted the same analysis on three student subsamples: the third of the students that had the lowest scores on the pre-test, the middle third of the students, and the third of the students that had the highest scores on the pre-test. The ANCOVAs showed that, after adjusting for the pre-test scores, there were no significant differences in post-test scores between the intervention group and the control group for either of these student subsamples ($F(1,184) = 0.59$, $p = 0.44$, $\eta_p^2 = 0.003$; $F(1,193) = 0.60$, $p = 0.44$, $\eta_p^2 = 0.003$; $F(1,180) = 1.38$, $p = 0.24$, $\eta_p^2 = 0.008$). The mean scores on the tests are presented in Table 1.

A partial correlation analysis was made to investigate a possible relationship between the number of new formative assessment activities implemented by the teachers in the intervention group and their students' achievement gains. Preliminary analyses showed no violation of the required assumptions of normality, linearity, and homoscedasticity. The partial correlation showed a weak, not statistically significant, positive partial correlation between the number of new formative assessment activities implemented by the teachers and the students' achievement on the post-test when controlled for the pre-test ($r = 0.15$, $n = 14$, $p = 0.62$). An inspection of the zero-order correlation ($r = 0.26$) suggests that controlling for the pre-test had some effect on the strength of the (non-significant) relationship between the two variables.

5. Discussion

5.1. Conclusion

The results of the study show that formative assessment, as implemented by the year-7 teachers who participated in the professional development programme, did not have a significant effect on students' achievement. The results showed no effects for either the whole intervention group or for any of the three student subsamples that differed in their achievement on the pre-test. There was also no significant correlation between the number of formative assessment activities implemented by the teachers and student achievement on the post-test when controlled for the pre-test scores.

The results of the study do not mean that formative assessment does not improve student achievement. Earlier studies have shown that practices of all four approaches to formative assessment can enhance achievement. Strongly teacher-led practices in which the teacher gathers information about student learning and provides feedback and instructional activities adapted to the identified learning needs (Hattie and Timperley, 2007; National Mathematics Advisory Panel, 2008; Yeh, 2009; Burns et al., 2010; Koedinger et al., 2010; Faber et al., 2017; Palm et al., 2017; Murphy et al., 2020), practices that focus on the student's proactive engagement in formative assessment practices such as peer-assessment and peer-feedback (Sanchez et al., 2017; Double et al., 2020) and self-assessment (Ross, 2006; Graham et al., 2015; Sanchez et al., 2017; Andrade, 2019), and practices that include all of these three approaches (Wiliam et al., 2004; Andersson and Palm, 2017) have all been shown to improve student achievement. The results of this study explicate that it is not the general approach to formative assessment in itself that is the decisive factor in whether the practice will affect student achievement (although the approach taken will provide different affordances for, and constraints upon, possible effects). Instead, they indicate that the way the approach is implemented is more essential to learning. This conclusion is consistent with reviews of the effects on student achievement from each approach to formative assessment, which all report positive average effects from the different approaches and large differences in effects sizes of the studies within each approach (Ross, 2006; Hattie and Timperley, 2007; National Mathematics Advisory Panel, 2008; Graham et al., 2015; Palm et al., 2017; Sanchez et al., 2017; Double et al., 2020). The results of the present study show that the particular ways in which the formative classroom practices were performed by the teachers were insufficient for improving student achievement. If the characteristics of the implemented activities had been different (for example if activities had included more specified instructions and practice for students about how to assess and give feedback to their

TABLE 1 Mean scores on the tests.

Group	*N*	Adjusted mean, post-test (standard error within parenthesis)	Unadjusted mean, post-test (standard deviation within parenthesis)	Mean, pre-test (standard deviation within parenthesis)
Formative assessment group				
All students	291	28.55 (0.39)	28.02 (11.19)	34.03 (9.42)
Lowest achievers	105	19.07 (0.62)	19.03 (8.36)	24.15 (4.37)
Medium achievers	96	27.71 (0.70)	27.80 (7.40)	34.52 (2.85)
Highest achievers	90	39.12 (0.68)	38.73 (7.49)	44.44 (4.62)
Control group				
All students	275	28.66 (0.40)	29.22 (10.61)	35.23 (9.56)
Lowest achievers	82	19.80 (0.71)	19.85 (7.38)	24.24 (4.44)
Medium achievers	100	28.47 (0.68)	28.39 (7.37)	34.53 (2.80)
Highest achievers	93	37.99 (0.67)	38.55 (7.78)	45.85 (4.88)

peers, or if activities would have more explicitly supported students in self-regulating their learning), there might have been a correlation between the number of implemented formative assessment activities and student achievement gains.

This study examined the effect of formative assessment practices on student achievement using measurements of actual student performance before and after the formative assessment intervention, and we have described the characteristics of these formative assessment practices. Such studies have been scarce (Schneider and Randel, 2010), and calls have been made to complement other types of research with this sort of study (Bennett, 2011; Kingston and Nash, 2011; McMillan et al., 2013). Our study addresses this need to empirically connect certain formative assessment practices with student achievement gains in different populations and contexts. In contrast to the present study, an earlier experimental study by Andersson and Palm (2017) did find positive effects when teachers developed formative assessment practices based on the big idea and five key strategies in the framework by Wiliam and Thompson (2008). In that study, formative assessment practices implemented by the year-4 teachers were based on the same framework as the formative assessment practices implemented by the year-7 teachers in the present study, and the specific practices implemented by both groups of teachers were also very similar. In the following, we will discuss possible reasons that the year-7 teachers' practices were inadequate for improving student achievement, and the characteristics of the implemented formative assessment practices that might explain the differences in the studies' outcomes. Further studies may explore these differences in more detail and empirically examine their possible effects.

5.2. Possible explanations for the non-effects on student achievement

The formative assessment practices of the year-7 teachers were in many ways similar to those of year-4 teachers (students approximately 10 years old) in a parallel study, where we saw that those practices did affect student achievement in mathematics (Andersson and Palm, 2017). The year-7 teachers' formative assessment practices included assessing the learning of all students' learning more often than they did before the PDP (mostly by using exit passes at the end of lessons), which enabled the teachers to adapt their instruction to the learning needs of *all* their students more frequently (and they did use this information to adjust feedback and instructional activities). Such practice is central to formative assessment, and should be beneficial for student achievement.

However, there were some differences between the practices of these two teacher groups that may have affected student achievement differently. For example, all year-4 teachers began to often let all students respond to daily whole-class questions on their mini-whiteboards, and those responses were followed by immediate modifications to instructional activities and feedback (Andersson and Palm, 2017); in contrast, only half of the year-7 teachers did so. Consequently, the year-4 teachers were more able to provide a practice that continuously adapted to their students' learning needs. In addition, the feedback of the year-4 teachers more often included detailed comments about what the students had done well and suggestions for improvement, which would enhance students' feelings of competence and therefore their motivation (Ryan and Deci, 2020), which would provide extended learning opportunities. Another difference between the practices of the two teacher groups is that it seems the year-7 teachers more often used questions more targeted towards 'basic knowledge' than towards conceptual understanding at various levels. The present study did not find any effects on student achievement for the students with the lowest pre-test scores, so the focus on basic skills does not seem to have helped these students learn more. However, more high-achieving students may have been able to benefit from a classroom practice that had been adapted to fit also their learning needs. To achieve significant overall effects on student achievement, it would be important for both questions and adjusted instruction to be targeted to different levels of knowledge and skills, so that the learning needs of all students in the class could be detected in the first place and then met accordingly.

However, assessing different levels of knowledge and skills, and adapting instruction to meet these different learning needs, would be much more difficult than only ensuring that all students obtain basics knowledge and skills in the curriculum. It is possible that these difficulties could be overcome using professionally developed tests as a means of gathering information about the students' progress and understanding. Then the teachers would not have to develop these items themselves, and items on such tests may be more thought-out and aimed towards capturing the true diversity of student understanding. Such items could be used during lessons to complement other ways of collecting information about students' learning needs. Indeed, several studies have shown that the use of such tests can improve student achievement (Yeh, 2009; Burns et al., 2010; Koedinger et al., 2010; Faber et al., 2017; Murphy et al., 2020). A disadvantage of relying on these tests may be that they limit the teachers' flexibility regarding when and how to assess their students, and when and how to take actions based on this assessment information (Palm et al., 2017). Using these kinds of tests as a main source for gathering information about students' learning could therefore hinder development of practices in which teachers could continuously create and capitalise upon what Black and Wiliam (2009, p. 10) call 'moments of contingency', in other words, to make timely adjustments to their teaching in order to meet their students' learning needs.

In their communication with students, the teachers also changed their emphasis away from the number of tasks to be solved in favour of focusing on the intended learning goals, and they started to present learning goals for each lesson. This sort of change may indeed be a first step towards teachers and students reaching a common understanding of the learning goals. However, the teachers did not go into detail with examples or more thorough descriptions of the learning goals, nor the criteria for attaining those goals at various levels, nor did they involve the students in active discussions and negotiations about the meaning of the goals. This change in emphasis towards the learning goals may be too superficial for the students to understand them well enough to use them as motivators, guidance in their own learning and guidance in their support of their peers' learning (Wiliam, 2007). Similarly, although some of the teachers started to hand over more responsibility for correcting diagnostic tests to the students, most of them did not teach the students how to regulate their own learning or how to assess and give feedback to their peers, which are approaches to formative assessment that can improve learning (Ross, 2006; Graham et al., 2015; Palm et al., 2017; Sanchez et al., 2017; Andrade, 2019; Double et al., 2020). Thus, the group of teachers in the present study did

implement practices that included all five key strategies, but most implemented activities were classified as pertaining to Key Strategies 1–3, so the focus was on the teacher as the responsible agent for the formative assessment process. This focus is similar to that identified in other studies aiming for the fourth approach to formative assessment (Jönsson et al., 2015; Wylie and Lyon, 2015). Not providing sufficient support for students to act as proactive agents in the formative assessment processes, either as self-regulated learners or peer assessors, misses one possible factor that could improve student achievement. Indeed, an appropriate use of Key strategies 4 and 5 is sometimes seen as signifying the "spirit" of formative assessment (Marshall and Drummond, 2006; DeLuca et al., 2019).

5.3. Limitations of the study and future research

It cannot be ruled out that the study would have produced different results if, for example, the time between the pre-test and post-test had been even longer, or if the student sample had included students with a longer history of participating in formative assessment practices. Such students might have been able to make more effective uses of the learning opportunities available to them *via* improved feedback based on more frequent assessments. In general, teachers' formative assessment practices might also improve during an implementation. Many teachers may need to start by implementing some parts of a formative assessment practice, and when feeling confident with these improvements, extend their practices further. With time, teachers may also improve the quality of the practice that they implemented. Both of these patterns are consistent with the learning continuum for teachers' implementation of formative assessment suggested by DeLuca et al. (2019), and effects on student achievement may sometimes manifest after a longer period of time than the time span in the present study. This suggests that longitudinal studies on the effects of formative assessment on student achievement would be valuable contributions to the research community.

A limitation of the study is that we did not specifically study the practices of the control group. Since the teachers in the intervention group were randomly selected, the assumption is that both groups of teachers had similar practices before the PDP, and that only the teachers in the intervention group changed their practices after the PDP. If that assumption was not correct (if for example the teachers in the intervention group would have shared their knowledge about formative assessment to their colleagues in the control group, or if the control group teachers had developed formative assessment practices for other reasons), this could have affected the results of the study. However, this scenario seems unlikely, because developing formative assessment practices most often requires strong long-term professional development support with ample time for learning and implementation (Heitink et al., 2016), and there were no other professional development initiatives going on at the participating schools at the time. In addition, the teachers said they had not shared any information with colleagues not teaching students in the intervention group (in fact they said it was difficult to even find time to collaborate with their colleagues who did teach these students). Another possibility is that the quality of mathematics teaching in Swedish schools is particularly high, and differences in achievement from teaching interventions would be easier to detect in countries where teachers do not engage in students' thinking as much. It is certainly possible that changes in teacher practices similar to those that were made in the current study might produce other results in other school contexts, but based on results of international comparative studies such as PISA Sweden does not stand out as an exceptional country when it comes to mathematics achievement (OECD, 2019).

Another limitation of this study is that it did not provide detailed specifications of some of the important characteristics of the teachers' formative classroom practices. Further specifications that would have been useful include more details about the teachers' feedback (e.g., how often each student received feedback and details about the teachers' suggestions for how to improve), the quality of the teachers' questions and tasks and the kinds of knowledge and skills they elicited, and details about the sort of adaptations to instructional activities teachers made in light of the assessment information they collected. Details about the interactions between the teachers and their students (for example, in feedback and support for peer-assessment and self-assessment) would also have been useful. Experimental studies connecting such specified characteristics to student outcomes would be valuable contributions to our understanding of the mechanisms underlying the impact of formative assessment and to the further development of a theory of formative assessment that is based on both theoretical and empirical evidence.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors upon request, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

This article is based on parts of a doctoral dissertation by Boström (2017).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Andersson, C., Boström, E., and Palm, T. (2017). Formative assessment in Swedish mathematics classroom practice. *Nord. Stud. Math. Educ.* 22, 5–20.

Andersson, C., and Palm, T. (2017). The impact of formative assessment on student achievement: a study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learn. Instr.* 49, 92–102. doi: 10.1016/j.learninstruc.2016.12.006

Andrade, H. (2019). A critical review of research on student self-assessment. *Front. Educ.* 4:87. doi: 10.3389/feduc.2019.00087

Baird, J., Hopfenbeck, T., Newton, P., Stobart, G., and Steen-Utheim, A. (2014). *State of the Field review: Assessment and Learning*. Oslo: Report for the Norwegian Knowledge Centre for Education, case number 13/4697. Availalbe at: http://forskningsradet.no

Bell, C., Steinberg, J., Wiliam, D., and Wylie, C. (2008). "Formative assessment and teacher achievement: two years of implementation of the keeping learning on track program" in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education* (New York, NY)

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Princ. Policy Pract.* 18, 5–25. doi: 10.1080/0969594x.2010.513678

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* 5, 7–74. doi: 10.1080/0969595980050102

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educ. Assess. Eval. Account.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Boström, E., (2017) Formativ bedömning: en enkel match eller en svår utmaning. Effekter av en kompetensutvecklingssatsning på lärarnas praktik och på elevernas prestationer i matematik. (Doktorsavhandling, Umeå universitet, Sverige) Hämtad från. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-135038

Boström, E., and Palm, T. (2019). Teachers' formative assesment practices: changes after a professional development programme and important conditions for change. *Assess. Matters* 13, 44–70. doi: 10.18296/am.0038

Boström, E., and Palm, T. (2020). Expectancy-value theory as an explanatory theory for the effect of professional development programmes in formative assessment on teacher practice. *Teach. Dev.* 24, 539–558. doi: 10.1080/13664530.2020.1782975

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., and Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educ. Meas. Issues Pract.* 31, 13–17. doi: 10.1111/j.1745-3992.2012.00251.x

Burns, M., Klingbeil, D., and Ysseldyke, J. (2010). The effects of technology-enhanced formative evaluation on student performance on state accountability math tests. *Psychol. Sch.* 47, 582–591. doi: 10.1002/pits.20492

Chen, F., and Andrade, H. (2018). The impact of criteria-referenced formative assessment on fifth-grade students' theater arts achievement. *J. Educ. Res.* 111, 310–319. doi: 10.1080/00220671.2016.1255870

Chen, F., Lui, A. M., Andrade, H., Valle, C., and Mir, H. (2017). Criteria-referenced formative assessment in the arts. *Educ. Assess. Eval. Account.* 29, 297–314. doi: 10.1007/s11092-017-9259-z

Cohen, L., Manion, L., and Morrison, K. (2011). *Research Methods in Education* (*7th*). New York, NY: Routledge.

DeLuca, C., Chapman-Chin, A., and Klinger, D. (2019). Toward a teacher professional learning continuum in assessment for learning. *Educ. Assess.* 24, 267–285. doi: 10.1080/10627197.2019.1670056

Desimone, L. (2009). Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educ. Res.* 38, 181–199. doi: 10.3102/0013189X08331140

Double, K. S., McGrane, J. A., and Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: a meta-analysis of control group studies. *Educ. Psychol. Rev.* 32, 481–509. doi: 10.1007/s10648-019-09510-3

Dunn, K. E., and Mulvenon, S. W. (2009). A critical review of research on formative assessment: the limited scientific evidence of the impact of formative assessment in education. *Pract. Assess. Res. Eval.* 14, 1–11. doi: 10.7275/JG4H-RB87

Faber, J. M., Luyten, H., and Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: results of a randomized experiment. *Comput. Educ.* 106, 83–96. doi: 10.1016/j.compedu.2016.12.001

Flórez, M. T., and Sammons, P. (2013). *Assessment for Learning: Effects and Impact*. Reading: CfBT Education Trust.

Gorard, S. (2007). The dubious benefits of mulit-level modeling. *Int. J. Res. Method Educ.* 30, 221–236. doi: 10.1080/17437270701383560

Graham, S., Hebert, M., and Harris, K. (2015). Formative assessment and writing. *Elem. Sch. J.* 115, 523–547. doi: 10.1086/681947

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., and Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educ. Res. Rev.* 17, 50–62. doi: 10.1016/j.edurev.2015.12.002

Hirsh, Å., and Lindberg, V. (2015). *Formativ Bedömning på 2000-Talet: En översikt av Svensk Och Internationell Forskning, Delrapport Från Skolforsk-Projektet [Formative Assessment in the 21st Century: An Overview of Swedish and International Research]*. Stockholm: Swedish Research Council.

Huang, F. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *J. Exp. Educ.* 84, 175–196. doi: 10.1080/00220973.2014.952397

Jönsson, A., Lundahl, C., and Holmgren, A. (2015). Evaluating a large-scale implementation of assessment for learning in Sweden. *Assess. Educ. Princ. Policy Pract.* 22, 104–121. doi: 10.1080/0969594X.2014.970612

Kane, M. (2016). Explicating validity. *Assess. Educ. Princ. Policy Pract.* 23, 198–211. doi: 10.1080/0969594X.2015.1060192

Kingston, N., and Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educ. Meas. Issues Pract.* 30, 28–37. doi: 10.1111/j.1745-3992.2011.00220.x

Koedinger, K., McLaughlin, E., and Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *J. Educ. Comput. Res.* 43, 489–510. doi: 10.2190/EC.43.4.d

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., and Cooper, H. (2021). A meta-analysis on the impact of grades and comments on academic motivation and achievement: a case for written feedback. *Educ. Psychol.* 41, 922–947. doi: 10.1080/01443410.2019.1659939

Lam, A., Ruzek, E., Schenke, K., Conley, A., and Karabenick, S. (2015). Student perceptions of classroom achievement goal structure: is it appropriate to aggregate? *J. Educ. Psychol.* 107, 1102–1115. doi: 10.1037/edu0000028

Maas, C., and Hox, J. (2004). Robustness issues in multilevel regression analysis. *Stat. Neerl.* 58, 127–137. doi: 10.1046/j.0039-0402.2003.00252.x

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educ. Psychol.* 47, 106–124. doi: 10.1080/00461520.2012.670488

Marshall, B., and Drummond, M. J. (2006). How teachers engage with assessment for learning: lessons from the classroom. *Res. Pap. Educ.* 21, 133–149. doi: 10.1080/02671520600615638

McMillan, J. H., Venable, J. C., and Varier, D. (2013). Studies of the effect of formative assessment on student achievement: so much more is needed. *Pract. Assess. Res. Eval.* 18, 1–15. doi: 10.7275/tmwm-7792

Murphy, R., Roschelle, J., Feng, M., and Mason, C. A. (2020). Investigating efficacy, moderators and mediators for an online mathematics homework intervention. *J. Res. Educ. Effect.* 13, 235–270. doi: 10.1080/19345747.2019.1710885

National Mathematics Advisory Panel. (2008). Chapter 6: Report of the Task Group on Instructional Practices. Available at: http://www.ed.gov/about/bdscomm/list/mathpanel/report/instructional-practices.pdf (Accessed September 12, 2014).

OECD (2019). "PISA 2018 Results" in *(Volume 1): What Students Know and Can Do* (Paris, France: PISA, OECD Publishing)

Palm, T., Andersson, C., Boström, E., and Vingsle, L. (2017). A review of the impact of formative assessment on student achievement in mathematics. *Nord. Stud. Math. Educ.* 22, 25–50.

Randel, B., Apthorp, H., Beesley, A., Clark, T., and Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *J. Educ. Res.* 109, 491–502. doi: 10.1080/00220671.2014.992581

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., and Miller, T. R. (2003). Peer-assisted learning interventions with elementary school studies: a meta-analytic review. *J. Educ. Psychol.* 95, 240–257. doi: 10.1037/0022-0663.95.2.240

Ross, J. (2006). The reliability, validity, and utility of self-assessment. *Pract. Assess. Res. Eval.* 11, 1–13. doi: 10.7275/9wph-vv65

Ryan, R., and Deci, E. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: definitions, theory, practices, and future directions. *Contemp. Educ. Psychol.* 61:101860:101860. doi: 10.1016/j.cedpsych.2020.101860

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., and Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: a meta-analysis. *J. Educ. Psychol.* 109, 1049–1066. doi: 10.1037/edu0000190

Schneider, C., and Randel, B. (2010). "Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment" in *Handbook of Formative Assessment*. eds. H. Andrade and G. Cizek (New York: Routledge), 251–276.

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

Skolverket. (2012). *TIMSS 2011: Svenska Grundskoleelevers Kunskaper i Matematik Och Naturvetenskap i ett Internationellt Perspektiv (Rapport 380)*. Stockholm: Skolverket.

Swaffield, S. (2011). Getting to the heart of authentic assessment for learning. *Assess. Educ. Princ. Policy Pract.* 18, 433–449. doi: 10.1080/0969594X.2011.582838

Timperley, H., Wilson, A., Barrar, H., and Fung, I. (2007). *Teacher Professional Learning and Development: Best Evidence Synthesis Iteration*. Wellington, New Zealand: Ministry of Education.

Wafubwa, R. N. (2020). Role of formative assessment in improving students' motivation, engagement, and achievement: a systematic review of literature. *Int. J. Assess. Eval.* 28, 17–31. doi: 10.18848/2327-7920/CGP/v28i01/17-31

Wafubwa, R. N., and Csíkos, C. (2022). Impact of formative assessment instructional approach on students' mathematics achievement and their metacognitive awareness. *Int. J. Instr.* 15, 119–138. doi: 10.29333/iji.2022.1527a

Wiliam, D. (2007). "Keeping learning on track: classroom assessment and the regulation of learning" in *Second Handbook of Mathematics Teaching and Learning*. ed. LesterF. K. Jr. (Greenwich, CT: Information Age Publishing), 1053–1098.

Wiliam, D. (2011). *Embedded Formative Assessment*. Bloomington, Indiana: Solution Tree Press.

Wiliam, D., Lee, C., Harrison, C., and Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assess. Educ. Princ. Policy Pract.* 11, 49–65. doi: 10.1080/0969594042000208994

Wiliam, D., and Thompson, M. (2008). "Integrating assessment with learning: what will it take to make it work?" in *The Future of Assessment: Shaping Teaching and Learning*. ed. C. A. Dwyer (Mahwah, NJ: Lawrence Erlbaum Associates), 53–82.

Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: a meta-analysis of educational feedback research. *Front. Psychol.* 10:3087. doi: 10.3380/fpsyg.2019.03087

Wylie, E. C., and Lyon, C. J. (2015). The fidelity of formative assessment implementation: issues of breadth and quality. *Assess. Educ. Princ. Policy Pract.* 22, 140–160. doi: 10.1080/0969594X.2014.990416

Yeh, S. S. (2009). Class size reduction or rapid formative assessment? A comparison of cost-effectiveness. *Educ. Res. Rev.* 4, 7–15. doi: 10.1016/j.edurev.2008.09.001