



Supporting primary school teachers' classroom assessment in mathematics education: effects on student achievement

Michiel Veldhuis^{1,2}  · Marja van den Heuvel-Panhuizen^{1,3}

Received: 16 October 2018 / Revised: 25 April 2019 / Accepted: 7 May 2019 /

Published online: 25 May 2019

© The Author(s) 2019, corrected publication 2019

Abstract

In a three-phase study, with a total of 40 third-grade teachers and their 830 students, teachers were supported to use classroom assessment techniques (CATs) to reveal their students' knowledge of number operations. In phase I, four teachers and 66 third-grade students participated in five monthly workshops in which CATs were co-designed and their use was discussed. In phase II, the first phase was replicated with four workshops with six different teachers and 148 third-grade students. In these two exploratory phases, we evaluated student achievement on a standardized national mathematics test in a pre-/posttest design and compared changes herein to changes in the national norm sample. In phase III, a control condition was added to the design to experimentally investigate the effect on student achievement with 30 teachers and 616 third-grade students. Teachers were randomly assigned to participate in 0, 1, 2, or 3 1-hour workshops. In all three phases, we found a significant increase in students' mathematics achievement scores on the standardized mathematics test. In phase III, the increase was significantly larger in the classes of teachers participating in three workshops than in classes with less workshops. Additionally, results from the analysis of classroom observations, feedback forms, and interviews indicate that teachers could easily integrate the CATs into their practice and could gather valuable information on their students. The results from the different phases of this study combined indicate that supporting teachers in their development and use of classroom assessment in mathematics may contribute to the improvement of students' mathematics achievement.

Keywords Classroom assessment · Mathematics education · Professional development · Student achievement · Teachers

The study reported on in this article was carried out in the ICA (Improving Classroom Assessment) project that was supported by a grant from the Netherlands Organization for Scientific Research (NWO MaGW/PROO: Project 411-10-750).

✉ Michiel Veldhuis
m.veldhuis@uu.nl

Extended author information available on the last page of the article

Introduction

Classroom assessment

Assessment, with the purpose of making informed decisions about how instruction should be continued, is embedded in teachers' teaching practice and is called formative assessment. During lessons, teachers need evidence about student learning to be adaptive to their students' specific learning needs (Wiliam 2007). There are many different ways of carrying out formative assessment, the type of formative assessment we focus on is the type that is completely in "the hands of teachers" (Van den Heuvel-Panhuizen and Becker 2003, p. 683) and is often called *classroom assessment* (e.g., Andrade and Brookhart 2019; Black and Wiliam 1998; Shepard 2000; Stiggins and Chappuis 2005). In the hands of the teachers, they decide when and how to assess and what to do with the assessment results they obtained by providing students with a carefully selected set of problems. Contrary to the past, when there were often concerns about the reliability of teachers' judgements of students' performances (Parkes 2013), the role of the teacher now includes gaining insights into their students' progress. Such a roll is seen as crucial for adapting their teaching to students' needs (Harlen 2007). Classroom assessments that are inseparably intertwined with instruction, such as asking questions, observing students, and giving quizzes or teacher-made written assignments, can provide insights into students' thinking and into what productive instructional steps might be taken next (Andrade and Brookhart 2019; Shepard et al. 2017). Notwithstanding the importance of the use of such assessment activities, teachers do not often report using them in practice (e.g., Frey and Schmitt 2010; Veldhuis et al. 2013). This might be due to such assessments having to be clearly linked to the taught, or to be taught, content for the realization of effective formative assessment (Hondrich et al. 2016). Collaborating with teachers and providing content-specific assistance is viewed as a fruitful way to improve teachers' abilities to effectively use formative assessment in their classrooms (Kim 2019; Yin and Buck 2019). Our study is designed to learn more about the feasibility and effectiveness of supporting primary school teachers in the development and use of domain- and topic-specific classroom assessment in mathematics.

Previous research on classroom assessment

Effects of classroom assessment on student achievement

In educational research, often large positive effects of teachers' use of classroom assessment on student achievement have been reported (studies reviewed in Black and Wiliam 1998, or more recently in Briggs et al. 2012, and in Kingston and Nash 2011). Notwithstanding the fact that scholars of these studies in most cases refer to classroom assessment or formative assessment when they discuss their research, the similarity of the operationalization they opted for is quite low; many different definitions and assessment methods have been used under the same umbrella term of

classroom assessment (see Veldhuis and Van den Heuvel-Panhuizen 2014). What strings studies on classroom assessment together, however—in addition to the terminology used—is that most interventions are focused on enhancing teachers' subject knowledge and promoting the use of assessments, thus allowing teachers to subsequently provide formative feedback to students. Formative feedback means “information communicated to the learner that is intended to modify the learner's thinking or behavior for the purpose of improving learning” (Shute 2007, p. 1). This type of feedback has been found to be most effective for motivating students and improving their learning (e.g., Hattie and Timperley 2007). In addition to the fact that the research projects on the effects of classroom assessment and their interventions were small scale, their comparability has been criticized because of the different conceptualizations of what classroom assessment entails (e.g., Bennett 2011). Even though the specificities of studies that have shown the effect of classroom assessment are different, their results do point to the effectiveness of the use of classroom assessment for improving students' mathematics achievement. On the basis of such empirical results, recently, in the USA, the National Council of Teachers of Mathematics (NCTM 2013) strongly endorsed teachers using classroom assessment strategies in their daily instruction in mathematics education. The basic idea behind the effectiveness of teachers' use of classroom assessment is that it can lead to teachers gaining more relevant and useful information on their students' understandings and skills. This allows them to subsequently better adapt their teaching to their students' needs, which in turn is expected to lead to improved student achievement. A recent study with 45 primary school teachers in Sweden (Andersson and Palm 2017) confirmed this line of reasoning with a yearlong intensive professional development program on using formative assessment strategies, resulting in students of these teachers significantly outperforming students in the control group on a mathematics posttest.

Strategies for using classroom assessment in mathematics

A number of scholars have focused on providing teachers with strategies for using classroom assessment in mathematics (e.g., Andersson and Palm 2017; Keeley and Tobey 2011; Leahy et al. 2005; Torrance and Pryor 2001; Wiliam 2011; Wiliam et al. 2004). These strategies for classroom assessment often concern activities that are familiar to teachers but that are now used with a clear assessment focus (e.g., Wiliam 2011). An example of such an assessment strategy is an all-students response system with multiple choice cards (ABCD cards); this means that the teacher poses a question that touches a key aspect of what is currently taught in class and to which all students respond individually by holding up a card. Teachers can use the information gathered in this way to go over a particular explanation or subject again, or instead move on; an instructional decision teachers make on a day-to-day basis. Such activities can be seen as an operationalization of the framework Wiliam and Thompson (2007) proposed, consisting of five key strategies that make up teachers' and students' formative assessment practice. These strategies are aimed at assisting teachers and students in establishing the following three pieces of information about the learner: where (s)he is going, where (s)he is right now, and how to get there (see also Stiggins et al. 2004). In this framework, the how-to-get-there part consists of the teacher providing formative feedback that moves learners forward

(Hattie and Timperley 2007). The other key strategies are related to sharing learning goals, making use of effective classroom discussions and learning tasks that elicit evidence of student understanding, and activating students as owners of, and resources for, their own learning (Wiliam and Thompson 2007, p. 63).

To investigate how teachers can acquire useful knowledge about their students' learning,

a number of studies have investigated the influence of interventions focusing on supporting teachers in their assessment practice. For example, in Wiliam et al. 2004), science and mathematics teachers were supported in their assessment practice over the course of two school years. This resulted in large learning gains, but this study—as it was not an experimental study—was focused on determining principles for practice and not so much on establishing the effect of the support. Another example is the investigation of the influence of a long-term (4 years) professional development program on formative assessment design in biology education on teachers' formative assessment abilities (Furtak et al. 2016). They found that teachers improved their assessment abilities on several aspects, such as question quality, interpretation of student ideas, and feedback quality, but surprisingly not on task quality; meaning that they did not provide students with higher quality tasks. In this sense, the professional development does not always lead to the envisioned results. A similar result was found in a large-scale study in the USA (Randel et al. 2016) in which a widely used program on classroom assessment was experimentally evaluated, and neither students' mathematics achievement nor teachers' assessment practices appeared to have been influenced. Another example is a study on formative assessment in science education: despite training teachers in the use of formative assessment, their students' achievement levels in science did not improve (Yin et al. 2008). These scholars hypothesized that this result was most probably due to a suboptimal implementation of the formative assessment strategies. A possible other reason could be that the aforementioned studies all aimed to assist teachers in their assessment by providing them with general strategies for formative assessment. This means that although they were meant to be used in the domain of mathematics (or science), there was no close relationship with the taught content when assessing the students. Instead, the focus was more on the format of the assessment techniques or the accompanying feedback. Formative assessment techniques that are closely connected to the taught mathematics have the potential to really inform teachers' further instruction. Such a content-dependent approach was chosen by Phelan et al. (2012). In their study, teachers were supported to assess students' learning in pre-algebra. To find out what had to be assessed, an expert panel was organized to map algebra knowledge and its prerequisites. This map was used to design the questions that could provide teachers with the necessary information, which turned out to have had a positive impact on students' learning (Phelan et al. 2012).

Features of high-quality teacher professional development

To ensure that teachers optimally implement what they learn about classroom assessment, a number of features of high-quality teacher professional development have to be carefully considered (see e.g., Garet et al. 2001). Although the direct usefulness of such lists of features has been questioned (e.g., Beswick et al. 2016; Kennedy 2016), they do provide actionable ideas to assist shaping professional development programs. For

example, as shown in the study by Phelan et al. (2012), it is important to focus on the mathematical content. However, as Kennedy's (2016) findings suggest, a professional development program should not exclusively focus on content knowledge, but should also help teachers to expose student thinking. Classroom assessments should be linked to the learning trajectories, the standards, the curriculum, and the textbooks. In connection with this, the professional development should ensure that teachers are aware of the mathematics teaching and learning trajectories (or learning progressions) of the grades they teach (Bennett 2011). To achieve this knowledge, mathe-didactical analyses (Van den Heuvel-Panhuizen and Teppo 2007) of the mathematical content to be taught should be carried out leading to an in-depth knowing of what concepts and skills are important, how models and strategies are related, and how they evolve over the years. Without this knowledge, worthwhile assessments are impossible. Secondly, having teachers collectively discuss and reflect on student work or reactions can support teachers' engagement in active learning (e.g., Lin 2006). Finally, the professional development should also take place over a prolonged period of time, because it often takes time for teachers to implement what they learned about classroom assessment (e.g., Black and Wiliam 1998) although changes have also been shown to occur rather rapidly (see Liljedahl 2010).

To establish a collaborative teacher learning community, gradualism, flexibility, choice, accountability, and support are deemed necessary (Wiliam 2007). Repeated meetings allow teachers to integrate what they have learned in their own practice and see their own practice in new ways, leading to new thinking. When teachers are also involved in the design of the classroom assessment, in collaboration with other teachers, it can raise their awareness of not only students' mathematics learning difficulties but also their decisions about remedial instruction (Lin 2006). Furthermore, for teachers to use the information gathered through their use of assessment, they have to be involved in the process of development of the assessment, have an active role in how they use the assessment, and be able to use the assessment information (Wilson and Sloane 2000, p. 191). Recently, Heitink et al. (2016) described prerequisites for the implementation of classroom assessment (or what they call assessment for learning) in teachers' classroom practice that reflect the previously mentioned features of effective professional development. Among the main prerequisites was that classroom assessment tasks should be meaningful and closely integrated into classroom instruction. Furthermore, teachers have to be able to interpret assessment information on the spot and the assessment should provide useful and constructive feedback that can be used for further instruction. Integrating these insights on professional development and the findings about the effectiveness of teachers' use of classroom assessment led us to set up our current study.

The present study

In the study described here, we focused on investigating the effect of supporting grade 3 teachers in the development and use of classroom assessment in their classrooms in mathematics on their students' achievement. In giving this support, we strived to integrate the previously mentioned features of classroom assessment in mathematics education and effective professional development. In order to keep the classroom assessment closely connected to the taught mathematics, the assessment was based on mathe-didactical analyses of the important mathematical content that was at hand in the

period of the school year when the support would be provided (see also Kim 2019). The mathematics textbook teachers used in their classrooms was the main resource for determining this content because, in the Netherlands, the textbook content can be considered the curriculum (e.g., Meelissen et al. 2012). In addition, to have the assessment tied with the teaching process, the use of classroom assessment was conceived as the use of classroom assessment techniques (CATs); short teacher-initiated assessment activities that reveal students' understanding of a particular mathematical concept or skill. Our main research question was *What effect does supporting third-grade teachers in the development and use of CATs in mathematics education have on their students' mathematics achievement?* How teachers use the different CATs could have an influence on the resulting effects on student achievement; therefore, we also qualitatively explored how teachers use and implement CATs in their classroom practice.

Method

Research design

Our study consisted of three phases carried out in three consecutive years from 2012 to 2014 (see Table 1 for the general planning of the three phases of the study). In phase I, we performed an explorative study with pre- and posttest from January to June 2012. A team of teachers and researchers performed mathe-didactical analyses of the mathematical content of that period of the school year were performed and, on the basis of these, CATs were developed and designed that were inspired by insights from research in mathematics education and formative assessment (see more in the “Material: CATs” section and the Appendix Table 4). The teachers participated in five 1-hour workshops with the first author, over the course of the second semester of grade 3 and used the CATs in between the workshops. The same research design was used in phase II that took place from January to June 2013 and in which the teachers participated in four 1-hour workshops.

In phase III, an experimental approach with pre- and posttest and control group was used. This phase took place from January to June 2014 with four conditions: a control

Table 1 General design of the three phases of the study from 2012 to 2014

	Condition	January	February	March	April	May	June
Phase I (2012)	–	Pretest	Workshop	Workshop	Workshop	Workshop	Workshop Posttest
Phase II (2013)	–	Pretest	Workshop	Workshop	Workshop	Workshop	Posttest
Phase III (2014)	Third experimental	Pretest	Workshop	Workshop	Workshop		Posttest
	Second experimental	Pretest	Workshop		Workshop		Posttest
	First experimental	Pretest	Workshop				Posttest
	Control	Pretest					Posttest

Pretest refers to the regular student-monitoring test mid-grade 3; Posttest refers to the regular student-monitoring test end-grade 3

(business-as-usual) condition, in which teachers did not partake in any workshops, and three experimental conditions in which the number of 1-hour workshops varied from one to three.

Participants

In phase I, four female third-grade teachers (and their 66 students) from four schools in a mid-sized town in the west of the Netherlands volunteered to participate. These teachers had classes of between 13 and 24 students. In their classes, three different textbooks were used, all inspired by an approach to mathematics education in which much attention is placed on the use of meaningful contexts for developing mathematical understanding and is known as Realistic Mathematics Education (RME) (see, e.g., Van den Heuvel-Panhuizen and Drijvers 2014).

In phase II, two male and four female third-grade teachers from five different schools participated (and their 148 students). The schools were situated in urbanized areas in the vicinity of two big cities in the west of the Netherlands, with highly mixed student populations. The classes of these teachers contained between 17 and 29 students. These schools were recruited by approaching a number of schools in the western region of the Netherlands by e-mail based on school data (such as postal address and e-mail) publicly available on the Internet. The involved teachers all used the same RME-based textbook: *De Wereld in Getallen* (The World in Numbers) (Huitema et al. 2009). The textbook was used as an inclusion criterion to control for the influence of the curriculum, as most teachers in the Netherlands follow their textbooks faithfully. The teachers participated in groups of three teachers in the four workshops.

In phase III, 30 third-grade teachers (and their 616 students) participated. These teachers worked at 25 different primary schools from all over the Netherlands; ranging from rural parts to densely populated areas. From the 33 classroom teachers who reacted positively to an e-mail request sent to schools, three teachers dropped out during the study: one for health reasons and the remaining two due to logistical concerns. The final sample contained five male and 25 female teachers who were randomly distributed over the conditions. As we expected, teachers that were supposed to participate in three workshops to be more prone to miss one or more sessions, the number of teachers and students in this condition was the largest by design.¹ See Table 2 for the numbers of teachers and students of the three phases. For phase III, this table also includes the distribution over the four conditions. Within each condition, we organized the teachers in small groups based on their schools' geographical locations. This means that in each experimental condition, there were at least two separate groups of teachers attending the workshops. Concerning the comparability of teachers in the different conditions, teachers from rural and more or less urban areas were present in all conditions and they all used the same textbook. Furthermore, the teachers in the different conditions were ignorant of the fact that there were other conditions with more or less frequent workshops.

¹ A few teachers indeed did not attend all workshops that they were supposed to. Six teachers attended two workshops instead of three and three teachers attended just one workshop instead of three. As these teachers attended fewer workshops, they could be considered as effectively being in another experimental condition.

Professional development: workshops

In phase I and II, during the second semester of grade 3, the teachers and the first author convened in monthly 1-hour workshops on the development and use of the CATs. In phase III, the frequency depended on the condition teachers were assigned to (see Table 1). The workshops were held at the schools of one of the participating teachers. The content and procedure of the workshops in the different phases of the study were identical and are described in more detail in the following.

In the workshops, teachers and researchers worked together in a stepwise procedure to develop the CATs. The first step was concerned with the determination of the important mathematics content (building blocks) of that period of the school year and after that, to think about ways to find out whether students had mastered the prerequisite skills and knowledge. The discussions during the workshops revolved around the mathe-didactical analysis of the important topics of that time period in the teaching and learning trajectory. These topics were addition and subtraction with crossing ten, knowledge of multiplication or division tables, and solving word problems; these are also prominently present in the teaching and learning trajectory of grade 3 in the Netherlands. After having identified and discussed these issues, the discussion turned to their (formative) assessment. From the second workshop onwards, teachers also shared their experiences of the preceding weeks: which CATs they used, why they used them, in what form, how their students reacted, what they thought of the activities, what information they collected by the CATs, and what they did as a follow-up with this information. Possible new CATs were discussed and the researcher distributed supporting material. In all workshops, attention was paid to the didactical reasons for using the techniques and how asking particular questions could give teachers access to a deeper level of students' skills and understanding. Moreover, it was discussed that by giving students feedback about the findings of the assessments, they could also become explicitly aware of their own understanding.

Material: CATs

In the workshops, nine CATs were developed and distributed consisting of short assessment activities of less than 10 min to the teachers. The CATs addressed important content of the teaching and learning trajectory of grade 3: operations with numbers up to 100 and 1000, the same topics that teachers identified as important based on their

Table 2 The number of teachers and third-grade students in the three phases of the study

	Condition	Number of teachers	Number of students	Average number of students per class
Phase I	–	4	66	16.5
Phase II	–	6	148	24.7
Phase III	Third experimental	10	207	20.7
	Second experimental	8	172	21.5
	First experimental	6	138	23.0
	Control	6	99	16.5
Total		40	830	20.8

experience and knowledge of the students' performance in class. The design of these CATs was inspired by principles of existing assessment techniques in mathematics education (see, e.g., Keeley and Tobey 2011; Leahy et al. 2005; Torrance and Pryor 2001; Wiliam 2011). These principles reflect the three questions students and teachers try to answer about students' mathematical understanding: where the student is going, where the students are right now, and how to get there (see Stiggins et al. 2004; Wiliam and Thompson 2007). To answer these questions, every CAT consisted of a classroom activity in which the teacher could get a quick overview of students' skills and knowledge of the relevant mathematical content. On the basis of this information, the teacher could then decide how to provide formative feedback to the students. More information about the content, format, and goal of the nine different CATs is given in the Appendix Table 4; in the following, we illustrate three CATs in more detail.

Example 1: CAT—*Crossing ten and more*

In this CAT, the teacher asks a series of questions to all students that can be answered quickly with “yes” or “no.” Students all have a red and a green card to show their answers, allowing the teacher to get an immediate overview of all students' responses. In our study, the teachers used this CAT (Fig. 1) for assessing whether students had ready knowledge about whether a total of two numbers was under or over ten, e.g., whether this is the case for the numbers 7 and 4. When asking these questions, adding or summation was purposefully not mentioned to avoid inciting students to calculate. Such instant number fact knowledge is needed to perform many numerical operations, such as additions and subtractions with two-digit numbers. For solving these problems, students have to be able to instantaneously identify whether crossing is the case, because this has consequences for the solution strategy. After crossing ten, the teacher can use this CAT with crossing 100 and 1000. In this way, the teacher can also assess whether the students understand the analogy between different number domains. For some students, 70 and 40 will be a new problem whereas others know immediately that what applied to 7 and 4 also applies to 70 and 40. Analogously, 700 and 400 will be new to some, but easy to those that understand the analogy.

Example 2: CAT—*Easy or difficult*

Another CAT, also related to students' knowledge of number facts and number operations, is *Easy or difficult*. Here, the main purpose is for the teacher to find out whether students are aware of the difficulties some number operations can have and whether they can reflect on these difficulties. In this CAT, students were presented with a series of problems and

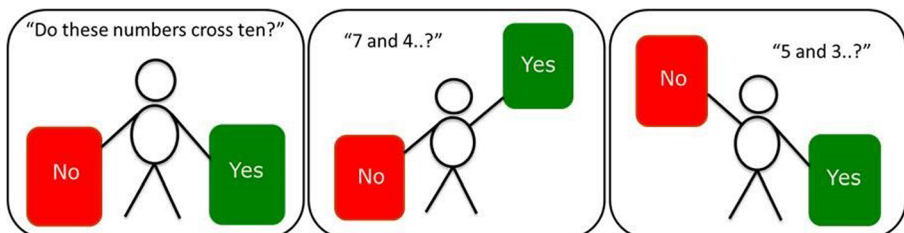


Fig. 1 CAT 1—Crossing ten and more

asked to identify whether they were, according to them, easy or difficult (see VVan den Heuvel-Panhuizen et al. 1995). They are given a worksheet containing two columns of similar problems differing on particular aspects. These aspects pertain to whether a ten is crossed or not (e.g., $12 + 9$ or $13 + 12$; $26 - 7$ or $35 - 4$). Other aspects are, for example, the size of the numbers (e.g., $20 + 40$ or $200 + 40$) or the order in which the numbers are presented (e.g., $54 + 20$ or $20 + 54$). Of each pair of problems, the student had to circle the easiest one, without calculating the answer. Important to note is that there are no right or wrong answers in this CAT, what is easy and what is difficult can differ between students. Afterwards, the students exchange their worksheets and discuss their reasoning, explaining differences or commonalities.

Example 3: CAT *Word problem difficulties*

This CAT consisted of the teacher setting up an experiment focusing on word problems (Fig. 2). Despite the high value attached to teaching students the ability to use mathematics to solve context problems, they often struggle with this. These difficulties can be due to a variety of reasons, for example, miscomprehension of the text, failure in transforming the problem situation into a mathematical problem, getting stuck in solving the mathematical problem itself, or a combination of these factors. For the teacher to find out where the problem lies for individual students, this assessment technique works as follows. Students solve a series of problems as word problems and later, the same problems in a different format, namely as bare number problems (e.g., Van den Heuvel-Panhuizen 1996). Then, the teacher can compare for every student and for the class as a whole, for every problem and for the total of problems, the results in the two formats.

Measures

Quantitative data

In all phases of the study, a pre-/posttest evaluation of students' mathematics achievement was carried out to investigate whether a learning effect following teachers' use of

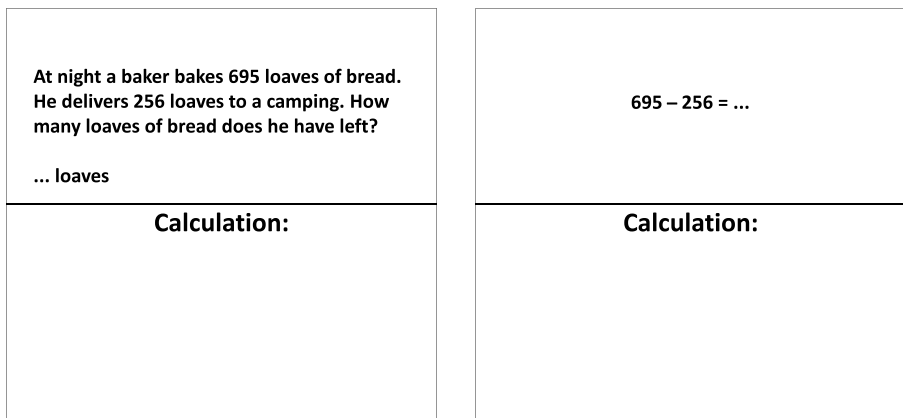


Fig. 2 Two worksheets of the CAT *Word problem difficulties*, presented as a word problem (on the left) and as a bare number problem (on the right)

CATs would be present. The pretest data consisted of the results from the nationally used midyear mathematics student-monitoring system test (in Dutch: *Cito Leerling Onderwijs Volg Systeem* [LOVS]); the results from the end of year test served as posttest data (Janssen et al. 2006). The teachers administered the tests, as per usual, in their own classes. These tests are used in most schools in the Netherlands and the items cover different mathematical topics that are important in grade 3, such as number knowledge, addition and subtraction until 1000, and division and multiplication with one- to three-digit numbers. The test scores are mathematics achievement scores calculated through item response theory models (for more information about the tests' construction, reliability, and validity, see Janssen et al. 2010). The reliability of the pretest is 0.93 and of the posttest 0.95, with a correlation of 0.96 (see Janssen et al. 2010). The same latent score scale continues from grade 1 until 6, with averages from midyear grade 1, 26.4 points, to 107.4, at the end of grade 6. As such, the "normal" achievement score (as defined by the norm reference sample) changes from midyear grade 3: 69.0 ($SD = 14.52$) to end of year grade 3: 74.1 ($SD = 14.48$), that is, a gain of + 5.1 points in one semester.

Qualitative data

In phase I and II, regular classroom observations of every teacher were conducted in between workshops, of which field notes were taken and summarized. These observations were complemented with short unstructured interviews with teachers and students. In phase III, due to logistical concerns, the regular classroom observations and interviews were impossible; only one classroom observation per teacher was done. In all phases, teachers were asked to register their evaluation of the CATs, what they learned from them, and whether they thought them to be useful, on a feedback form. Also, what teachers said during the discussions in the workshops was noted.

Data analysis

The pretest and posttest mathematics achievement data of the three phases were analyzed descriptively (M , SD , and correlations) in general, and per condition in phase III. Gain scores were calculated to compare the learning gains of the students in the different phases and compare their gains with the national norm reference sample. Although there is quite some discussion as to whether the use of null hypothesis significance testing is warranted (e.g., Trafimow and Marks 2015), we report the results of analysis of covariance of the posttest scores with pretest scores as a covariate, to give the reader, in addition to the descriptive statistics, an idea about the statistical significance of differences.

The qualitative data analysis was in the first place aimed at illustrating how teachers took up the techniques in their classrooms. To go a small step further than just illustration, all data from the observations, informal interviews, discussions, and feedback forms were thematically grouped into the following categories, reflecting important aspects of the formative assessment cycle: (1) teachers' use and adaptations of the CATs, (2) teachers' (and students') ideas about the usefulness of the CATs, and (3) teachers' instructional decisions following using the CATs. From these groupings, we formulate more general descriptions of how the teachers took up the CATs in their practice.

Results

Effects on students' mathematics achievement

In all three phases of the study, there were no significant differences between the mathematics achievement scores of students in the different conditions on the pretest. The average mathematics achievement of students increased from midyear (pretest) to end-of-year (posttest) testing (see Table 3). About 90% of the students in phase I and II showed improvement (defined as a positive gain score) from pretest to posttest. The mean differences between pretest and posttest and the effect sizes in phase I (+ 9.7, $d = 0.70$) and II (+ 7.6, $d = 0.53$) were notably larger than in the national norm reference sample (+ 5.1, $d = 0.36$).

In the experimental part of the study, phase III, this effect on students' mathematics achievement was experimentally evaluated. In the control condition, the first experimental condition, and the second experimental condition, the proportion of students (79%, 81%, and 82%, respectively) showing improvement was lower than that in the third experimental condition (89% of the students improved). From the descriptive statistics in Table 3, it becomes clear that students in the control condition seemed to improve slightly more (gain of + 6.6 points) than those in the first (+ 5.4) and second (+ 5.6) experimental conditions, but students in the third experimental condition improved the most by far (+ 8.2, $d = 0.59$). Using an analysis of covariance (ANCOVA) with condition as between-subjects factor, posttest score as dependent variable and pretest score as covariate showed that this difference is also statistically significant ($F(3, 601) = 3.8, p = .010, \eta_p^2 = .019$). Post-hoc pairwise comparisons showed that the differences on the adjusted posttest means (based on the scores on the covariate: the pretest score) were also significant between the third experimental condition ($M_{adj} = 80.02$) and all other conditions (second experimental $M_{adj} = 77.69, p = .003, d = 0.17$; first experimental $M_{adj} = 77.96, p = .013, d = 0.15$; control $M_{adj} = 78.13, p = .040, d = 0.13$). Repeating this analysis with only those teachers that

Table 3 Pre- and posttest mathematics achievement scores, gain scores, and effect sizes for phase I, II, and III and the national norm reference sample

Condition	Pretest			Posttest			Mean gain ^a	Positive gain ^b	d^c	
	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>				
Phase I –	69.8	13.9	66	79.6	14.1	66	+ 9.7	91%	0.70	
Phase II –	71.4	14.2	146	79.0	15.0	146	+ 7.6	88%	0.53	
Phase III	Third experimental	71.4	14.6	207	79.5	13.2	206	+ 8.2	89%	0.59
	Second experimental	72.1	14.4	170	77.5	13.1	171	+ 5.6	82%	0.39
	First experimental	74.5	11.5	137	79.9	11.9	136	+ 5.4	81%	0.46
	Control	70.0	15.3	97	75.9	15.4	98	+ 6.6	79%	0.43
National norm reference	69.0	14.5		74.1	14.5		+ 5.1		0.36	

^a Mean gain is the mean of the differences between the scores on the pretest and the posttest; as of some students, there were missing data this is not exactly equal to the difference between the averages on pre-/posttest

^b Positive gain represents the percentage of students that improved from pre- to posttest

^c d is the effect size from pre- to posttest calculated as the mean difference divided by the pooled standard deviation

remained in their original conditions, to check for a possible confounding effect of attrition or level of commitment, yielded comparable results.²

Teachers' implementation and ideas about the feasibility of the CATs

Teachers' use and adaptations of the CATs

The teachers used the CATs in their classes at moments that they considered them to be useful, which on average came down to teachers using every technique at least once or twice. Of the three exemplary CATs, the CAT *Crossing ten and more* seems to be quite straightforward, as became clear from the feedback forms, the teachers generally only made minor adaptations to this CAT, if they made any. They paraphrased the wording or adapted some numbers, sometimes extended the number of questions or changed a context, more often they slightly adapted the setting of the CAT. For example, teacher H (from phase II) interpreted the use of this CAT as a game. He considered it to be “nonsense to be the only one doing the work” and let different students come up with the problems to present to the other students. This adaptation was valuable to this teacher, as it allowed him to not only assess the students giving the responses but also the strategies of the students asking the questions. For the other two CATs, teachers did not adapt the format or content, they used them as described in “[Material: CATs](#)”.

Teachers' (and students') ideas about the usefulness of the CATs

On the feedback forms, all the teachers who mentioned *Crossing ten and more* (18 teachers) wrote that this technique gave them interesting information and that they liked using it. More specifically, they reflected on the insights they gained into students' understanding and the possibility of the use of this whole-classroom response system to help students who are normally less prone to provide answers to become less anxious. For example, teacher M (from phase III) explained that this CAT gave her quick “insight in students' level of automatization³” making her “very enthusiastic.” Another teacher, teacher JR (from phase III) wrote that she got the valuable information that “also high-performing students can have difficulty with quick decisions about whether two numbers cross 100 or 1000”. Moreover, teacher M (from phase III) noticed that “a normally anxious girl could also show her understanding now”.

For the CAT *Easy or difficult*, the teachers explained during the meetings and on the feedback forms (eight teachers on the nine forms about this CAT) that students mostly pointed to the problems in which the numbers did not cross ten and those where the digits 7 and 8 appeared as the difficult ones. Students were very engaged in this activity, in which the assignment in the mathematics lesson asked them to reflect upon what they found easy or difficult, without having to perform calculations.

² Excluding the data from students of teachers that did not attend all workshops and thus switched conditions led to a reduced number of teachers ($N=21$) and students ($N=428$). This resulted in comparable results with the third experimental condition ($M_{adj}=79.8$, $N=206$) outscoring the other conditions (second experimental $M_{adj}=77.9$, $N=60$; first experimental $M_{adj}=78.9$, $N=60$; control $M_{adj}=77.9$, $N=96$). The main effect for condition with these data was, however, not statistically significant ($F(4,421)=1.7$, $p=.162$).

³ Automatization in this context is aimed at making the procedure of, for example, finding the results of (simple) additions automatized, which can be followed by and result in memorization

All teachers agreed that the CAT *Word problem difficulties* provided them with valuable information on their students' strategies when solving word problems. Most teachers expected, as they explained during the meetings, students to struggle more with the word problems than the bare number problems, before performing the experiment. Afterwards, many remarked on the feedback forms (six teachers on the eight forms about this CAT) and in the interviews that students' performance on either type of problem format depended on the student and the type of operation and wording that were used in the problem. For example, teacher L (from phase III) wrote: "Marvelous experiment! [...] Now, I know what strategy every student uses [...] and in which situation." Teacher F (from phase III) wrote: "Students were thinking about how they solved the problems. You see very different ways of thinking, giving you insight in how they solve them. You get to understand how they think."

Teachers' instructional decisions following using the CATs

After using the CAT *Crossing ten and more*, teachers integrated their observations on the students in their further instruction. Overall, teachers related that they found that their high-achieving students would hesitate more on some of these number pairs than they expected, indicating that their automatization was not on the same level as their regular performance in class or on other more standard assessments would indicate. Some teachers would then provide these students with supplementary automatization exercises, like teacher N (from phase III) who wrote on the feedback form that "now I am going to practice more with these children and they will get the multiplication tables to study as homework." As another example, teacher J (from phase I) identified a type of problem with which most students struggled, and she wrote two examples of this on the blackboard to refer to in her further instruction.

When the teachers discussed CAT *Easy or difficult* in the classroom, students explained why and how they decided whether a problem was easy or difficult; many students identified crossing the ten as the main determinant for the difficulty of a problem. Teachers noticed that they could identify differences in approaches through this CAT. For example, teacher M (from phase II) wrote that she "found differences between children: some just randomly pick, others know exactly why they chose." In the subsequent workshop, teacher C (from phase III) really liked this CAT as "many students identified crossing ten but also how they explained their reasoning was very interesting."

Using the results of the *Word problem difficulties* technique, teachers adapted their further instruction to the specific needs of their students. Most teachers, as they explained in the meetings and interviews, reflected with the students upon the different characteristics of how the problem was presented, and of course the similarities; they also let students compare their own work on the different ways of presenting the same problems. In doing this, students were able to not only find out whether they had used different strategies for the different problem formats but also that the only difference between these tasks was that the way of presenting had changed.

Teachers' overall impression on using the CATs

An overall finding was that the teachers liked to use the CATs and that they interpreted them in their own way to adapt them to their practice. Even though teachers operated

diversely in their classrooms and flexibly organized the implementation of different assessment techniques, this adaptive use did not counteract teachers' perceived usefulness of the techniques. For example, teacher L (from phase III) wrote on the feedback form that "[using the CATs] provided [her] with a lot of information [about the students]." Teacher JT (from phase III) said that working with the techniques made him "think about [his] practice: what am I going to do and why am I testing?" Teachers also underscored that they noticed their students becoming more prone to verbalize their thinking process when solving a problem.

A general finding the teachers shared about their students in the interviews, the meetings, and on the feedback forms was that the students referred to the CATs as "mathematics games." They were in the teachers' perception very motivated to participate in, for them, atypical classroom activities such as the CAT *Crossing ten or more* with the red and green cards or the CAT *Easy or difficult*, but also in the CAT *Word problem difficulties*, which contains exercises that are not very different from those they normally have to do. After using the CATs, two teachers explained that their students explicitly mentioned that working on mathematics in this way made it much more fun.

Discussion and conclusion

Discussion

The feasibility of the CATs for the teachers combined with an indication for improvement in students' mathematics achievement are the main results of our study. Teachers found the CATs useful for gaining insights into students' understandings. Additionally, an improvement in students' mathematics achievement was observed. In the experimental third phase of our study, we found that students of teachers that participated in three 1-hour workshops on the development and use of CATs had significantly larger score gains than students from teachers that had participated in fewer workshops. Together, these results provide evidence for the beneficial effects of supporting teachers' use of classroom assessment in mathematics on students' mathematics achievement, more particularly, that teachers' frequent participation in workshops on the development and use of CATs led to larger improvement of their students' mathematics achievement. This confirms the findings of other scholars (e.g., Andersson and Palm 2017; Black and Wiliam 1998; Briggs et al. 2012) that teachers' use of classroom assessment is associated with improved student learning.

The gain scores in the different phases (see Table 3) can be explained by the fact that teachers "had time to try out ideas in their own classroom, bring their experiences back to the community of practice, and collaboratively work to refine their assessment tools and strategies" (Suurtamm and Koch 2014, p. 283). The learning gain was quite large considering that the professional development only took three to five workshops of about 1 hour. Interestingly, the average gain in students' achievement scores of the teachers that participated in three and four workshops was slightly smaller (respectively + 8.2, $d = 0.59$ and + 7.6, $d = 0.53$) than those of students of teachers that had five workshops (+ 9.7, $d = 0.70$). Of course, due to the sample size and design of our study, the beneficial effect of one or two extra workshops is not

unequivocally demonstrated, but it does illustrate that teacher participation in at least three workshops was necessary to observe an influence on students' mathematics achievement. An explanation as to why teachers' participation in three workshops in the experimental phase III, or in even more workshops in phase I and II, and the subsequent use of CATs was associated with increased student achievement is the following. By repeatedly participating in the work during the workshops, doing the mathe-didactical analyses of the relevant mathematical domains and thinking about ways to find out their students' learning in these domains, teachers became more aware of two things: the teaching and learning trajectory and formative assessment strategies. Through this combined awareness, teachers could have become more prone to use the CATs they developed during the workshops in their classroom in an effective way—we have to admit that this chain of reasoning is partly speculation, as we did not explicitly assess teachers' knowledge. These teachers used the CATs to really single out a particular, in their eyes important, understanding or skill of students to look into and use that gathered information to shape their ensuing instructional decisions. By doing this, teachers were becoming more knowledgeable about students' understanding of relevant concepts and, as such, could adapt their teaching in such a way that students received the instruction that they needed to progress in their learning. An explanation for the lack of an effect in the conditions with fewer workshops could be that these teachers lacked the time in between the meetings to implement what they had discussed in the different workshops and the reflective discussions in subsequent meetings.

The development of the CATs provided teachers with ways to first investigate their students' learning of mathematics and then promote this and further engage students in their own learning process. This echoes the good teaching practice, which Ginsburg (2009) voiced when discussing formative assessment in relation to mathematics education. An important aspect of the mathe-didactical analyses underlying the CATs is that they can direct teachers' and, when the CATs are used, students' thinking related to a particular mathematical topic, thus providing valuable and didactically useful information about their students to teachers (Erickson 2007).

The use of the CATs for mathematics was truly in the hands of the teachers as they could freely adapt the CATs to fit their practice. Through the adaptations teachers made, following their development, they could develop ownership of the CATs; they became an integral part of the teachers' educational practice. Openness and adaptability of the implementation of an educational intervention have been found to improve teachers' feelings of ownership of, and involvement in, the intervention (Suuramm and Koch 2014). Teachers' possible ownership of the CATs gives a strong indication for their sustainability: teachers are probably more likely to continue using CATs in the future.

Developing and using the assessment techniques are but the first step; the next step is for teachers to integrate the assessment techniques into their practice and use the gathered information. Upon their first encounter with the CATs, teachers thought them just to be "interesting mathematics activities" but were unsure whether using them would have any effect on their students or their own instruction. As teacher A (from phase III) wrote on a feedback form after using the first techniques: "I don't see how these techniques can help students improve. Only doing these exercises is not very useful I think." However, when the teachers had more experiences with the CATs, they did underscore the fact that after some time,

it would become easier to further integrate the assessment techniques in their daily practice. This sentiment echoes the findings of one of the earliest international studies on teachers' use of classroom assessment in mathematics (Shepard et al. 1996), where teachers also needed time to integrate this new approach into their practice.

Limitations and further research

A first point of contention of the reported effects is what exactly affected students' mathematics achievement. Some could argue that the step from a (small) number of professional development sessions and activities to implement in classrooms to looking for effects on changes in student mathematics achievement over a semester is too big. However, as the teachers were randomly distributed over the conditions and the main difference between the practices of the teachers in the different conditions only concerned their participation in the workshops and subsequent use of the CATs, the differences in performance are most probably due to the intervention. Furthermore, it could have been that it was not teachers' development and use of the CATs and their acting upon the gathered information, but more the content of the CATs and the practice students got in these domains, that influenced student achievement. We cannot fully exclude this possibility but do see three strong arguments against this explanation. First, the teachers generally used the CATs equally often (every CAT once or twice) in the different conditions. Secondly, the practice students got through the exercises in the CATs came in the place of the regular teaching content (that often would touch upon the same topics), so students did not get supplementary practice in the experimental conditions. Thirdly, teachers reported that they often used the information to adapt their instruction based on the information they obtained from the CATs. These three facts support the conclusion that teachers' use of the CATs seems to have led to the improvement of students' mathematics achievement.

Another issue that might have influenced the results is the teachers' voluntary participation in the study. It could be that our sample was not representative of the Dutch population of teachers, and, for example, the teachers in this sample were overly motivated for mathematics education or formative assessment. Also, teachers' prior didactical knowledge or mathematical knowledge for teaching was not investigated; it could have been that this also influenced the effects on students' achievement and teachers' use of the CATs. Furthermore, due to practical considerations, such as teachers changing schools or moving to a different grade level, it was impossible in the current study to investigate long-term effects with a follow-up. It would be interesting to see if these beneficial effects of teachers' use of CATs continue to show in student achievement in the ensuing school years. However, investigating this would require a very large sample of teachers and students with possibilities to control for all kinds of environmental effects in a longitudinal design. In phase III, there were also a number of teachers who participated in fewer workshops than they were supposed to, which resulted in changing experimental condition group. It could be that these teachers were slightly less motivated. This lower level of motivation could have led them to continue to "make little use of assessment formatively to help the learning

process” (Harlen 2005, p. 209) and would help to explain that their students’ improvement was comparable with the usual (i.e., as shown by the national reference norm). However, most occurrences of nonattendance were due to external reasons, such as sick children or an administrative matter to attend to. Thus, the motivational explanation for the smaller effect on students in these conditions does not necessarily hold ground. Also, performing the same analyses with only the data of those teachers that remained in their original conditions actually gave the same results in terms of effect size and directions.

In earlier research, it was often argued that to get teachers to become owners of the assessment techniques and use them as such, it was needed to have sustained programs of professional development (e.g., Andersson and Palm 2017; Black and Wiliam 1998). For further research and theory development, it is important to investigate the specific factors that influenced teachers’ implementation of CATs in their mathematics teaching practice, and the relation with student performance. We did not explicitly focus on the precise actions teachers undertook on the basis of the assessment information they gathered—only their reports of the planned actions were analyzed. Observing and analyzing their actual practice would be a very interesting follow-up study, because it has been found that teachers are better at inferring students’ levels of understanding than at deciding about the next steps in instruction on the basis of assessment information (Heritage et al. 2009).

Conclusion

Bearing the results of this study in mind, a probable chain of events to explain the effects on student achievement is that teachers participating in three (or more) workshops pay more attention to the didactical underpinnings of the use of the techniques. Teachers might have become more positively inclined to use the assessment techniques, and the information gathered by them, in their further teaching due to having more reflective discussions with other teachers about the mathe-didactical analyses of the mathematical content to be taught according to the textbook and the results of using the techniques. This form of teacher learning community (e.g., Lave and Wenger 1991) has also been advocated to use with teachers developing their classroom assessment skills (e.g., Suurtamm and Koch 2014; Wiliam et al. 2004). In its current form, with teachers participating in merely three (or more) meetings on the development and use of CATs, and the effect this shows on students’ mathematics achievement clearly suggest merit in continuing investigation of these techniques and assist teachers in using them. The awareness of students’ mathematics skills and knowledge teachers develop while thinking about the CATs can be of use in mathematics teacher education, for example, in “support[ing] beginners’ work on two crucial elements of mathematics teaching: unpacking mathematics and attending to students thinking” (Sleep and Boerst 2012, p. 1039).

Let us finish this discussion with the reminder that knowing what students know is quintessential in teachers’ practice: for teachers to help their students, “the mental movement must be known before it can be directed” (Dewey 1904, p. 262). The developed CATs for mathematics education clearly helped teachers to get to know their students’ “mental movement” and direct it towards further learning, as evidenced by the improved mathematics achievement of their students.

Appendix

Table 4 Descriptions of the classroom assessment techniques (CATs) for mathematics in grade 3

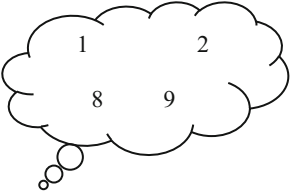
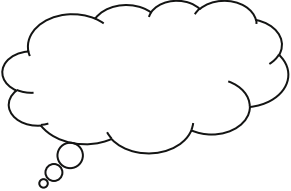
Title	Format	Description	Example	Goal
1 Crossing ten and more	Red and green cards	The teacher calls a series of number pairs and asks: "Do these numbers cross 10/100/1000? Yes or No?" Students use the cards to respond instantly. The green card means Yes and the red card means No.	"7 and 4" *cards* "1 and 8" *cards* ... "70 and 40" *cards* "700 and 400" *cards*	Assess whether students have ready knowledge about whether a total of two numbers crosses 10/100/1000 and whether they understand the analogies.
2 Crossing ten and more	Red and green cards	The teacher asks: "Is the difference bigger than 5/10/50/100?" and calls a series of number pairs.	"9 and 2" *cards* "15 and 7" *cards*	Assess whether students have ready knowledge about whether the difference between two numbers is bigger than 5/10/50/100 and whether they understand the analogies.
3 Crossing ten and more	Red and green cards	The teacher asks: "Is [a series of numbers] a multiple of 4?". This question is asked for various multiplication tables.	"Is 32 in the table of 8?" *cards* "Is 44 in the table of 8?" *cards*	Assess whether students have ready knowledge of the multiplication/division tables.
4 Easy or difficult	Worksheet and (class/group) discussion	On a worksheet with two columns of problems, students have to circle which of two problems is easiest—without calculating the result. When finished, they discuss and explain their reasoning to their neighbor and in class.	Which is easiest to you? "11 + 2 or 13 + 12" "26 - 7 or 35 - 4"	Assess whether students are aware of the difficulties some number operations contain and whether they can reflect on these difficulties.
5 Clouds	Worksheet and (class/group) discussion	On a worksheet on which clouds are printed filled with numbers, students have to connect two or three numbers that add up to 10, 100, or 1000. When finished students exchange their work with their neighbor and discuss differences in approach.		Assess whether students have ready knowledge about what numbers complement each other equaling 10/100/1000 and see the analogy with other numbers.
6 Make your own clouds	Empty worksheet (for own production)	On a worksheet on which empty clouds are printed, students can fill in pairs or another number of numbers that add up		Assess the insight of students in the combinations of numbers equalling

Table 4 (continued)

Title	Format	Description	Example	Goal
		to 10, 100, or 1000, or a different number. Exchange work with peer and discuss differences in approach.		10/100/1000 and creating exercises.
7 Word problem difficulties	Classroom experiment	Students solve a series of problems first in word problem format, then the same problem but as a bare number problem. After class, the teacher compares student work on the different presentational formats.	“Charly saved 680 euro, a computer costs 1000; how much does he still need?” “1000–680 = “	Find out why students have difficulty (or not) with word problems. Do they have difficulties with understanding the text or with doing the calculation? Assess differences between students’ solution strategies for different formats.
8 What could have been the question?	Worksheet (for own production)	Students are presented multiplications above ten, for each multiplication they have to think of a possible question the teacher could ask, for which the problem at hand could give the answer? Students write down possible questions, and answer them, followed by a whole-class discussion.	“6 × 16” Examples of questions: “How much is..?” “How can you calculate..?” “Problems with the same outcome?”	Assess students’ awareness of the (limited) types of questions one can ask about a problem and the type of strategies students come up with related to multiplication.
9 Find your errors and correct them	Worksheet	The teacher has corrected the work of a student and returns the work to the student saying: “Of these 20 questions you made mistakes on 5, find them and correct them.”		By this activity, teachers can assess whether students have insight in the underlying mathematics of their own mistakes. Students have to actively engage with their mistakes and the learning material.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Andersson, C., & 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. *Learning and Instruction, 49*, 96–102.
- Andrade, H. L., & Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*. <https://doi.org/10.1080/0969594X.2019.1571992>.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.
- Beswick, K., Anderson, J., & Hurst, C. (2016). The education and development of practising teachers. In K. Makar, S. Dole, J. Visnovska, M. Goos, A. Bennison, & K. Fry (Eds.), *Research in mathematics education in Australasia 2012–2015* (pp. 329–352). Singapore: Springer Science+Business Media.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education: Principles, Policy & Practice, 5*(1), 7–74.
- Briggs, D. C., Ruiz-Primo, M. A., Furtak, E. M., Shepard, L. A., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice, 31*(4), 13–17.
- Dewey, J. (1904). The relation of theory to practice in education. In C. A. McMurry (Ed.), *Third yearbook of the National Society for the scientific study of education* (pp. 9–30). Chicago: Chicago University Press.
- Erickson, F. (2007). Some thoughts on “proximal” formative assessment of student learning. *Yearbook of the National Society for the Study of Education, 106*, 186–216.
- Frey, B. B., & Schmitt, V. L. (2010). Teachers’ classroom assessment practices. *Middle Grades Research Journal, 5*(3), 107–117.
- Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de Leon, V., Morrison, D., & Heredia, S. C. (2016). Teachers’ formative assessment abilities and their relationship to student learning: findings from a four-year intervention study. *Instructional Science, 44*(3), 267–291.
- Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915–945.
- Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: children’s minds, teachers’ minds. *Human Development, 52*(2), 109–218.
- Harlen, W. (2005). Teachers’ summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal, 16*(2), 207–223.
- Harlen, W. (2007). *Assessment of learning*. London: Sage.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
- Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review, 17*, 50–62.
- 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.
- Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (2016). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education: Principles, Policy & Practice, 23*(3), 353–376.
- Huitema, S., Erich, L., Van Hijum, R., Van de Wetering, M. et al. (2009). *De wereld in getallen – vierde editie* [The world in numbers – fourth edition]. Den Bosch, the Netherlands: Malmberg.
- Janssen, J., Scheltens, F., & Kraemer, J-M. (2006). *Primair onderwijs. Leerling- en onderwijsvolgsysteem. Rekenen-wiskunde groep 5* [primary education. Student and educational monitoring system. Mathematics grade 3]. Arnhem: Cito.
- Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8* [Scientific justification of the mathematics test for Grade 1 until Grade 6]. Arnhem, the Netherlands: Cito.

- Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment: 75 practical strategies for linking assessment, instruction, and learning*. Thousand Oaks: Corwin.
- Kennedy, M. (2016). How does professional development improve teaching. *Review of Educational Research*, 86(4), 945–980.
- Kim, H. J. (2019). Teacher learning opportunities provided by implementing formative assessment lessons: Becoming responsive to student mathematical thinking. *International Journal of Science and Mathematics Education*, 17, 341–361.
- Kingston, N., & Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37.
- Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge.
- Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: minute-by minute and day by day. *Educational Leadership*, 63(3), 18–24.
- Liljedahl, P. (2010). Noticing rapid and profound mathematics teacher change. *Journal of Mathematics Teacher Education*, 13(5), 411–423.
- Lin, P. J. (2006). Conceptualizing teachers' understanding of students' mathematical learning by using assessment tasks. *International Journal of Science and Mathematics Education*, 4(3), 545–580.
- Meelissen, M. R. M., Netten, A., Drent, M., Punter, R. A., Droop, M., & Verhoeven, L. (2012). *PIRLS en TIMSS 2011. Trends in leerprestaties in Lezen, Rekenen en Natuuronderwijs* [PIRLS and TIMSS 2011. Trends in achievement in reading, mathematics and science]. Nijmegen/Enschede, the Netherlands: Radboud University/Twente University.
- National Council of Teachers of Mathematics (2013). Formative assessment: a position of the National Council of Teachers of Mathematics. NCTM. Retrieved from https://www.nctm.org/uploadedFiles/Standards_and_Positions/Position_Statements/Formative%20Assessment1.pdf. Accessed Jul 2013.
- Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 107–123). Thousand Oaks: Sage..
- Phelan, J., Choi, K., Niemi, D. N., Vendlinski, T., Baker, E. L., & Herman, J. (2012). The effects of POWERSOURCE© assessments on middle-school students' math performance. *Assessment in Education: Principles, Policy & Practice*, 19(2), 211–230.
- Randel, B., Aphorip, H., Beesley, A. D., Clark, T. F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *The Journal of Educational Research*, 109(5), 491–502.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14.
- Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice*, 15(3), 7–18.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2017). *Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment*. Paper presented at the NCME Special Conference on Classroom Assessment and Large-Scale Psychometrics: The Twain Shall Meet, Lawrence, KS.
- Shute, V. J. (2007). *Focus on formative feedback*. Princeton: Educational Testing Service.
- Sleep, L., & Boerst, T. A. (2012). Preparing beginning teachers to elicit and interpret students' mathematical thinking. *Teaching and Teacher Education*, 28(7), 1038–1048.
- Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice*, 44(1), 11–18.
- Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: doing it right – using it well*. Portland: Assessment Training Institute..
- Suurtamm, C., & Koch, M. J. (2014). Navigating dilemmas in transforming assessment practices: experiences of mathematics teachers in Ontario, Canada. *Educational Assessment, Evaluation and Accountability*, 26(3), 263–287.
- Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: using action research to explore and modify theory. *British Educational Research Journal*, 27(5), 615–631.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2.
- Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht: CD-β Press/Freudenthal Institute, Utrecht University.
- Van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactical model for assessment design in mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 689–716). Dordrecht: Kluwer Academic Publishers.

- Van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic Mathematics Education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 521–525). Dordrecht: Springer.
- Van den Heuvel-Panhuizen, M., & Teppo, A. (2007). Tasks, teaching sequences, longitudinal trajectories: about micro didactics and macro didactics. In J. H. Woo, H. C. Lew, K. S. Park, & D. Y. Seo (Eds.), *Proceedings of the 31st Conference of the IGPME* (Vol. 1, p. 293). Seoul: PME.
- Van den Heuvel-Panhuizen, M., Middleton, J. A., & Streefland, L. (1995). Student-generated problems: easy and difficult problems on percentage. *For the Learning of Mathematics*, 15(3), 21–27.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). Primary school teachers' assessment practice in mathematics education. *PLoS One*, 9(1), e86817.
- Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen, J., & Eggen, T. J. H. M. (2013). Teachers' use of classroom assessment in primary school mathematics education in the Netherlands. *Cadmo*, 21(2), 35–53.
- Wiliam, D. (2007). Keeping learning on track: classroom assessment and the regulation of learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1053–1098). Greenwich: Information Age Publishing.
- Wiliam, D. (2011). *Embedded formative assessment*. Bloomington: Solution Tree Press.
- Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: 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: Lawrence Erlbaum Associates.
- Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education: Principles, Policy and Practice*, 11(1), 49–65.
- Wilson, M., & Sloane, K. (2000). From principles to practice: an embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208.
- Yin, X., & Buck, G. A. (2019). Using a collaborative action research approach to negotiate an understanding of formative assessment in an era of accountability testing. *Teaching and Teacher Education*, 80, 27–38.
- Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Tomita, M., Furtak, E. M., Brandon, P. R., & Young, D. B. (2008). On the measurement and impact of formative assessment on students' motivation, achievement, and conceptual change. *Applied Measurement in Education*, 21(4), 335–359.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Michiel Veldhuis^{1,2} · Marja van den Heuvel-Panhuizen^{1,3}

¹ Freudenthal Group, Department of Pedagogical and Educational Sciences, Faculty of Social and Behavioural Sciences & Freudenthal Institute, Faculty of Science, Utrecht University, PO Box 80140, 3508 Utrecht, TC, Netherlands

² iPabo University of Applied Sciences, Amsterdam, Netherlands

³ Nord University, Bodø, Norway