

University of Groningen

Testing the effectiveness of classroom formative assessment in Dutch primary mathematics education

van den Berg, Marian; Bosker, Roel J.; Suhre, Cor J. M.

Published in:
School Effectiveness and School Improvement

DOI:
[10.1080/09243453.2017.1406376](https://doi.org/10.1080/09243453.2017.1406376)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

van den Berg, M., Bosker, R. J., & Suhre, C. J. M. (2018). Testing the effectiveness of classroom formative assessment in Dutch primary mathematics education. *School Effectiveness and School Improvement*, 29(3), 339-361. <https://doi.org/10.1080/09243453.2017.1406376>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Testing the effectiveness of classroom formative assessment in Dutch primary mathematics education

Marian van den Berg, Roel J. Bosker & Cor J.M. Suhre

To cite this article: Marian van den Berg, Roel J. Bosker & Cor J.M. Suhre (2018) Testing the effectiveness of classroom formative assessment in Dutch primary mathematics education, *School Effectiveness and School Improvement*, 29:3, 339-361, DOI: [10.1080/09243453.2017.1406376](https://doi.org/10.1080/09243453.2017.1406376)

To link to this article: <https://doi.org/10.1080/09243453.2017.1406376>



© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 06 Dec 2017.



Submit your article to this journal [↗](#)



Article views: 1389





View Crossmark data [↗](#)

ARTICLE



Testing the effectiveness of classroom formative assessment in Dutch primary mathematics education

Marian van den Berg ^{a,b}, Roel J. Bosker ^a and Cor J.M. Suhre ^c

^aGION education/research, University of Groningen, Groningen, The Netherlands; ^bCedin, Drachten, The Netherlands; ^cDepartment of Teacher Education, University of Groningen, Groningen, The Netherlands

ABSTRACT

Classroom formative assessment (CFA) is considered to be a fundamental part of effective teaching, as it is presumed to enhance student performance. However, there is only limited empirical evidence to support this notion. In this effect study, a quasi-experiment was conducted to compare 2 conditions. In the treatment condition, 17 teachers implemented a CFA model containing both daily and weekly goal-directed instruction, assessment, and immediate instructional feedback for students who needed additional support. In the control condition, 17 teachers implemented a modification to their usual practice. They assessed their students' mastery of learning goals on the basis of half-yearly mathematics tests, and prepared weekly pre-teaching sessions for groups of low-achieving students. The posttests showed no significant differences in student performance between the 2 conditions after controlling for student and teacher characteristics. The degree of implementation of the CFA model, however, appeared to be positively related to the 5th-grade students' performance.

ARTICLE HISTORY

Received 8 December 2016
Accepted 13 November 2017

KEYWORDS

Classroom formative assessment; instructional effectiveness; primary education; mathematics

Introduction

Basic mathematical knowledge and skills are prerequisites to fully participate in today's society (Organisation for Economic Co-operation and Development [OECD], 2014). Unfortunately, in many countries primary school students' mathematics performance is below expectation or is declining (OECD, 2014, pp. 50–56, 2016, pp. 181–184). To improve students' mathematical abilities, governments, researchers, and teachers often turn their attention to instructional practices that include the use of elements such as goal setting, assessment, adaptive teaching, grouping strategies, feedback, and reinforcement, as they are generally considered to be effective in enhancing student performance (Good, Wiley, & Flores, 2009; Reynolds et al., 2014; Scheerens, 2016). Particularly the use of formative assessment as a combination of three of these elements, namely, goal setting, assessment, and feedback, has gained renewed interest (Conderman & Hedin, 2012; Inspectie van het Onderwijs [Dutch Inspectorate of Education], 2010; Mandinach, 2012). Formative assessment refers to the process of gathering and analyzing information about the students' understanding of a learning goal to provide

CONTACT Marian van den Berg  m.van.den.berg@rug.nl

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

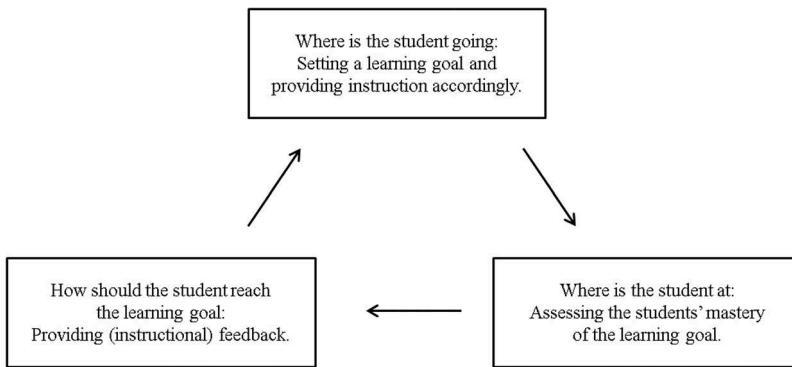


Figure 1. Three elements of formative assessment.

instructional feedback that moves students forward in their learning process (Black & Wiliam, 2009; Callingham, 2008; Shepard, 2008). It is thus a process (see Figure 1) consisting of three key elements, “goal setting for instruction”, “assessment”, and “instructional feedback”, that is used to promote adaptive teaching (Black & Wiliam, 2009; Wiliam & Thompson, 2008).

A formative assessment cycle is started by setting clear learning goals for the students. Formulating clear learning goals is a precondition for the subsequent assessment to take place, as it determines which knowledge and skills are going to be taught and assessed (Ashford & De Stobbeleir, 2013; Locke & Latham, 2006; Marzano, 2006). Furthermore, it enables drawing conclusions about students’ levels of mastery (Moon, 2005).

The subsequent assessment is used to gather information about the possible gaps in students’ current knowledge and skills and those described by the learning goals. The assessment should also provide information about a student’s zone of proximal development. This zone of proximal development is best described as the level of competence that a student can reach with the help of a more competent other (Vygotsky, 1978). This kind of information is crucial for providing effective instructional feedback aimed at closing students’ gaps in knowledge and skills (Hattie & Timperley, 2007; Moon, 2005; Shepard, 2005). A teacher can provide instructional feedback by means of scaffolding within the zone of proximal development. This entails that a teacher supports a student in completing a task by, for instance, modeling it, providing explicit instruction, or providing visual representations (Pfister, Moser Opitz, & Pauli, 2015). This support is gradually released up until the point that the student can perform the task on his or her own (Wood, Bruner, & Ross, 1976). Once the instructional feedback has been provided and all students have mastered the learning goal, the formative assessment cycle starts all over again. On the basis of the place, timing, and purpose of the above-mentioned elements, several types of formative assessment can be distinguished, ranging from the use of standardized test data aimed at differentiated instruction to self-assessments to enhance students’ self-regulated learning.

Although there appears to be a consensus on the value added of formative assessment in promoting student performance, there is much debate about which form of formative assessment is most effective. There appears to be no clear empirical evidence about what works in which way for students of different ages (cf. Dunn &

Mulvenon, 2009; Kingston & Nash, 2011; McMillan, Venable, & Varier, 2013). There is especially little known about the effectiveness of formative assessment in mathematics education. A meta-analysis by Kingston and Nash (2011) about the effect of formative assessment on student performance has reported on only five studies regarding mathematics education. These studies yielded a mean effect size of .17 with 95% confidence intervals ranging from .14 to .20 ($n = 19$). This is considered to be a small effect size (Cohen, 1988).

There is, however, reason to believe that formative assessment consisting of frequent assessments in combination with timely instructional feedback is effective in enhancing student performance. In fact, a minimal frequency of two times a week has been reported to yield effect sizes no smaller than between .80 and .85 and percentile point gains of 29.0 to 30.0 (Bangert-Drowns, Kulik, & Kulik, 1991; Fuchs & Fuchs, 1986). This may explain why a common formative assessment practice in which half-yearly standardized tests are used to create differentiated instruction plans often has a nonsignificant or small effect on student performance (cf. Keuning & Van Geel, 2016; Ritzema, 2015): The time span between the assessment and the instructional feedback may be too long (Conderman & Hedin, 2012).

Rationally, it makes sense to suppose that frequent assessments are effective in improving student achievement, considering that it entails timely and continuous feedback. Perhaps, therefore, many teachers use more ongoing assessments to gain additional information about students' understanding for the purpose of instructional decision making. This kind of formative assessment, called classroom formative assessment (CFA), is regarded as a promising means to enhance student performance, as it is used during a lesson cycle (Conderman & Hedin, 2012). CFA requires the teacher to assess the students' mastery of a particular learning goal during the lesson and provide immediate instructional feedback. By providing instructional feedback during the learning activity, students' misconceptions are corrected as quickly as possible allowing for an uninterrupted learning process. CFA should be particularly effective in enhancing student performance in the domain of mathematics, as mastery of new topics hinges on previously acquired mathematical knowledge and skills. Therefore, it seems reasonable to assume that a teacher should continuously assess the students' mastery and provide instructional feedback to prevent students from developing knowledge gaps.

Often, CFA consists of an interaction between the teacher and students to allow for decision making during instruction (cf. Heritage, 2010; Leahy, Lyon, Thompson, & Wiliam, 2005; Shepard, 2000). Assessment techniques such as questioning (preferably with the aid of answering cards) and classroom discussions are used to provide the teacher with an impression of the *class's* understanding of the learning goal. It is debatable, however, whether these interactive assessment techniques provide teachers with sufficient insight into students' individual difficulties. For example, not all students may participate actively in the discussion, resulting in a lack of insight into these students' skills and proficiencies. Furthermore, in general these interactive techniques tend to generate an unstructured overload of information that is difficult to translate into instructional feedback for each individual student. Without specific information about students' individual difficulties, the provision of adequate instructional feedback is problematic (Ateh, 2015). It thus seems that teachers should *frequently* apply classroom formative assessments during the lessons that allow them to gain insight into each *individual* student's struggles.

Although the use of CFA can be considered to be an essential professional skill for teachers to master (Assessment Reform Group, 2002), teachers often experience problems in implementing it. Teachers find it particularly difficult to use the three aspects – goal setting, assessment, and instructional feedback – in a coherent manner. For instance, teachers tend to assess their students' understanding without setting clear goals and criteria for success (Antoniou & James, 2014) or do not provide adequate feedback based on the information gathered during the assessment (Furtak et al., 2008; Wylie & Lyon, 2015). It thus seems that in order for teachers to effectively implement CFA, they should be provided with adequate training and coaching on the job.

In this article, we assess the potential value added of a CFA practice by comparing student performance on mathematics tests in two conditions. In the treatment condition (CFA condition), 17 teachers from seven schools used a CFA model in which frequent assessments of each student's mastery were applied to allow for specific instructional feedback during the mathematics lessons, whilst in the control condition 17 teachers from eight different schools implemented a modification to their usual practice of analyzing their students' half-yearly standardized mathematics tests to plan instructional support for low-achieving students. To diminish implementation issues, the teachers were intensively trained and coached during a professional development program (PDP). Both conditions and their differences will be discussed in more detail in the study design section. The questions we seek to answer in this article are:

- (1) To what extent do teachers in a CFA condition use goal-directed instruction, assessment, and immediate instructional feedback more frequently than teachers in a control condition?
- (2) To what extent is the teachers' use of the CFA model effective in enhancing students' mathematics performance?
- (3) To what extent does a higher degree of implementation of the CFA model lead to higher student performance?

We expected that, as a result of our PDP, teachers in the CFA condition would use goal-directed instruction, assessment, and immediate instructional feedback more frequently than the teachers in the control setting. Given the *frequency* of the assessments and immediate instructional feedback for *all* students who need it, we expected that the students in the CFA condition would outperform the students in the control condition on mathematics tests. Additionally, we expected that there would be a positive relationship between the degree of implementation and student performance. For the latter two research questions, we also investigated whether the CFA model was more or less effective for students with different mathematics abilities.

Study design

We conducted a quasi-experimental pretest-posttest design in which 34 teachers participated in either the CFA condition or the control condition. In the following paragraphs, we will first explain how our participants were selected, and specify their characteristics. Next, we will describe the outline of our two conditions. Finally, we will present the instruments and data analysis methods that were used.

Selection and participants

In the school year preceding the study, 24 schools were randomly assigned to the CFA condition and the control condition. These schools were accommodating single-grade classes in Grades 4 and 5, and worked with one of two sets of mathematics curriculum materials (named “Plus” and “The World in Numbers”). The curriculum materials consisted of, amongst others, learning trajectories, tests, guidelines for instruction, and differentiated assignments. After the schools were randomly assigned to either the CFA or the control condition, they were informed about the study and asked if their fourth- and fifth-grade teachers were willing to take part. During a meeting, a researcher (four researchers participated in the study) provided the school leader and interested teachers with information about the project. Together, they tried to establish coherence in the goals of the project and those of the school and discussed what was expected of the teachers during the project. Once both issues were agreed upon, the teachers could take part in the project.

This resulted in 13 teachers from five schools willing to take part in the CFA condition and six teachers from three different schools willing to participate in the control condition. As we aimed for a minimum of 16 teachers per condition, we repeated the procedure with another 10 schools, resulting in four additional teachers (two schools) taking part in the CFA condition and six more teachers (three schools) in the control condition. Since we had enough teachers for the CFA condition but fell short of four teachers in the control condition, we contacted four more schools to ask if their teachers wanted to take part in our study as well. This resulted in the participation of five more teachers (two schools) in the control condition.

Our final sample thus consisted of 34 teachers from 15 schools. Nine fourth-grade and eight fifth-grade classes from seven schools participated in the CFA condition. The teachers of these classes implemented the CFA model in their mathematics lessons. Eight of them were male and nine female. Another nine fourth-grade classes and eight fifth-grade classes from eight different schools functioned as the control group. As in the CFA condition, eight of their teachers were male and nine female. The teachers in the CFA condition were on average 46.94 years old ($SD = 9.19$) and had 21.06 years of teaching experience ($SD = 9.91$). In the control condition, the teachers were on average 44.76 years old ($SD = 14.06$) and had 21.76 years of teaching experience ($SD = 14.37$). The groups of teachers did not differ significantly from each other with regard to these characteristics (age: $t(32) = -.534, p = .597$; teaching experience: $t(32) = .167, p = .869$). Our sample consisted of proportionally more male teachers than there are in the general population of Dutch primary school teachers. As there is no empirical evidence that gender plays a role in teachers’ use of formative assessment practices, we did not consider this to be an issue. The average age of the teachers in our sample was sufficiently comparable to that of the population of Dutch primary school teachers (Ministerie van Onderwijs, Cultuur en Wetenschap [Dutch Ministry of Education, Culture and Science], 2014).

The effectiveness of the CFA model was tested at the student level. In total, 873 students participated in the study. Due to illness, 38 of them failed to take the pretest (4% of the sample), while another 49 students (6% of the sample) did not take the posttest, adding up to 10% of missing data ($n = 87$). Of the 87 students whose test results were not included in the data analyses, 53% was in the CFA condition. Our final

sample contained 786 students, of which 381 were in the CFA condition, consisting of 53.4% boys. In the control condition, there were 405 students, of which 55.2% was a boy. The difference in gender between the two groups was not significant with $\chi^2(1) = .42$, $p = .52$. The results of the 381 students in the CFA condition were used to determine the relationship between the degree of CFA implementation and student performance in mathematics.

Comparing two conditions

A common method of establishing the effectiveness of an intervention is comparing a business-as-usual control condition with a treatment condition. This practice, however, has some drawbacks. First, it is susceptible to the so-called Hawthorne effect. This kind of effect entails that teachers and/or students in the treatment condition may behave differently than those in the control group because of the specific treatment they are given. This change in behavior may lead to an effect of the intervention that is thus biased (Shadish, Cook, & Campbell, 2002). Second, when using a business-as-usual control condition, we would have no real knowledge of or influence on what happened in those schools. Some teachers, for instance, could already be using (elements of) the CFA model, which would influence the results of the study. As many schools are hesitant to take part in a study and engage in observations, testing, and interviews without receiving anything in return, we provided a low-impact intervention to be able to observe the teachers in class. Thus, in order to diminish the chance of a Hawthorne effect, and to have some insight into what teachers in the control schools were doing during the mathematics lessons, we created interventions for both the treatment (CFA) condition and the control condition.

Classroom formative assessment condition

In the CFA condition, the teachers implemented a model that was based on frequent use of the formative assessment cycle described in the introduction. To support the teachers during the implementation process, the model was embedded in two commonly used sets of mathematics curriculum materials. The teachers could use the materials to identify the learning goals, consult guidelines for instruction (e.g., explanation of mathematical procedures, such as ciphering, or suggestions for mathematical representations, such as number lines or an abacus), and draw from readily available assignments to assess the students' mastery of the goals. To facilitate frequent assessments followed by immediate instructional feedback, the CFA model consisted of four daily CFA cycles and a weekly cycle. [Figure 2](#) illustrates this procedure.

On a daily basis, the teachers were expected to decide upon one learning goal that was going to be taught and assessed at a class level. They had to provide a short whole-group instruction that was focused on this goal and use an appropriate mathematical procedure or representation. Hereafter, the teacher assessed each student's mastery of the learning goal by giving him or her a specific task related to the learning goal, such as adding up to a hundred by means of jumping on a number line. The teachers observed the students while they were working on this task. This approach is an efficient way of assessing individual students in a class of approximately 25 students (Ginsburg, 2009). To allow the teachers to gain more insight into the students' issues and provide more

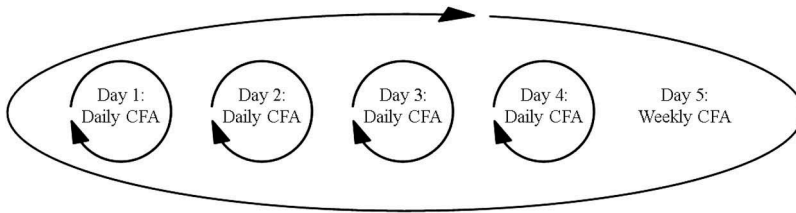


Figure 2. The CFA model consisting of four daily CFA cycles and one weekly cycle.

effective immediate instructional feedback, the students were expected to write down their mathematical procedures or use mathematical representations when making the assignments (Heritage & Niemi, 2006; Kouba & Franklin, 1995). On the basis of these assessments, the teachers selected students who showed an insufficient understanding of the task. These students received immediate instructional feedback, which took place in the form of small-group instruction. This setting allowed the teachers to scaffold the students' learning, for example, by addressing the prior knowledge required to master the learning goal, or introducing other mathematical representations or concrete materials. As the selection of the students was dependent on the assessment during the lesson, the students who needed instructional feedback could *vary per lesson*.

At the end of the week, the teachers assessed the students' understanding of the learning goals once more by means of a quiz on the digital whiteboard. Each quiz consisted of eight multiple-choice questions based on the four learning goals that had been covered during the weekly program. The questions were all developed by the researchers and presented in the same format as the tasks in the students' textbooks. The three incorrect answer possibilities were based on common mathematical errors, such as procedural errors ($34 + 46 = 70$) or place value errors ($34 + 46 = 98$) (Kraemer, 2009; Ryan & Williams, 2007). The multiple-choice questions enabled the teachers to detect the students' misconceptions that they had not identified and corrected during the daily CFA cycles (Ginsburg, 2009). To enhance the participation in the quiz, the students had to answer the questions by means of a clicker (voting device). Their answers were then digitally stored by a classroom response system (Lantz, 2010). After each question, the teacher and students could see the class's performance on a response frequency chart. On the basis of this information, the teacher could choose to provide immediate instructional feedback by making use of an animation (e.g., jumping on a number line) below the question. This animation showed a step-by-step solution to the problem that was posed in the multiple-choice question. Prior to the study described in this article, six quizzes were piloted in three different schools to check whether the quizzes were user friendly, unambiguous, and at an appropriate level of difficulty. An example of a multiple-choice question including instructional feedback is depicted in Figure 3.

At the end of the quiz, the teachers were expected to analyze the students' recorded scores to decide which students needed some more instructional feedback.

To ensure an optimal implementation of the CFA model, the teachers took part in a PDP led by a certified educational coach. Three certified educational coaches from an external consultancy bureau participated in the study. The PDP started with a *small-*

1 Do the math

€ 868,05 – € 293,68 =

A. € 635,63
B. € 574,47
C. € 575,37
D. € 574,37

€ 8 6 8 , 0 5
€ 2 9 3 , 6 8 -
€ _____ ?

5 - 8 =

Multiple choice question

Instructional feedback

Figure 3. An example of a multiple-choice question and instructional feedback from a fifth-grade quiz.

group workshop in which the teachers were made aware of the *coherence* between the goals of the innovation and those of the school (Desimone, 2009) by discussing, for instance, that the teachers should assess the students' mastery of the learning goal themselves, instead of relying solely on their students to come to them with questions. It was important to make the teachers feel that the rationale behind the innovation was in line with their own teaching practices, as this would increase the chances that the teachers would implement the innovation on a continuous basis (Lumpe, Haney, & Czerniak, 2000). The workshop was also used to discuss possible *barriers* (e.g., school schedules, the quality of the curriculum materials, and the time available for planning and reflection) and the *support* (e.g., whether the teachers had all the necessary mathematical materials, such as coins and bills or fraction cubes, available in their classroom to provide small-group instruction) required to optimize the implementation process (Penuel, Fishman, Yamaguchi, & Gallagher, 2007). Hereafter, the workshop focused on realising *active learning* with a focus on *content* (Garet, Porter, Desimone, Birman, & Yoon, 2001; Heller, Daehler, Wong, Shinohara, & Miratrix, 2012) by having the teachers first watch a video featuring a best-practice example of the CFA model. Then, the teachers prepared a few lessons according to the CFA model and received feedback from each other and the coach. The teachers and coach also discussed which mathematical representations and procedures could best be used for instruction of particular learning goals. To support the teachers, they were provided with an example of a lesson plan, an overview of the mathematical learning trajectories for their year grade including mathematical representations and procedures to be used during (small-group) instruction, and a manual for the classroom response system. Finally, the teachers were expected to practice with the classroom response system after a demonstration.

The workshop was followed up with *on-site* professional development. There is evidence that a focus on how to use the innovation in one's own practice is effective in accomplishing teacher change (Darling-Hammond, 1997; Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009). The on-site practice was combined with *coaching on the job*. This allowed the teachers to receive timely individualized feedback on their use of the model, which should

help teachers change their teaching behavior (Birman, Desimone, Porter, & Garet, 2000; Grierson & Woloshyn, 2013). The coaching on the job included both an observation and a reflective conversation between the teacher and the coach, in which the teacher would self-evaluate his or her lesson and compare these findings to those of the coach. Throughout the PDP, the teachers were stimulated to evaluate their use of the CFA model together, as *collective participation* motivates teachers to discuss and solve practical problems collectively (Little, 1993; Porter, Garet, Desimone, Yoon, & Birman, 2000) and enhances the sustainability of an educational innovation (Coburn, 2004; Fullan, 2007; Moss, Brookhart, & Long, 2013). For instance, in addition to the initial workshop a team meeting was organized half-way through the intervention in which the teachers and coach evaluated the use of the CFA model and discussed specific difficulties collectively. The PDP lasted a full school year. This is considered to be sufficient for teachers to practice with an innovation (Birman et al., 2000).

Control condition

In the control condition, the teachers implemented a model that was based on a currently favored practice in The Netherlands, in which teachers use half-yearly standardized tests to monitor the students' progress. On the basis of the test results, a number of ability groups are then formed within class, of which often the high-achieving students are allowed to skip the whole-group instruction while the low-achieving students always receive extra small-group instruction after the whole-group instruction regardless of the learning goal that is discussed (Inspectie van het Onderwijs, 2010). The teachers in the control condition also used the half-yearly mathematics test results to assess the low-achieving students' mastery of the goals, which is comparable to the common practice mentioned above. Contrary to the currently favored practice in The Netherlands, this information was not used for providing extra small-group instruction but to prepare weekly pre-teaching sessions for these students. This procedure is represented in Figure 4.

The teachers entered the low-achieving students' responses to the standardized test in an Excel macro. The macro identified these students' specific mathematical problem domains (e.g., adding, subtracting, or metrics). For instance, when a student would answer only 3 out of 10 questions within the domain of metrics correctly, the macro would highlight this domain as a problem area. The macro also provided the teachers with pre-teaching plans for the low-achieving students in need of instructional support for a particular learning goal within a problem domain during the upcoming semester.

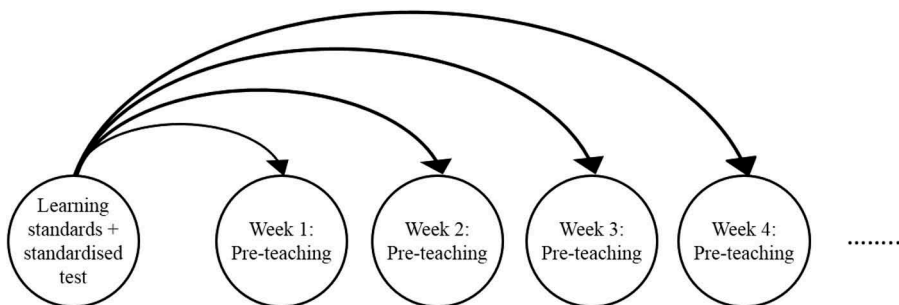


Figure 4. The model used in the control condition consisting of half-yearly standardized test results as input for weekly pre-teaching sessions for low-achieving students during a semester.

The pre-teaching plans contained the student's problem domain, the learning goal related to this domain, and the related curriculum materials that the teachers should use during the pre-teaching sessions.

The teachers in the control condition taught their daily lessons as described in the curriculum materials. At the end of the week, the selected low-achieving students received pre-teaching on the upcoming lessons. Pre-teaching is considered just as effective in enhancing the mathematics performance of low-achieving students as small-group instruction after whole-group instruction (Carnine, 1980; Lalley & Miller, 2006), which is a common practice in The Netherlands (Inspectie van het Onderwijs, 2010). Compared to the CFA condition, the control condition can thus be viewed as a low-impact intervention.

The teachers in the control condition also took part in a PDP. However, this PDP was less intensive compared to the PDP in the CFA condition. The teachers in the control condition participated in three small-group workshops led by a certified educational coach, in which they discussed the coherence between the goals of the innovation and those of the school, possible barriers to overcome, and support needed to allow for a successful implementation process. During the workshops, the teachers learned how to use the Excel macro to interpret the standard-based test results of their low-achieving students and to formulate pre-teaching plans. On the basis of these plans, the teachers prepared the pre-teaching sessions for the upcoming week by determining which mathematical representations or procedures could best be used. The teachers received the Excel macro including a manual, an example of a lesson plan, and an overview of the mathematical learning trajectories for their year grade including mathematical representations and procedures to be used during the pre-teaching sessions. Comparable to the PDP in the treatment condition, the PDP of the control condition was spread over a period of a full school year. Table 1 shows both conditions together for comparison purposes.

Procedure

At the beginning of the school year, the students in all classes took a paper-and-pencil mathematics pretest on the learning goals covered the school year before. The researchers administered the test following a fixed protocol. In addition, they observed the teachers' usual mathematics teaching practices to determine whether the teachers were already using elements of the CFA model and to establish any initial differences between the teachers' teaching practices in both conditions. After these observations, the professional development programs started as described in the sections on the two conditions. First, the teachers in both conditions attended a workshop. During the third and fourth weeks, those in the CFA condition were coached on the job four times, twice

Table 1. Overview of the CFA condition and the control condition.

	CFA condition	Control condition
Teaching practice	Daily and weekly goals Daily and weekly assessments Daily small-group instruction to varying groups of students	Learning standards Half-yearly (and monthly) assessments Weekly pre-teaching sessions to preset groups of low-achieving students
Professional development program	One workshop Coaching on the job Team meeting	Three workshops

during a daily mathematics lesson and twice during a weekly quiz. After these 2 weeks, the teachers were expected to carry out the CFA model by themselves.

Halfway through the school year, the researchers observed the teachers in both conditions again. In addition, the CFA teachers were coached on the job during one lesson after which they participated in a school team meeting. The teachers in the control condition followed a similar workshop to the first where they analyzed the half-yearly test results and prepared a pre-teaching plan.

For the remainder of the school year, the teachers in the CFA condition were coached on the job for a minimum of two lessons. The teachers in the control condition attended a final workshop in which they evaluated the use of the half-yearly tests and the pre-teaching sessions. In addition, they analyzed the half-yearly tests to make a pre-teaching plan for the upcoming school year. Finally, at the end of the school year, the researchers observed the teachers in both conditions during a mathematics lesson. Then, the students took a paper-and-pencil mathematics posttest covering the learning goals taught during the project. Again, the test was administered according to a testing protocol.

Instruments

Observation instrument

To determine to what extent the teachers used goal-directed instruction, assessment, and immediate instructional feedback on a daily basis, three mathematics lessons per teacher were observed. During each of these lessons, the researchers scored the teacher's activities (instruction, assessment, feedback, or classroom management) per minute on a time interval scoring sheet. Observation data about goal-directed instruction, assessment, and immediate instructional feedback were used to construct implementation scores of these elements. The scores were based on the following key features which could be observed during the lesson:

Goal-directed instruction:

- Feature 1: The teacher provides a short introduction.
- Feature 2: The teacher provides an instruction of a maximum of 20 min that is focused on one learning goal.
- Feature 3: The teacher uses a mathematical representation or procedure that is in accordance with the learning goal.

Assessment:

- Feature 1: The teacher assesses the students' work during seat work for 2 (class size: 15 to 20 students) to 6 (class size: more than 20 students) min before providing immediate instructional feedback.
- Feature 2: The teacher's primary focus lies on assessing the students' work rather than on responding to the students' questions.

Immediate instructional feedback:

- Feature 1: The teacher provides the selected students with instructional feedback immediately after the assessment.

- Feature 2: The teacher uses a mathematical representation that is in accordance with the learning goal.
- Feature 3: The teacher assesses the selected students' mastery of the learning goal after the immediate instructional feedback.
- Feature 4: The teacher spends at least 5 min on providing immediate instructional feedback about the learning goal and re-assessing the student's mastery (5 min was considered to be the minimum amount of time to perform these actions).

As the number of features that were observed for the three elements of the CFA model differed, the scores on each scale were transformed into proportions. For instance, as regards goal-directed instruction, if a teacher started the lesson with a short introduction (Feature 1) and provided instruction about one learning goal for a maximum of 20 min (Feature 2), but did not use an appropriate mathematical representation or procedure (Feature 3), then this teacher would score two out of three points. This score was transformed into a proportion of .67. As a result of the transformation of the scores into proportions, all scales ran from 0 to 1. The reliability of the observation instrument was high with a Cronbach's α of .85. The teachers were observed by four different researchers. The inter-observer reliability for multiple raters among the researchers was good with $\kappa = .759$ and $p < .001$ (Siegel & Castellan, 1988).

Registration of quiz data

The quiz data stored by the classroom response system were used to check how often the teachers in the CFA condition administered the weekly quiz and analyzed the students' quiz results. The teachers gave mathematics lessons for 21 weeks, while the remaining weeks were "test weeks" within the curriculum. The classroom response system therefore had to store the results of 21 quizzes. The data were used to determine the proportion of quizzes administered and analyzed by the teachers.

Implementation scale complete CFA model

The scores on the teachers' daily use of goal-directed instruction, assessment, and immediate instructional feedback were transformed to proportions. As a result, we were able to combine these proportions regarding the second and third observations and the proportions pertaining to the use of the weekly quizzes and report. This resulted in one implementation scale for the complete CFA model. The scale ran from 0 to 1, as the use of CFA during the lessons and the weekly quizzes were all based on proportions. With a Cronbach's α of 0.83, its reliability was good.

Mathematics tests

The students took two mathematics tests: a pretest at the beginning of the project and a posttest at the end. As we wished to determine whether the teachers' use of the CFA model was effective in enhancing the students' understanding of the learning goals, we developed new mathematics tests that primarily focused on the topics that were taught in both conditions during the project. The newly developed tests also prevented teaching to the test, as the items – contrary to the items of the standardized tests – would be unknown to both the teachers and the students. To construct our tests, we analyzed what kind of domains (e.g., multiplying and dividing) and subskills (e.g., multiplications

with a one-digit and a three-digit number) were covered in both sets of curriculum materials. The domains that were found to be present in both sets were:

- numeracy;
- adding and subtracting;
- multiplying and dividing;
- money (not in the fifth-grade posttest);
- time;
- measurement and geometry;
- fractions, ratio (both not in the fourth-grade pretest), and percentages (only in the fifth-grade posttest)/

We used a matrix in which we crossed these domains with the subskills to ensure content validity. The number of questions about a domain within a test was based on the number of times the domain was taught in both sets of curriculum materials. All developed tests consisted of open-ended and multiple-choice questions comparable to the tasks in the students' textbooks. Figure 5 depicts two example questions of the fifth-grade posttest.

The psychometric qualities of all tests were examined by calculating p values, corrected item-total correlations, and Cronbach's alpha values. Table 2 shows that the internal consistency of all tests was high, with Cronbach's alphas ranging from .81 to .84. The tests appeared to be quite difficult, particularly the fifth-grade posttest with a mean difficulty of $p = .37$ ($SD = .15$). Nonetheless, the tests discriminated well between students with high and low mathematics ability levels, with corrected item-total correlations of between .13 and .56. One item in the fifth-grade pretest with a corrected item-total correlation below .10 was deleted, as such items discriminate poorly between students' abilities (cf. Nunnally & Bernstein, 1994). This resulted in a pretest containing 26 items with a high internal consistency of Cronbach's $\alpha = .84$. We wished to use the pre- and posttest scores of both the fourth- and the fifth-grade students in our statistical analyses. Therefore, we transformed all of the student test scores to z scores per grade.

$37 \times 255 =$	Every day Linda delivers 179 papers. How many papers in total does she deliver in 28 days?
Answer:	Answer:

Figure 5. Two examples of questions in the fifth-grade posttest.

Table 2. Psychometric qualities of all tests.

	n	Cronbach's α	p values		corrected item-total correlations
Fourth-grade pretest	25	.82	$M = .53$	$SD = .22$.20–.50
Fourth-grade posttest	24	.81	$M = .48$	$SD = .23$.13–.52
Fifth-grade pretest	26	.84	$M = .50$	$SD = .17$.19–.56
Fifth-grade posttest	24	.83	$M = .37$	$SD = .15$.19–.52

Statistical analyses

We described the median proportion, and the first and third quartiles for each element of the CFA model, to find out to what extent the teachers in both conditions used elements of the CFA model prior to the intervention. Furthermore, to determine whether the teachers in the CFA condition used goal-directed instruction, assessment, and immediate instructional feedback more frequently than the control teachers, we applied a Mann-Whitney *U* test to compare the teachers' scores on these elements during the observations. The scores for the second and third observations were combined to create an average score for the use of each element during the intervention.

The initial student performance differences between the two conditions, the effect of the CFA model on the students' posttest performance, and the effect of the degree of implementation on the students' posttest performance were all estimated by means of a multilevel regression analysis using MLwiN (Rasbash, Browne, Healy, Cameron, & Charlton, 2015). As the students were nested within classes, we performed a two-level analysis to take the variability at both the class and the student level into account. The effects of both the CFA model and the degree of implementation were corrected for the influence of the students' pretest scores, their gender (using a dummy variable with boy as the reference category), their year grade (using a dummy variable with fourth grade as the reference category), the classes' mean pretest scores, and the teachers' years of teaching experience. Furthermore, we explored possible differential effects of the CFA model and the degree of implementation, for instance, whether the teachers' use of the model was more effective for low-achieving than for high-achieving students or whether it was more effective in Grade 4 than in Grade 5.

Results

Frequency of use of daily CFA

Frequent use of goal-directed instruction, assessment, and immediate instructional feedback is considered to be effective in enhancing student performance. This is why, as a result of the PDP, the CFA teachers were expected to apply these elements more frequently during their lessons than the control teachers. In both conditions, the teachers were observed three times over the course of the intervention to test this hypothesis. One teacher in the control condition was not observed during the first observation round due to illness.

Table 3 provides an overview of the first and third quartiles and the median scores of the teachers in both conditions regarding their daily use of the three elements. The

Table 3. Teachers' daily use of goal-directed instruction, assessment, and immediate instructional feedback (All scores are proportions).

Condition	Elements of daily CFA	N	Observation 1			Observation 2			Observation 3				
			Q1	Mdn	Q3	N	Q1	Mdn	Q3	N	Q1	Mdn	Q3
CFA	Goal-directed instruction	17	.67	1.00	1.00	17	.67	1.00	1.00	17	.67	1.00	1.00
	Assessment		.00	.00	.25		.00	.50	1.00		.25	1.00	1.00
	Immediate instructional feedback		.00	.00	.00		.00	.50	.63		.00	.50	.75
Control	Goal-directed instruction	16	.67	.67	1.00	17	.33	.67	1.00	17	.33	.67	1.00
	Assessment		.00	.00	.00		.00	.00	.00		.00	.00	.00
	Immediate instructional feedback		.00	.00	.00		.00	.00	.00		.00	.00	.00

Note: Q1 refers to Quartile 1, Mdn (Median) refers to Quartile 2, and Q3 refers to Quartile 3.

results of the first observation indicate that the teachers in both the CFA condition and the control condition provided goal-directed instruction. This result was expected, as both sets of curriculum materials are focused on one learning goal per lesson. Neither in the CFA condition nor in the control condition did the teachers frequently use assessments or immediate instructional feedback prior to the project. Mann-Whitney U tests show that there were no significant differences between the two conditions as regards the use of goal-directed instruction ($U = 123.00$, $p = .61$), assessment ($U = 130.00$, $p = .76$), and immediate instructional feedback ($U = 127.50$, $p = 0.54$) prior to the intervention. During the project, the teachers did not differ in their use of goal-directed instruction ($U = 94.00$, $p = 0.07$). However, there were significant differences as regards their use of assessment ($U = 22.00$, $p < 0.001$) and immediate instructional feedback ($U = 32.00$, $p < 0.001$). These results indicate that, as intended, the CFA teachers assessed their students' mastery of the learning goal and subsequently provided immediate instructional feedback more often during the lessons than the teachers in the control condition. The results do not imply, however, that the teachers in the control condition did not provide any instructional feedback during their lessons. In 48% of the 50 observations, the control teachers provided small-group instruction to a preset group of low-achieving students selected on the basis of the half-yearly standardized mathematics tests. This finding appears to be in line with the usual practice in The Netherlands.

Student performance on the mathematics tests

We used the students' test scores to determine the effectiveness of the CFA model. Table 4 shows the mean pre- and posttest scores and standard deviations for the fourth- and fifth-grade students in both conditions. A multilevel regression analysis showed that the fourth-grade students' in the CFA condition scored significantly higher on the pretest than the students in the control condition, with $t(401) = 1.93$ and $p = 0.03$. There were no significant differences in the pretest scores of the fifth-grade students between the two conditions, with $t(383) = 0.77$ and $p = 0.22$. Because of the significant difference in the fourth-grade students' pretest scores, we used the students' pretest scores as a covariate in our statistical analyses.

Comparing two conditions: CFA versus a control setting

Table 5 depicts a summary of the multilevel models that we tested for the prediction of the students' posttest scores. The empty model with the students' posttest score as a

Table 4. Fourth- and fifth-grade students' pre- and posttest scores.

Grade	Condition	N	Pretest			Posttest		
			Scale	Mean	SD	Scale	Mean	SD
4	CFA	196	0–25	14.03	4.70	0–24	11.86	4.55
	Control	206		12.81	4.82		10.94	4.59
5	CFA	185	0–26	13.37	5.58	0–24	9.01	4.96
	Control	199		12.78	5.29		8.87	4.91

Note: The pre- and posttests are not comparable tests. The results cannot be used to establish gain in student performance.

Table 5. Multilevel models predicting students' mathematics posttest scores.

	Models							
	Empty model		Covariate Model		Main Effect Model		Interaction Model	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed part								
Intercept	-0.001	0.053	0.086	0.093	0.098	0.102	0.097	0.102
Pretest			0.713*	0.026	0.713*	0.026	0.712*	0.036
Girl			0.098	0.050	0.098	0.050	0.099	0.050
Fifth grade			-0.017	0.076	-0.017	0.076	-0.017	0.076
Mean pretest score class			-0.034	0.128	-0.023	0.134	-0.023	0.134
Experience teacher			-0.006	0.003	-0.006	0.003	-0.006	0.003
CFA condition					-0.023	0.079	-0.023	0.079
CFA*Pretest							0.002	0.050
Random part								
Variance at class level	0.055	0.023	0.028	0.012	0.028	0.012	0.028	0.012
Variance at student level	0.942	0.049	0.466	0.024	0.466	0.024	0.466	0.024
Deviance	2212.663		1659.460		1659.379		1659.378	
No. of groups	34		34		34		34	
No. of students	786		786		786		786	

* $p < 0.001$.

dependent variable and a random intercept shows that the variability at the student level (0.942) was much higher than that at the class level (0.055). The intra-class correlation coefficient, which indicates similarity among individuals in the same class, is $\rho = 0.055/0.997 = 0.055$. This coefficient is quite low, and thus indicates that the differences within a class were relatively large compared to the differences among classes. In our covariate model, we added our five covariates: the students' pretest scores, their gender, their year grade, the classes' mean z score on the pretest, and the teachers' years of teaching experience. As a result, the deviance decreased by 553.203. Compared to a chi-squared distribution with five degrees of freedom, this decrease indicates that the covariate model fitted our data significantly better than the empty model with $p < 0.001$. In this model, the students' pretest scores appeared to be the only covariate that had a significant positive effect on the students' posttest scores. Next, we added the teachers' participation in the CFA condition as an explanatory variable, which resulted in our main effect model. Adding this variable did not lead to a significantly better model fit ($\chi^2 = .081$, $df = 1$, $p = .776$). Finally, we tested all interactions between the teachers' participation in the CFA condition and the covariates. Adding these interactions did not result in an increased model fit. These findings indicate that in this study the teachers' participation in the CFA condition did not enhance student performance.

The effect of degree of implementation of CFA on student performance

In order to explore whether differences in the degree to which the teachers implemented the CFA model were related to the students' mathematics performance, we analyzed the test data of the students in the CFA condition (381 students in 17 classes) using multilevel models comparable to those in the previous section. We used the implementation scale that included both the teachers' daily use of goal-directed instruction,

Table 6. Fourth- and fifth-grade teachers' implementation of the CFA model in proportions.

Grade	<i>N</i>	Min	Max	Mean	<i>SD</i>
4	9	.42	.91	.75	.16
5	8	.26	.77	.57	.14
4 and 5 combined	17	.26	.91	.67	.17

assessment, and immediate instructional feedback and their weekly use of the quizzes and reports.

Table 6 shows the minimum, maximum, mean, and standard deviation for the teachers' implementation of the CFA model in each separate grade and in both grades combined. The minimum (.26) and maximum scores (.91) in both grades combined show that the spread in implementation makes it worthwhile to investigate whether the degree of implementation had an effect on student performance.

The results of the multilevel model analyses are shown in Table 7. Our empty model indicates that in the CFA condition the variability at the student level (0.945) was much higher than the variability at the class level (0.048). Next, we added our five covariates (the students' pretest scores, their gender, the grade they were in, the classes' mean *z* score on the pretest, and the teachers' years of teaching experience) to the empty model. This led to a significantly better model fit with $\chi^2 = 271.943$, $df = 5$, $p < .001$. The students' pretest scores were a significant predictor for the students' posttest scores. Hereafter, we added the degree of implementation to determine whether it had a significant effect on student performance. However, the model fit was not increased ($\chi^2 = .855$, $df = 1$, $p = .355$). Finally, we added interactions among the covariates and the degree of implementation. We found one significant interaction effect between the students' year grade and the degree of implementation on student performance. Adding this interaction to the model resulted in an increased model fit ($\chi^2 = 7.323$, $df = 1$, $p = .007$).

As we can see in Figure 6, this interaction effect indicates that the degree of implementation has had a significant positive effect on student performance in Grade 5. The slightly negative effect of the degree of implementation on the fourth-grade students' performance was not significant.

Table 7. Multilevel models predicting students' mathematics performance.

	Models							
	Empty model		Covariate Model		Main Effect Model		Interaction Model	
	Coefficient	<i>SE</i>	Coefficient	<i>SE</i>	Coefficient	<i>SE</i>	Coefficient	<i>SE</i>
Fixed part								
Intercept	0.057	0.073	0.169	0.149	-0.098	0.318	0.549	0.347
Pretest			0.713*	0.036	0.713*	0.036	0.713*	0.036
Girl			0.070	0.071	0.067	0.071	0.060	0.071
Fifth grade			-0.004	0.119	0.061	0.136	-1.185*	0.424
Mean pretest score class			-0.069	0.233	-0.063	0.226	0.078	0.188
Experience teacher			-0.010	0.006	-0.009	0.006	-0.007	0.005
Implementation					0.346	0.369	-0.654	0.446
Implementation*Fifth grade							1.895*	0.622
Random part								
Variance at class level	0.048	0.031	0.037	0.020	0.037	0.019	0.015	0.012
Variance at student level	0.945	0.070	0.457	0.034	0.457	0.034	0.458	0.034
Deviance	1072.658		800.715		799.860		792.537	
No. of groups	17		17		17		17	
No. of students	381		381		381		381	

* $p < 0.05$.

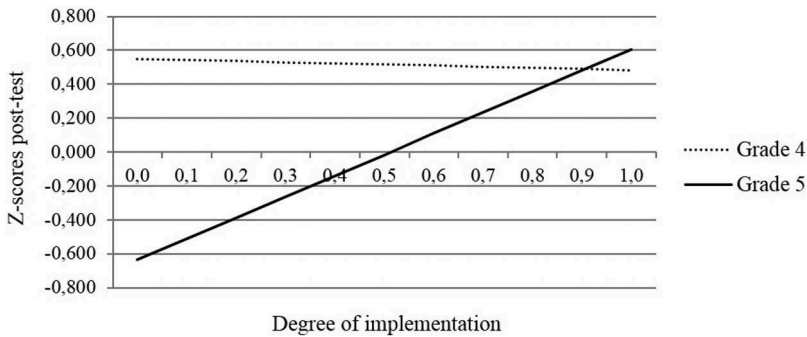


Figure 6. The effect of degree of implementation on the students' mathematics posttest scores in Grades 4 and 5.

Conclusion and discussion

Goal setting, assessment, and providing instructional feedback are, amongst others, considered to be key elements of instructional effectiveness in general (Good et al., 2009; Reynolds et al., 2014; Scheerens, 2016) and formative assessment in particular (Black & Wiliam, 2009; Wiliam & Thompson, 2008). Our study focused on the value added of a CFA model, in which these elements are used on a daily and weekly basis, on student performance. Our analyses show that, as intended, the teachers in the CFA condition used assessments and immediate instructional feedback significantly more often during their lessons than the control teachers. However, this did not result in significant differences in student performance between the two conditions after controlling for student, class, and teacher characteristics. This lack of an effect might be the result of using a condition in which teachers made a modification to their usual practice instead of a business-as-usual control condition. The modification of providing pre-teaching to low-achieving students based on their half-yearly test results may have had an effect in itself. In addition, our observations showed that about half of the teachers in the control condition provided their low-achieving students with small-group instruction during the lesson, as is common practice in The Netherlands. If these teachers combined this common practice with the weekly pre-teaching sessions, this may also have had a positive effect on the low-achieving students' performance.

Still, regardless of the number of times that the control teachers provided the low-achieving students with instructional feedback, in the CFA condition *all students* were assessed and provided the necessary immediate instructional feedback *on a daily basis*. Therefore, the CFA condition should still have been more effective in enhancing student performance. Perhaps a more plausible explanation for the absence of an effect is that although the CFA teachers made use of assessments and immediate instructional feedback more often than the control teachers, they did not do so as frequently as intended. Therefore, the extent to which these elements were used may have been too low to result in significantly better student performance results. Furthermore, the teachers may not have applied the CFA model in an effective manner. Studies have shown that in assessing their students, teachers find it difficult to pinpoint precisely what the problems are, and to instantly decide what help is required (Even, 2005; Heritage, Kim, Vendlinski,

& Herman, 2009). Perhaps, the CFA teachers were not able to properly select those students who needed immediate instructional feedback, to correctly identify their errors, or to determine their zone of proximal development. This may have led to a mismatch between the students' misconceptions and the teachers' instructional feedback (Furtak, Morrison, & Kroog, 2014; Heritage et al., 2009; Schneider & Gowan, 2013), which would have been detrimental to the effectiveness of the CFA model. Unfortunately, we did not qualitatively analyze how the teachers used goal-directed instruction, assessment, and immediate instructional feedback and how the students responded to the immediate instructional feedback, which makes it difficult to draw definite conclusions about the quality of use.

Finally, we investigated whether the degree of implementation of the CFA model was related to student performance. It turned out that the degree of implementation had no significant main effect on the students' mathematics posttest scores. However, there was an interaction effect of the degree of implementation and the students' year grade on student performance. This interaction effect implied that a higher degree of implementation resulted in higher student performance, but only in Grade 5. This result may have been generated by the posttests that were used. The fifth-grade posttest contained task items that were much more complex than those of the fourth-grade posttest. Task complexity has been identified as a moderator of the effect of feedback (Kluger & DeNisi, 1996). This implies that students can show more of what they have learned from the instructional feedback in difficult tasks. The effect of the immediate instructional feedback in our study is therefore perhaps better noticeable in Grade 5. The effects of goal-directed instruction and assessment in Grade 5 may also be more visible because of the test's task complexity (Kingston & Nash, 2011).

The above-described findings seem to indicate that the CFA model as implemented in this study does not lead to enhanced student performance. The small positive effect of the degree of implementation on the fifth-grade students' performance may be an indication that the degree and quality of implementation plays a role in the effectiveness of the CFA model.

In drawing these conclusions, it is important to keep in mind that there are some limitations to our study that may have influenced our results. First, because of the large differences between our two conditions in terms of the intervention and the intensity of the PDP, we decided to assign each school to one of the two conditions before asking them to take part in the study. This approach may have led to selection bias, as some schools may have been more inclined to participate in the condition allotted to them than other schools. Therefore, we cannot be certain that our sample is representative for other teachers and students in The Netherlands or abroad.

Second, although we assume that the teaching practice in the control condition resembled the usual practices in The Netherlands, we cannot be sure that this was in fact the case. Because we did not include a business-as-usual control condition in our study, we therefore cannot make definite claims about the effectiveness of the CFA model (and the control setting) in comparison to the usual practice. It is thus advisable for future research to add a real business-as-usual control condition.

Third, as mentioned above, we did not evaluate how skilled the teachers were in applying the CFA model, what difficulties they encountered during the process, and how the students reacted to the immediate instructional feedback. It would be worthwhile to qualitatively

study these aspects. The results could be used to amend the PDP, if necessary. Catering this support exactly to the teachers' needs would improve the use of the CFA model, rendering the analysis of its effectiveness in increasing student performance more reliable.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The study was supported by a grant from the Netherlands Organisation for Scientific Research (NWO-PROO), The Hague.

Notes on contributors

Marian van den Berg is a PhD candidate at the Groningen Institute for Educational Research (GION education/research) of the University of Groningen. Her main research project focuses on primary mathematics teachers' use of classroom formative assessment (CFA). The project consists of studies regarding the development, implementation, and effectiveness of a CFA model. She is currently working as an educational advisor and researcher at Cedin in Drachten, The Netherlands.

Roel J. Bosker is full professor of education and director of the Groningen Institute for Educational Research (GION education/research) at the University of Groningen, The Netherlands. He specializes in evidence-based education, inequality of educational opportunities, and multilevel statistical modeling. Currently, he leads a research and development team that is adapting the well-tested Success for All program to the Dutch educational context.

Cor J.M. Suhre is a senior researcher in the teacher education department of the University of Groningen. His research interests include the assessment of computational thinking in secondary education, the contribution of computer-assisted instruction to improve problem solving in Physics and Mathematics, and the professional development of teachers. Currently, he is involved in instrument development research aimed at the assessment of critical and creative thinking skills in university-degree programs.

ORCID

Marian van den Berg  <http://orcid.org/0000-0002-4292-2535>

Roel J. Bosker  <http://orcid.org/0000-0002-1495-7298>

Cor J.M. Suhre  <http://orcid.org/0000-0001-5687-758X>

References

- Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability*, 26(2), 153–176. doi:10.1007/s11092-013-9188-4
- Ashford, S. J., & De Stobbeleir, K. E. M. (2013). Feedback, goal setting, and task performance revisited. In E. A. Locke & G. P. Latham (Eds.), *New developments in goal setting and task performance* (pp. 51–64). New York, NY: Routledge.

- Assessment Reform Group. (2002). *Assessment for Learning: 10 principles. Research-based principles to guide classroom practice*. Retrieved from <https://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf>
- Ateh, C. M. (2015). Science teachers' elicitation practices: Insights for formative assessment. *Educational Assessment, 20*(2), 112–131. doi:10.1080/10627197.2015.1028619
- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research, 85*(2), 89–99. doi:10.1080/00220671.1991.10702818
- Birman, B. F., Desimone, L., Porter, A. C., & Garet, M. S. (2000). Designing professional development that works. *Educational Leadership, 57*(8), 28–33.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. doi:10.1007/s11092-008-9068-5
- Callingham, R. (2008). Dialogue and feedback: Assessment in the primary mathematics classroom. *Australian Primary Mathematics Classroom, 13*(3), 18–21.
- Carnine, D. (1980). Preteaching versus concurrent teaching of the component skills of a multiplication algorithm. *Journal for Research in Mathematics Education, 11*(5), 375–379. doi:10.2307/748628
- Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education, 77*(3), 211–244. doi:10.1177/003804070407700302
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Conderman, G., & Hedin, L. (2012). Classroom assessments that inform instruction. *Kappa Delta Pi Record, 48*(4), 162–168. doi:10.1080/00228958.2012.733964
- Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. New York, NY: National Commission on Teaching & America's Future.
- Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Stanford, CA: National Staff Development Council and the School Redesign Network at Stanford University.
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199. doi:10.3102/0013189X08331140
- Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7), 1–11.
- Even, R. (2005). Using assessment to inform instructional decisions: How hard can it be? *Mathematics Education Research Journal, 17*(3), 45–61. doi:10.1007/BF03217421
- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199–208. doi:10.1177/001440298605300301
- Fullan, M. (2007). *The new meaning of educational change* (4th ed.). New York, NY: Teachers College Press.
- Furtak, E. M., Morrison, D., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education, 98*(4), 640–673. doi:10.1002/sce.21122
- Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*(4), 360–389. doi:10.1080/08957340802347852
- Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945. doi:10.3102/00028312038004915
- Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: Children's minds, teachers' minds. *Human Development, 52*(2), 109–128. doi:10.1159/000202729
- Good, T. L., Wiley, C. R. H., & Florez, I. R. (2009). Effective teaching: An emerging synthesis. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (pp. 803–816). New York, NY: Springer.
- Grierson, A. L., & Woloshyn, V. E. (2013). Walking the talk: Supporting teachers' growth with differentiated professional learning. *Professional Development in Education, 39*(3), 401–419. doi:10.1080/19415257.2012.763143

- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487
- Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, 49(3), 333–362. doi:10.1002/tea.21004
- Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers.
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31. doi:10.1111/j.1745-3992.2009.00151.x
- Heritage, M., & Niemi, D. (2006). Toward a framework for using student mathematical representations as formative assessments. *Educational Assessment*, 11(3), 265–282. doi:10.1080/10627197.2006.9652992
- Inspectie van het Onderwijs. (2010). *Opbrengstgericht werken in het basisonderwijs: Een onderzoek naar opbrengstgericht werken bij rekenen-wiskunde in het basisonderwijs* [Data-driven decision-making in primary education: A study after data-driven decision-making in primary mathematics education]. Utrecht: Author.
- Keuning, T., & Van Geel, M. J. M. (2016). *Implementation and effects of a schoolwide data-based decision making intervention: A large-scale study* (Doctoral dissertation). Retrieved from <http://dx.doi.org/10.3990/1.9789036541190>
- Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37. doi:10.1111/j.1745-3992.2011.00220.x
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. doi:10.1037/0033-2909.119.2.254
- Kouba, V. L., & Franklin, K. (1995). Multiplication and division: Sense making and meaning. *Teaching Children Mathematics*, 1(9), 574–577.
- Kraemer, J.-M. (2009). *Balans over de strategieën en procedures bij het hoofdrekenen halverwege de basisschool* [Overview of mathematical strategies and procedures half-way primary education]. Arnhem: Cito.
- Lalley, J. P., & Miller, R. H. (2006). Effects of pre-teaching and re-teaching on math achievement and academic self-concept of students with low achievement in math. *Education*, 126(4), 747–755.
- Lantz, M. E. (2010). The use of “Clickers” in the classroom: Teaching innovation or merely an amusing novelty? *Computers in Human Behavior*, 26(4), 556–561. doi:10.1016/j.chb.2010.02.014
- Leahy, S., Lyon, C., Thompson, M., & William, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, 63(3), 19–24.
- Little, J. W. (1993). Teachers’ professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis*, 15(2), 129–151. doi:10.3102/01623737015002129
- Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, 15(5), 265–268. doi:10.1111/j.1467-8721.2006.00449.x
- Lumpe, A., Haney, J. J., & Czerniak, C. M. (2000). Assessing teachers’ beliefs about their science teaching context. *Journal of Research in Science Teaching*, 37(3), 275–292. doi:10.1002/(SICI)1098-2736(200003)37:3<275::AID-TEA4>3.0.CO;2-2
- Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, 47(2), 71–85. doi:10.1080/00461520.2012.667064
- Marzano, R. J. (2006). *Classroom assessment & grading that work*. Alexandria, VA: Association for Supervision and Curriculum Development.
- McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research & Evaluation*, 18(2), 1–15.
- Ministerie van Onderwijs, Cultuur en Wetenschap. (2014). *Kerncijfers 2009–2013 onderwijs, cultuur en wetenschap* [Key figures 2009–2013 education, culture and science]. Den Haag: Author. Retrieved from <https://www.rijksoverheid.nl/documenten/jaarverslagen/2014/05/21/ocw-kerncijfers>
- Moon, T. R. (2005). The role of assessment in differentiation. *Theory into Practice*, 44(3), 226–233. doi:10.1207/s15430421tip4403_7

- Moss, C. M., Brookhart, S. M., & Long, B. A. (2013). Administrators' roles in helping teachers use formative assessment information. *Applied Measurement in Education*, 26(3), 205–218. doi:10.1080/08957347.2013.793186
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
- Organisation for Economic Co-operation and Development. (2014). *PISA 2012 results: What students know and can do (Volume I, Revised edition, February 2014): Student performance in mathematics, reading and science*. Paris, France: Author. doi:10.1787/9789264208780-en
- Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results (Volume 1). Excellence and equity in education*. Paris, France: Author. doi:10.1787/9789264266490-en
- Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, 44(4), 921–958. doi:10.3102/0002831207308221
- Pfister, M., Moser Opitz, E., & Pauli, C. (2015). Scaffolding for mathematics teaching in inclusive primary classrooms: A video study. *ZDM Mathematics Education*, 47(7), 1079–1092. doi:10.1007/s11858-015-0713-4
- Porter, A. C., Garet, M. S., Desimone, L., Yoon, K. S., & Birman, B. F. (2000). *Does professional development change teaching practice? Results from a three-year study*. Washington, DC: U.S. Department of Education.
- Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2015). MLwiN Version 2.35 [Computer software]. Bristol: Centre for Multilevel Modelling.
- Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, 25(2), 197–230. doi:10.1080/09243453.2014.885450
- Ritzema, E. (2015). *Professional development in data use: The effects of primary school teacher training on teaching practices and students' mathematical proficiency* (Doctoral dissertation). Groningen: GION, University of Groningen.
- Ryan, J., & Willams, J. (2007). *Children's mathematics 4–15: Learning from errors and misconceptions*. Maidenhead: Open University Press/McGraw-Hill Education.
- Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Dordrecht: Springer. doi:10.1007/978-94-017-7459-8
- Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education*, 26(3), 191–204. doi:10.1080/08957347.2013.793185
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14. doi:10.3102/0013189X029007004
- Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership*, 63(3), 66–70.
- Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). Mahwah, NJ: Erlbaum.
- Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York, NY: McGraw-Hill.
- Vygotsky, L. S. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- William, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NY: Erlbaum.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x
- Wylie, E. C., & Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice*, 22(1), 140–160. doi:10.1080/0969594X.2014.990416