

**Improving classroom assessment  
in primary mathematics education**

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

This research was carried out in the context of the Dutch Interuniversity Centre for Educational Research.

**ico**

Veldhuis, Michiel

Improving classroom assessment in primary mathematics education / M. Veldhuis – Utrecht: Freudenthal Institute for Science and Mathematics Education, Faculty of Science, Utrecht University / FIsme Scientific Library (formerly published as CD-β Scientific Library) no 90, 2015

Dissertation Utrecht University. With references. Met een samenvatting in het Nederlands.

ISBN: 978-90-70786-32-8

Keywords: mathematics education / classroom assessment / primary school teachers / assessment techniques / student achievement / professional development

Cover design: Vormgeving Faculteit Bètawetenschappen

Cover illustration: Wouter & Michiel Veldhuis

Printed by: Drukkerij Wilco, Amersfoort

© 2015 Michiel Veldhuis, Utrecht, the Netherlands.

**IMPROVING CLASSROOM ASSESSMENT  
IN PRIMARY MATHEMATICS EDUCATION**

**VERBETERING VAN HET TOETSGEBRUIK  
BIJ REKENEN-WISKUNDE IN HET BASISONDERWIJS**

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van  
de rector magnificus, prof. dr. G. J. van der Zwaan, ingevolge het besluit van  
het college voor promoties in het openbaar te verdedigen  
op woensdag 24 juni 2015 des middags te 2.30 uur

door  
Michiel Veldhuis

geboren op 17 februari 1986  
te Amsterdam

Promotor: Prof. dr. M. H. A. M. van den Heuvel-Panhuizen

## Table of contents

<b>Chapter 1</b>	<b>Introduction</b>	9
<b>Chapter 2</b>	<b>Teachers' use of classroom assessment in primary school mathematics education in the Netherlands</b> <i>M. Veldhuis, M. van den Heuvel-Panhuizen, J. Vermeulen &amp; T. Eggen</i> <i>Published in CADMO, 2013, 21(2), 23-53</i>	25
<b>Chapter 3</b>	<b>Exploring teachers' use of classroom assessment techniques in primary mathematics education</b> <i>M. Veldhuis &amp; M. van den Heuvel-Panhuizen</i> <i>Submitted</i>	49
<b>Chapter 4</b>	<b>Supporting primary school teachers to improve their assessment practice in mathematics: Effects on students' learning</b> <i>M. Veldhuis &amp; M. van den Heuvel-Panhuizen</i> <i>Submitted</i>	75
<b>Chapter 5</b>	<b>Primary school teachers' assessment profiles in mathematics education</b> <i>M. Veldhuis &amp; M. van den Heuvel-Panhuizen</i> <i>Published in PLOS ONE, 2014, 9(1), e86817</i>	111
<b>Chapter 6</b>	<b>Summary and discussion</b>	147
	<b>Samenvatting (summary in Dutch)</b>	165
	<b>Dankwoord</b>	175
	<b>Curriculum Vitae</b>	178
	<b>List of publications related to this thesis</b>	179
	<b>List of presentations related to this thesis</b>	180
	<b>FIsmE Scientific Library</b>	182
	<b>ICO Dissertation Series</b>	187



*pour Clémentine & Théo*



# **Chapter 1**

## **Introduction**

## Introduction

### 1. Teaching needs assessment

A primary school teacher is preparing her mathematics lesson and reads in the teacher guide of her mathematics textbook that today she has to introduce a new problem to her students, namely, how to perform calculations like  $77-29$ . Now, this does not come as a surprise to her, because she is very well aware of the teaching and learning trajectory of primary mathematics and because of this knowledge she is also aware that for her students to be able to perform such calculations they need to have acquired a number of skills and a particular level of understanding beforehand. These prerequisites are among others the basic skills of being able to make jumps of ten (e.g., on the number line, from 77 to 67 to 57...) and jumping to a multiple of ten (e.g., from 77 to 70 to 60...). Essentially these skills imply that the students are able to structure the calculation instead of using a counting strategy (cf. Treffers, Van den Heuvel-Panhuizen, & Buys, 1999; Van den Heuvel-Panhuizen, 2008). Furthermore, if students are already familiar with calculating by a compensation strategy (e.g.,  $77-30+1$ ...) then it might be worthwhile to build onto these skills as well.

To be able to build onto these skills and fit her classroom instruction to what her students already know, the teacher needs to have knowledge about whether students indeed have reached any of these prerequisites. Thus, before moving on with instruction she is confronted with the task of collecting information about her students' skills and understanding. For collecting this information, she can use a multitude of approaches. The teacher can use for this almost anything that teachers usually have in their repertoire of classroom activities, such as asking a (series of) question(s), giving (parts of) a standardized test or teacher-designed assignments, observing students working, inspecting student work, or letting students design their own exercises, and so on. All these activities through which a teacher can gather information on students' knowledge and skills are called *assessment* activities. These activities are a sine qua non for teaching, because by means of assessment teachers can "meet students where they are" (Schoenfeld, 2014, p. 407). This 'meeting of students' also implies that the assessment should not only reveal what students do not know and cannot do, but also what students know and can do (cf. Pellegrino, Chudowsky, & Glaser, 2001). In this way, points of action for teaching are provided.

The aforementioned broad interpretation of assessment can also be found in the work of Popham (2000) who considers assessment as the process “by which educators use students’ responses to specially created or naturally occurring stimuli in order to make inferences about students’ knowledge, skills, or affective status” (p. 3). However, what inferences are made on the basis of students’ responses might differ depending on the purpose the teacher has in mind with the assessment. If the main purpose of the assessment is for a teacher to use the assessment results to get access to students’ skills and understanding in an effort to improve further instruction and move students’ learning forward, then we call this assessment *classroom assessment* (cf. De Lange, 1999), as this term clearly locates the practice of this type of assessment in the classroom, initiated by the teacher in interaction with the students.

## **2. Classroom assessment**

Classroom assessment has been given many different names such as ‘informal assessment’, ‘instructionally embedded assessment’, ‘didactical assessment’, ‘assessment for learning’, or ‘formative assessment’ (as, for example, described in Van den Heuvel-Panhuizen & Becker, 2003, p. 698). What all these names have in common is that they refer to an assessment that enables the teacher to make informed decisions about further instruction and consequently leads to instruction that adequately fits to the possibilities and the needs of the students. Evidently, if the teacher adapts her instruction to the students’ needs, as identified through classroom assessment, the students will be more apt to learn; an insight already voiced by the philosopher John Locke (1693) in the 17th century: “[w]e must observe children carefully for ‘favorable seasons of aptitude and inclination’ and teach the child when he is in tune” (p. 53). In addition to the adapted instruction the teacher can also use the information gathered from the classroom assessment to provide worthwhile feedback to the students. This feedback is focused on the task at hand and on helping the student move forward towards the learning goals (cf. Hattie & Timperley, 2007). Teachers using classroom assessment can regularly provide explicit and implicit feedback to their students on the basis of the results of classroom assessment. On top of that, students also receive immediate feedback from their own participation, and that of their peers, in the classroom assessment activities. Moreover, feedback that students receive in this fashion has been linked to increased motivation for learning (e.g., Dweck, 1986).

In fact, classroom assessment is a continuous process. After –or even while– the teacher is providing adapted instruction or feedback to students, the sequence starts all over again; because for every new decision about instruction the teacher needs to have precise information about students’ current skills and understanding to be able to continue.

Taking into account this process of informed instructional decision making and giving feedback to students it stands to reason that classroom assessment could be associated with improved students’ learning. Indeed, many studies found evidence that classroom assessment, known under its different monikers, is an effective way to improve students’ achievement (see among many others, Black & Wiliam, 1998; Crooks, 1988; Scheerens & Bosker, 1997; Wiliam, Lee, Harrison, & Black, 2004). Certainly, critique has been given on its definition (e.g., Bennett, 2011), and size of the positive effects on student learning (e.g., Briggs, Ruiz-Primo, Furtak, Shepard & Yin, 2012; Kingston & Nash, 2011; McMillan, Venable, & Varier, 2013), but all who criticized these studies agree that teachers’ use of classroom assessment is related with increased student learning. Policymakers have also embraced its use, such as, for example, the US National Council of Teachers of Mathematics (NCTM, 2013): “The NCTM strongly endorses the integration of formative assessment strategies into daily instruction” (p. 1).

### **3. Other approaches to assessment**

Classroom assessment, as described here, plays a central and important role in education and is completely intertwined and embedded in teachers’ teaching practice. Teachers can use a whole range of activities assessing their students in their daily classroom practice that are simply part of their teaching practice. However, not all teachers are aware that these practices can be used with a specific assessment focus. Often assessment is only associated with the use of (standardized) assessment instruments, such as externally developed tests, which results in the term assessment for many involved in education not having a positive connotation (as evidenced in many weblogs, newspaper articles, and the like). Due to the recent strong emphasis on test results many educators, teachers, students, and their parents are voicing concerns about the great number and kind of tests students encounter during their educational career. There have even been voices calling for students and teachers to opt out of tests

completely (cf., Heitin, 2015). The problem with movements like these is that there is a risk that teachers turn away from assessment in general, which is not a development that will make education necessarily stronger or student achievement better. Therefore, it is very important to make a clear distinction between the different kinds of assessment. Assessment is neither good nor bad; this completely depends on the type of purpose for which a particular assessment is used.

In general, according to their function, three types of assessment can be distinguished: assessment with an evaluative purpose, with a summative purpose, and with a formative purpose (cf., Wiliam & Thompson, 2008). Assessment used for evaluative purposes, is often called evaluation and is mainly used by administrators to evaluate the quality of teachers, schools, districts, or even countries. This type of assessment is called evaluative, as it truly is an evaluation of educational outcomes of a particular curriculum, program, or policy. This evaluative use of assessment results mainly differs from the summative use of assessment information through its higher aggregation level. Summative assessment is the type of assessment that is aimed at making decisions about a qualification, selection, or certification of students (e.g., Bloom, 1969; Scriven, 1967). These assessments are mainly used at the end of a school year or semester, to test whether students master the skills learned in that schooling period, in order to decide whether they can continue to the next grade, for example. More often than not these tests are graded with reference to a relative norm, whether the norm is some kind of external criterion of minimum level of mastery, or a reference group such as the other students in class. Students receive a relative position in the form of a grade on most of these tests. When policymakers, teachers, parents, or students talk about assessment or tests, the tests in assessment used for summative purposes are those that they are mostly referring to. These are the high-stakes tests included in many educational systems. Such summative or evaluative use of tests or assessment instruments has put assessment in a bad light with stakeholders. Often students' results on tests used in these assessments are not used to help the student forward in their learning process, but only to judge teacher or school quality, in the case of evaluative assessment purposes, or, with a summative purpose, to decide whether a student can receive a passing grade or relative ranking.

Classroom assessment that has by definition a formative purpose, as described earlier, clearly differs from these summative or evaluative practices. One way of looking at this difference is by the simple temporary perspective of the purpose, as Wiliam (2000) describes:

Summative assessments are best thought of as retrospective. The vast majority of summative assessments in education are assessments of what the individual has learnt, knows, understands, and can do. In contrast, formative [classroom] assessment can be thought of as being prospective. (p. 14)

This prospective quality of assessment with a formative purpose is also clearly present in the following characterization: “Formative assessment is designed to extend and encourage learning; summative assessment is used to determine how much students have learned, with little or no emphasis on using results to improve learning” (McMillan, 2007, p. 7). In short, the main difference is that assessment with a formative purpose is focused on helping students’ learning, whereas assessment with a summative purpose is aimed at judging students.

#### **4. Focus of this PhD thesis**

The picture of a teacher’s classroom assessment practice in a primary mathematics classroom sketched in the beginning of this chapter was a hypothetical picture giving a first impression of what this PhD thesis is about. The goal of the research project that formed the basis of this thesis was to go beyond such a hypothetical picture and to provide insight into the actual classroom assessment practice of Dutch primary school teachers in mathematics. Taking into account the previously described effectiveness of teachers’ use of classroom assessment for improving students’ mathematics achievement studying the classroom assessment practice of teachers may provide valuable directions for increasing the quality of mathematics education in primary school.

Studying the classroom assessment practice of teachers means that the focus of the study is much broader than the use of tests for summative or evaluative purposes. In this study classroom assessment is understood as completely intertwined and embedded in teachers’ teaching practice. For teachers to teach students in a good way, they have to know what students can and cannot do. The spotlight of this thesis lies on precisely this part of teachers’ teaching practice: the moment in the instructional cycle where teachers think about what

approach to choose that reveals (assesses) their students' skills and understanding best. Our endeavor was to investigate what assessment practices teachers currently use and propose new approaches. These approaches are based on activities that are close to teachers' teaching practices, but are now used with a specific assessment purpose and as such enhance teachers' assessment repertoire and their awareness of the possibilities assessment provides them.

For classroom assessment to fit in the teachers' mathematics teaching practice, it needs to be epistemologically consistent with the didactics of mathematics education (Van den Heuvel-Panhuizen & Becker, 2003) and should thus correspond to the approach to education as reflected in the adhered learning theory and the curriculum that is taught (Shepard, 2000). The didactical approach primary teachers generally follow in mathematics education in the Netherlands is (more or less loosely) based on the ideas of Realistic Mathematics Education (RME) (e.g., Freudenthal, 1991; Van den Heuvel-Panhuizen, 2000; Van den Heuvel-Panhuizen & Drijvers, 2014). This is, for example, evidenced in the RME-oriented mathematics textbooks and accompanying teacher guides on which primary school teachers rely quite loyally for most of their lessons (Scheltens, Hemker, & Vermeulen, 2013). The main principles of RME are, as follows, described by Van den Heuvel-Panhuizen (2000):

[M]athematics must be connected to reality, stay close to children's experience and be relevant to society, in order to be of human value. Instead of seeing mathematics as a subject to be transmitted, Freudenthal stressed the idea of mathematics as a human activity. Mathematics lessons should give students the 'guided' opportunity to 're-invent' mathematics by doing it. (p. 3)

Following these principles, classroom assessment in the context of realistic mathematics education should provide "a rich environment in which students have an opportunity to demonstrate what they know" (van den Heuvel-Panhuizen & Becker, 2003, p. 712) and in this way support teachers in providing students well-adapted instruction.

For the purpose of this PhD research project several classroom assessment techniques that fitted to Dutch primary school teachers' didactical approaches and practices in mathematics were developed. These techniques were inspired by existing techniques that have been used in previous research (e.g., Dekker & Feijs, 2006; Wiliam, Lee, Harrison, & Black, 2004), described in practice-

## Chapter 1

oriented work (e.g., Keeley & Tobey, 2011; Wiliam, 2011a), and theory about assessment in RME (e.g., De Lange, 1999; Van den Heuvel-Panhuizen, 1996). These assessment techniques were designed to fit closely to the learning and teaching trajectory of primary school mathematics (Van den Heuvel-Panhuizen, 2008) and as such provide teachers with information about key understandings of their students so that their learning could be moved forward.

The main research question of this PhD thesis was:

*Does supporting teachers in improving their classroom assessment practice in mathematics contribute to students' mathematics achievement?*

## 5. Structure of this PhD thesis

This PhD thesis comprises a number of journal articles formatted as chapters, each focusing on a different part of this research project on the improvement of classroom assessment in primary mathematics education. Table 1 illustrates the structure of this thesis and provides the title and research question for every chapter.

Table 1  
*Structure of this Thesis*

	Title	Research question
Chapter 1	Introduction	
Chapter 2	Teachers' use of classroom assessment in primary school mathematics education in the Netherlands	What is the current assessment practice of primary school teachers in mathematics in the Netherlands?
Chapter 3	Exploring teachers' use of classroom assessment techniques in primary mathematics education	How do primary teachers take up classroom assessment techniques in mathematics?
Chapter 4	Supporting primary school teachers to improve their assessment practice in mathematics: Effects on students' learning	What is the effect on student learning of supporting primary teachers' use of classroom assessment techniques in mathematics?
Chapter 5	Primary school teachers' assessment profiles in mathematics education	Can teachers' assessment practice in primary school mathematics education be described by means of assessment profiles?
Chapter 6	Summary and discussion	

As said earlier assessment is a sine qua non for teaching. Teachers have to assess their students otherwise they cannot make instructional decisions about them. However, what is less clear is how they really assess. In order to live up to the promising verb in the title of this PhD thesis ('improving'), teachers' current assessment practice in primary mathematics education needed to be established. The results of this investigation are provided in *Chapter 2*. The establishment of teachers' assessment practice for mathematics was long overdue, as the last time the 'current' situation of classroom assessment practice in mathematics in primary education in the Netherlands was investigated on a large scale was over 25 years ago (Janssens, 1986). Therefore a large-scale survey with an online questionnaire on teachers' assessment practice in primary school mathematics was set up. The online questionnaire consisted of four different parts, with scales on teachers' background characteristics, their mathematics teaching practice, their assessment practice (methods and purposes), and their beliefs on assessment. In this way the following sub-questions were answered: (i) Which assessment instruments and techniques do teachers use to gain insight about the mathematical skills of their students? (ii) How, how often, and why do they collect this information? (iii) What is the relationship between the teachers' ideas about assessment, mathematics, the teaching of mathematics, and their assessment practice?

As a second step, after having identified what teachers reportedly were doing – and not doing – the focus was on developing an intervention focused on changing teachers' assessment practice, as described in *Chapter 3*. Paul Black (1990), one of the main researchers of classroom assessment in the UK, wrote:

A teacher who can record a pupil's performance over time and in several contexts, and who can discuss idiosyncratic answers in order to understand the thinking that might lie behind them can build up a record of far better reliability than any external test can achieve. However, in order to do this, teachers need help from substantial programmes aimed to support teacher assessment with resources of questions, procedures, and in-service training. (p. 25)

Therefore in our effort to investigate classroom assessment, the intervention was based on experiences on professional development on formative assessment in teacher learning communities (e.g., Suurtamm & Koch, 2014; Wiliam, 2011b), since "teachers' capacity to implement AfL [i.e. classroom assessment] can be enhanced by opportunities to work together planning, trying out and

evaluating new ideas“ (Swaffield, 2011, p. 438). Also theory on assessment in realistic mathematics education (e.g., De Lange, 1999; Van den Heuvel-Panhuizen, 1996) and ideas on formative assessment in mathematics (e.g., Hodgen & Van den Heuvel-Panhuizen, 2013) were used. The two small-scale studies described in this chapter had as main purpose the investigation of the feasibility of teachers’ implementation of classroom assessment techniques for mathematics in their practice. These assessment techniques were among others inspired by Dylan Wiliam and his colleagues’ work (cf., Leahy, Lyon, Thompson, & Wiliam, 2005; Wiliam, 2011a). In addition to the feasibility of the classroom assessment a first exploration of the effectiveness of teachers’ use of the classroom assessment techniques was also made.

Then, when the feasibility and a first indication for the effectiveness of these assessment techniques had been established, the effect on student achievement of supporting teachers in their use of the classroom assessment techniques was experimentally investigated. In particular the differential effect on student learning of an experimentally varied number of workshops on the use of classroom assessment techniques was our focus. In *Chapter 4* the results of this quasi-experimental study are discussed.

Furthermore, in *Chapter 5*, the results of a secondary analysis of the questionnaire data from *Chapter 2* are reported. This analysis was focused on the identification of profiles of teachers’ assessment practice. The rationale for distinguishing assessment profiles of teachers is that these can contribute to our theoretical understanding of assessment as teachers carry it out. In addition, knowledge about these assessment profiles could help us in a practical sense with designing tailor-made courses for professional development that fit the teachers’ needs. The characterizations of assessment practice and the identification of the four assessment profiles that emerged from the analysis are described as well as what they can be used for.

Finally, in *Chapter 6* the findings from the four studies are summarized and connected to each other. Furthermore, implications of the usefulness of classroom assessment techniques for the educational practice of mathematics in primary school are discussed. The chapter is concluded by providing suggestions for further research into classroom assessment.

## References

- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5-25.
- Black, P. J. (1990). APU science: The past and the future. *School Science Review*, 72(258), 13–28.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
- Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Taylor (Ed.), *Educational evaluation: New roles, new means: the 68<sup>th</sup> yearbook of the National Society for the Study of Education (part II)* (pp. 26-50). Chicago, IL: University of Chicago Press.
- 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.
- Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438–481.
- Dekker, T., & Feijs, E. (2006). Scaling up strategies for change: change in formative assessment practices. *Assessment in Education: Principles, Policy & Practice*, 17(3), 237-254.
- De Lange, J. (1999). *Framework for classroom assessment in mathematics*. Madison, WI: NICLA/WCER.
- Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist*, 41(10), 1040-1048.
- Freudenthal, H. (1991). *Revisiting Mathematics Education. China Lectures*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.
- Heitin, L. (2015, January 27). Activists share strategies for 'opting out' of tests. *Education Week*. Retrieved from <http://www.edweek.org/>
- Hodgen, J., & Van den Heuvel-Panhuizen, M. (2013). Improving assessment in school mathematics. In P. Andrews & T. Rowland (Eds.), *MasterClass in Mathematics Education: International Perspectives on Teaching and Learning* (pp. 27-38). London: Bloomsbury.

- Janssens, F. (1986). *De evaluatiepraktijken van leerkrachten. Een beschrijvend onderzoek naar het evalueren tijdens het rekenen in het primair onderwijs* [The evaluation practices of teachers. Descriptive research into the evaluation of mathematics in primary education]. Arnhem: Cito.
- Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment: 75 practical strategies for linking assessment, instruction, and learning*. Thousand Oaks, CA: Corwin.
- Kingston, N., & Nash, B. (2011). Formative assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, 30(4), 28-37.
- Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom Assessment: Minute by minute, day by day. *Educational Leadership*, 63(3), 19-24.
- Locke, J. (1693). *Some thoughts concerning education*. London: Churchill.
- McMillan, J. H. (2007). *Assessment essentials for standards-based education*. Thousand Oaks, CA: Corwin Press.
- 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). Retrieved from <http://pareonline.net/getvn.asp?v=18&n=2>
- National Council of Teachers of Mathematics (2013). *Formative assessment: a position of the National Council of Teachers of Mathematics*. NCTM. Retrieved from <http://www.nctm.org/about/content.aspx?id=37990>
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. W. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- Popham, W.J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. Needham, MA: Allyn and Bacon.
- Scheerens, J. & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Elsevier.
- Scheltens, F., Hemker, B., & Vermeulen, J. (2013). *Balans van het rekenwiskundeonderwijs aan het einde van de basisschool 5* [Audit of mathematics education end of primary school]. Arnhem, the Netherlands: Cito.
- Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? A story of research and practice, productively intertwined. *Educational Researcher*, 43(8), 404-412.

- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné and M. Scriven (Eds.), *AERA Monograph Series on Curriculum Evaluation Vol. 1 - Perspectives on Curriculum Evaluation* (pp. 39-83). Chicago: Rand McNally.
- Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.
- 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.
- Swaffield, S. (2011) Getting to the heart of authentic Assessment for Learning. *Assessment in Education: Principles, Policy & Practice*, 18(4), 433-449.
- Treffers, A., Van den Heuvel-Panhuizen, M., & Buys, K. (Eds.) (1999). *Jonge kinderen leren rekenen. Tussendoelen annex leerlijnen, hele getallen onderbouw basisschool*. Groningen: Wolters-Noordhoff.
- Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht, the Netherlands: CD-β Press / Freudenthal Institute, Utrecht University.
- Van den Heuvel-Panhuizen, M. (2000). *Mathematics education in the Netherlands: A guided tour. Freudenthal Institute Cd-rom for ICME9*. Utrecht: Utrecht University.
- Van den Heuvel-Panhuizen, M. (Ed.) (2008). *Children Learn Mathematics: A Learning-Teaching Trajectory with Intermediate Attainment Targets for Calculation with Whole Numbers in Primary School*. Rotterdam: Sense Publishers.
- 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*. London: Springer.
- Wiliam, D. (2000). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.), *Assessment and the Curriculum* (pp. 165–181). Greenwich, CT: JAI Press.

- Wiliam, D. (2011a). *Embedded Formative Assessment*. Bloomington, IN: Solution Tree.
- Wiliam, D. (2011b). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3-14.
- Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice*, 11(1), 49-65.
- Wiliam, D., & Thompson, M. (2008). 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, NJ: Lawrence Erlbaum Associates.



## Chapter 2

### **Teachers' use of classroom assessment in primary school mathematics education in the Netherlands**

Michiel Veldhuis<sup>a</sup>, Marja van den Heuvel-Panhuizen<sup>a,b</sup>, Jorine A. Vermeulen<sup>c</sup>, & Theo J. H. M. Eggen<sup>c</sup>

*<sup>a</sup>Freudenthal Institute for Science and Mathematics Education, Utrecht University, <sup>b</sup>Department of Pedagogical and Educational Sciences, Utrecht University, <sup>c</sup>Cito/University of Twente*

Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen, J. A., Eggen, T. J. H. M. (2013). Teachers' use of classroom assessment in primary school mathematics education in the Netherlands. *CADMO*, 21(2), 35-53.

## **Teachers' use of classroom assessment in primary school mathematics education in the Netherlands**

### **Abstract**

This paper reports on a survey of the classroom assessment practices of Dutch primary school teachers in mathematics education. We investigated, using an online questionnaire, how teachers collect information on their students' progress and how teachers' assessment methods, purposes, and beliefs about the usefulness of assessment are related. In total 960 teachers at 557 Dutch primary schools responded to the questionnaire. Observation-based assessment methods of questioning, observing, and correcting written work, were the most frequently – that is weekly – applied methods, whereas instrument-based methods, particularly using textbook tests and student monitoring tests were employed several times a year. Teachers used assessment mainly for formative purposes and they considered the assessment methods they used themselves as most relevant.

*Keywords:* Classroom assessment; primary school; mathematics education; survey study.

## 1. Introduction

An important challenge teachers face is to create a learning environment in their classroom in which students can develop skills and conceptual understanding. To establish such an environment it is essential that teachers have a good understanding of their students' current stage of development: What skills and what level of comprehension do they have? Without this knowledge teachers' teaching might be out of sync with their students' learning progress. To gather this indispensable information, teachers must assess their students regularly. This assessment by teachers is often called "classroom assessment" (e.g., Airasian, 1991; Black & Wiliam, 1998; Brookhart, 1997; Shepard, 2000). It includes "the kind of assessment that can be used as part of instruction to support and enhance learning" (Shepard, 2000, p. 4).

International studies have shown that teachers use a wide range of methods to collect information about their students' learning (e.g. Mavrommatis, 1997; Remesal, 2011; Suurtamm, Koch, & Arden, 2010). To find out students' skills and comprehension level, teachers can use methods ranging from standardized tests and tests that come with a textbook, to asking questions and observing students while they are working.

The assessment method teachers choose to reveal their students' learning processes depends on several factors. A first factor that has been found to affect this choice is teachers' beliefs concerning classroom assessment (Dixon, Hawe, & Parr, 2011). In particular, Brown (2004) found teachers' perceived usefulness of assessment was linked to the assessment they reported using. These beliefs may be specific for mathematics education, because not all types of knowledge and skills are equally important to assess (Anderson, 2003).

A second main factor in choosing a particular assessment method, beside beliefs, concerns the assessment purpose teachers have in mind (Suurtamm et al., 2010). A major distinction in this respect is Scriven's (1967) division in assessment with a summative and assessment with a formative purpose (e.g., Black & Wiliam, 1998). Summative assessment is mostly applied at the end of a teaching period, usually for reasons of certification and accountability. For example, teachers can use summative assessment when they have to compose an end-of-year report to decide whether students can move up to a next grade

level. Formative assessment is meant to obtain information about students' learning process to make informed decisions on how to design the learning environment so that learning can be optimized (e.g., Wiliam, 2011). Another purpose of assessment is diagnostic assessment, this is a special form of formative assessment in which assessment is used to obtain detailed information about individual students' prior knowledge, ways of reasoning, use of strategies, and misconceptions (Crisp, 2012; Keeley & Tobey, 2011).

A further determining factor of using particular assessment methods is the view on education in which the assessment takes place. Assessment by teachers is "an essentially interactive process, in which the teacher can find out whether what has been taught has been learned, and if not, to do something about it" (Wiliam, 2007, p. 1054). This means teachers' assessment is closely interwoven with their teaching and classroom assessment can be considered as an integrated part of instruction. Therefore the methods used for assessment should correspond to the approach to education as reflected in the adhered learning theory and the curriculum that is taught (Shepard, 2000). As a consequence the reform that took place in the last decades in teaching primary school mathematics has also caused a new approach to assessment (Romberg, Zarinnia, & Collis, 1990). The foci in instruction and assessment have shifted from a paradigm where a more or less behaviorist learning theory and a curriculum with emphasis on algorithms, together with standardized testing were paramount, to a new paradigm with a cognitive constructivist learning theory and a reformed curriculum emphasizing reasoning, modeling, and problem solving, and the use of classroom assessment (Romberg et al., 1990; Shepard, 2000).

Although countries have different perspectives in this change in assessment practice, "[c]ommon to all these visions is the notion of assessment as a positive tool for learning and an interconnected part of teaching and learning" (Berry, 2011, p. 89), and that there is a trend of "reducing excessive use of tests and examinations, and using assessment to understand and support learning as well as using student information to improve teaching" (ibid, p. 99). Even so, it is also agreed that the complementary role of assessment by teachers is still underdeveloped and much effort should be put to find the best way to engage and empower teachers to enhance their assessment practice (e.g., Gardner, Harlen, Hayward, & Stobart, 2011).

## 2. Present study

Finding ways to optimize teachers' classroom assessment practice was the underlying rationale for this study through which we – as a start – wanted to gain knowledge of the current state of affairs in primary school mathematics education. Similar to many other countries, in the Netherlands, in the last decades, a significant change has occurred in the approach to mathematics education, known as *Realistic Mathematics Education* (RME). A characteristic feature of RME (e.g., Van den Heuvel-Panhuizen & Drijvers, 2014) is that realistic situations serve as a starting point to instigate the development of mathematical concepts, tools, and procedures and that in a later stage mathematical knowledge gradually becomes more formal and general, and less context-specific. Furthermore, RME adheres to a constructivist approach to learning. Students are given an active role in the learning process, which means, for example, that there is much attention for students' own ways of problem solving.

Although the consequences for assessment of this RME approach have been explored in several studies (e.g., De Lange, 1987; Van den Heuvel-Panhuizen, 1996), not much is known about teachers' classroom assessment practice. The last study into primary school teachers' mathematics assessment practice in the Netherlands was done by Janssens (1986) over 25 years ago. The results from this survey which only included teachers' assessment of students' written work in mathematics, pointed to an almost exclusive use of textbook tests, a lack of teacher knowledge about assessment, and a mostly erroneous use of standardized (psychological) tests. There was no mention of more informal assessment methods such as portfolio use or practical assignments. In addition to Janssens' (1986) study, later investigations were carried out only on the general assessment practice in primary school (e.g., Blok, Otter, & Roeleveld, 2001) or the purposes of assessment in secondary school (e.g., Segers & Tillema, 2011), but there has been no large-scale survey of current assessment practice of primary school teachers in mathematics education.

The present study was set up to close this knowledge gap. Specifically, we aimed to investigate what assessment methods teachers use, and how often they use assessment for summative, formative, and diagnostic purposes. We also considered teachers' beliefs about assessment, as it has been found that these affect the success of educational reform projects (Brown, 2004). Finally, a further

aim of the study was to know more about the relation between teachers' assessment beliefs and their use of assessment methods. In sum, this paper is organized around the following two research questions:

1. How can primary school teachers' mathematics assessment practice be characterized in terms of assessment methods used, frequency, and purposes?
2. How is teachers' use of assessment methods related to the purposes for which they assess, and to their beliefs on assessment?

### **3. Method**

#### **3.1 Participants and procedure**

To answer the research questions, in January 2012 a link to an online questionnaire (see Section 3.2) was successfully sent by e-mail to 5094<sup>1</sup> primary schools for regular education in the Netherlands. School directors were asked in the accompanying e-mail to transfer the link to the teachers of their school. The purpose of the questionnaire was described to be: Finding out how primary teachers gather information on their students' skills and knowledge in mathematics. The teachers who were willing to respond to the questionnaire were promised a set of digital mathematical exercise material as a reward. In February 2012, we sent a renewed request to participate in our study to all schools that did not fill in the questionnaire after the first request. During the two-month period that the questionnaire was online, 1228 (partly) filled in questionnaires were returned. In total 883 teachers filled in all the items of the questionnaires. Of the 267 schools that informed us on their reasons for not participating most (98%) mentioned a lack of time or interest. To avoid bothering the schools we refrained from doing a non-response study. We had contacted schools twice already and on reception of our reminder some schools clearly stated that – because of the large amount of e-mails they receive daily – they did not wish to be contacted again.

---

<sup>1</sup> The Dutch Ministry of Education provided us with a list containing the e-mail addresses (6848 in total) of all primary schools in The Netherlands. However, due to changes in the addresses not all e-mails could be delivered.

The final sample included 960 teachers at 557 different schools, who filled in at least one question about their assessment practice. Of the sample of teachers 83.7% were female, and the mean age was 41.4 years ( $SD = 11.6$ ). Except for the kindergarten grades, all grades were equally represented (approximately 20% for each grade). Only 34.2% of participants were full-time teachers, teaching a class five days a week. On their education we found that the majority (84.2%) had the regular qualification that is required to become a primary school teacher in the Netherlands, i.e. a degree at a teacher education college for primary school teachers (*PABO* in Dutch). Some (10.1%) also had a university bachelor's degree, and some others (9.6%) even had a master's degree. About a quarter of teachers (24.6%) had 1 to 6 years work experience, another quarter (26.5%) had been working for 7 to 13 years, the next quarter (25.0%) involved the teachers who had been working 14 to 25 years, and finally the remaining quarter (23.9%) had 25 years or more experience.

To investigate the representativeness of the sample we compared the background characteristics with available national statistics ("Statistieken ArbeidsMarkt Onderwijs Sectoren", 2010) and the sample characteristics of the National Assessment of Educational Achievement (in Dutch *PPON*; Scheltens, Hemker, & Vermeulen, 2013). Almost all variables, including age, gender, geographical location school, urbanisation level school, textbook use, education, religious denomination school, and amount of working days followed approximately the same distribution as the national statistics. Teachers in our sample – obtained through an open invitation by e-mail – reproduced a number of general characteristics and as such seem to be quite representative of the population of primary school teachers in the Netherlands. Nevertheless, it remains possible that participating teachers were special in other aspects; they could for instance have been positively biased towards assessment in their responses on the questionnaire. The purpose of our survey, however, was neutral and in only asking teachers to inform us anonymously about their assessment practice, this potential positive bias most probably did not have a detrimental influence on the reliability of teachers' responses.

### **3.2 Online teacher questionnaire**

The online questionnaire we developed for this study consisted of 40 questions, subdivided in four parts respectively addressing teachers' (i) background characteristics, (ii) mathematics teaching practice, (iii) assessment practice, and (iv) beliefs on assessment.

#### *3.2.1 Background characteristics*

Teachers' background characteristics were assessed by 13 questions asking, among other things, about their age and prior education.

#### *3.2.2 Mathematics teaching practice*

The part about their mathematics teaching practice contained 15 questions about, for example, the textbook used, the degree of adapting instruction to students' level, and the time dedicated to mathematics.

#### *3.2.3 Assessment practice*

Assessment practice was dealt with in two questions containing lists of possible assessment methods and purposes. The primary school teachers were specifically asked to report on their assessment practices with regard to mathematics education. These methods and purposes were deduced from literature on classroom assessment (Black & Wiliam, 1998; Mavrommatis, 1997; Stiggins & Bridgeford, 1985; Suurtamm, Koch, & Arden, 2010). Examples of assessment methods in the list were portfolios, asking questions, and standardized tests, and examples of purposes included the establishment of learning goals, investigating why students make mistakes, and providing feedback to students about their learning. Communicating how one solves a mathematical problem is seen as very important in the development and acquirement of mathematical insight, and should also be assessed if one wants to improve students' performance (e.g., Van den Heuvel-Panhuizen, 1996). The use of scrap paper is a means for students to communicate their solution strategy. Therefore, the stimulation of scrap paper was included as an assessment purpose. Moreover, teachers were given the opportunity to add assessment methods and purposes that were not listed. Furthermore, teachers were asked to rate on a six-point-scale (1 = *Rarely to never*, 2 = *Yearly*, 3 = *A few times a year*, 4 = *Monthly*, 5 = *Weekly*, 6 = *A few times a week*) how often they used particular methods, and how often they used them for particular purposes.

### 3.2.4 Beliefs on assessment

To measure the teachers' beliefs on the usefulness of classroom assessment we included ten translated and adapted statements from Brown's (2004) Teachers' Conceptions of Assessment (COA-III) questionnaire. Teachers were asked to rate to what extent they agreed with the statements on a four-point-scale (1 = *Completely disagree*, 2 = *Disagree*, 3 = *Agree*, 4 = *Completely agree*). Teachers also had to rate the importance of the assessment of various knowledge domains (e.g., factual knowledge) and types of skills (e.g., application) on a four-point-scale (1 = *Very unimportant*, 2 = *Unimportant*, 3 = *Important*, 4 = *Very important*). Finally, teachers were asked the importance they allotted to the assessment methods used by selecting the four most and least relevant assessment methods.

A pilot study with six primary school teachers was conducted to investigate the clarity of the questions, the suitability of the answering format, and the length of the questionnaire. The feedback from this pilot was used to make the final version of the questionnaire.

### 3.3 Construction of the variables

Before analyzing the data we grouped the assessment methods into observation-based methods and instrument-based methods. In observation-based methods teachers gather data through direct observation of student behavior. In an instrument-based method a physical (or digital) instrument is used –often supplementing teachers' direct observation– to obtain assessment data. Most standardized assessment instruments are not developed by teachers, but by external agencies for standardized testing or textbook publishers. These instruments usually have prescribed purposes as well as a prescribed administration frequency.

The frequency of assessment use was calculated as the sum of the frequencies of every assessment method. The frequency of instrument-based assessment is the sum of the frequencies of the assessment methods in this category (4 items), just as the frequency of the observation-based assessment is the sum of the frequencies of the observation-based methods (7 items). The number of assessment purposes was constructed as the sum of the number of purposes that teachers pursued monthly or more often. Summative, formative, and diagnostic

## Chapter 2

purposes represent the number of summative, formative, or diagnostic purposes that teachers pursued monthly or more often.

The scales that were constructed (see Table 1) showed overall fairly high reliability ( $\alpha > .78$ ), with the exception of the frequency of assessment method use, which was lower, probably because of the above mentioned prescribed frequency for some methods.

Table 1  
*Construction and Internal Consistency of Scale Variables*

Scale	Number of items	Response options	$\alpha$	Example
Frequency of assessment method use	11	Six-point-scale from <i>Rarely to never</i> until <i>A few times a week</i>	.52	“How often do you collect students’ scrap paper?”
Importance of assessing skills and knowledge	10	Four-point-scale from <i>Very unimportant</i> until <i>Very important</i>	.82	“How important is assessing your students’ level of factual knowledge?”
Perceived usefulness of assessment (Shortened COA-III)	10	Four-point-scale from <i>Completely disagree</i> until <i>Completely agree</i>	.80	“Assessment helps me to assess to what extent students understand my lessons?”
Frequency of use of assessment purposes	12	Six-point-scale from <i>Rarely to never</i> until <i>A few times a week</i>	.79	“How often do you use assessment to investigate the reasons for students’ errors?”

## 4. Results

To come to a description of characteristics of teachers' assessment practice in terms of methods used, frequency, and purpose (research question 1), we computed the relative frequencies of each of these aspects. To answer research question 2, we computed Spearman's rank correlations to describe how the use of particular assessment methods, their purposes, and the perceived usefulness are related.

### 4.1 The use of assessment methods for primary school mathematics

Asking questions, correcting written work, and observing students were reported as the most used assessment methods. Collecting scrap paper, letting student give presentations, and keeping portfolios of student work were mentioned as the least used methods. Figure 1 shows the relative frequencies of the use of assessment methods based on teachers' report of how often they used the listed methods. For clarity, in Figure 1, the actual response format in the six-point scale was merged into three main categories (*Yearly* = *Rarely to never* and *Yearly*; *Monthly* = *Several times a year* and *Monthly*; *Weekly* = *Weekly* and *Several times a week*); however, the original response format was used in further analysis.

The teachers' answers revealed that with respect to the observation-based methods, most teachers asked questions, made observations, and corrected their students' written work weekly or more often (from 77% to 91%), whereas some (19%) only observed their students monthly. Letting students present their work and keeping portfolios was rare (~80% yearly), whereas letting students do practical assignments was a monthly activity for over half of the teachers (57%). With respect to instrument-based methods we found that almost all teachers used tests from the textbook and from a student monitoring system monthly (> 85%), as would be expected by the regulations of these tests. However, assessing students with either teacher- (24%) or student-developed problems (22%) was not very common on a weekly basis.

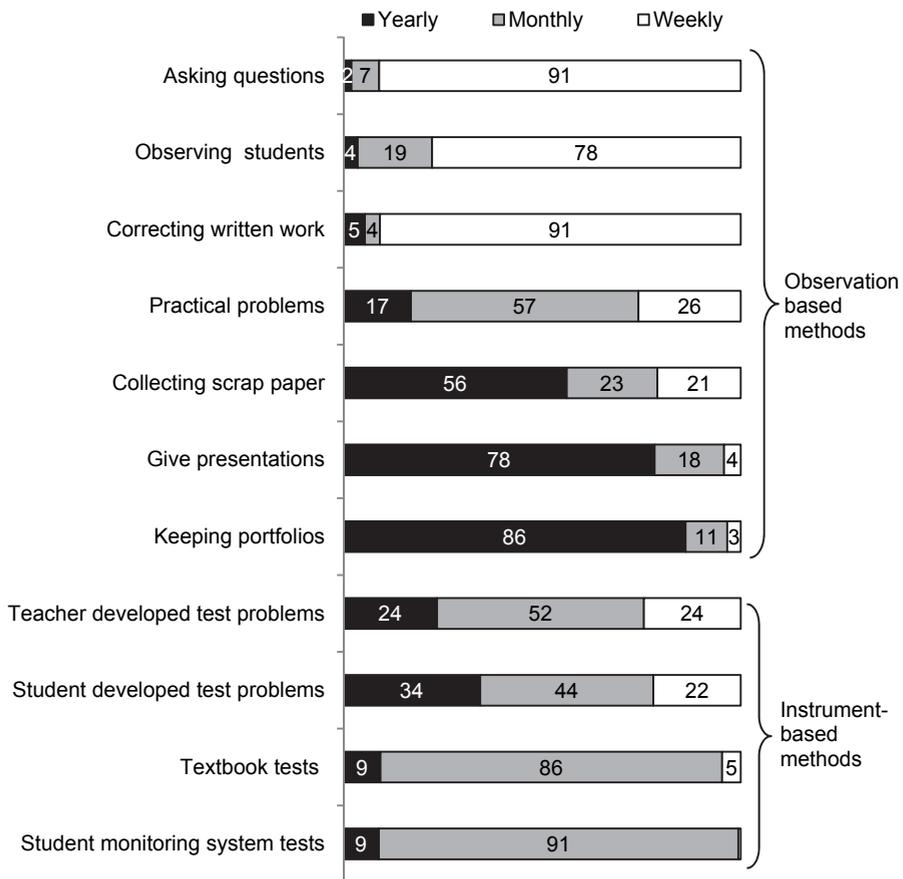
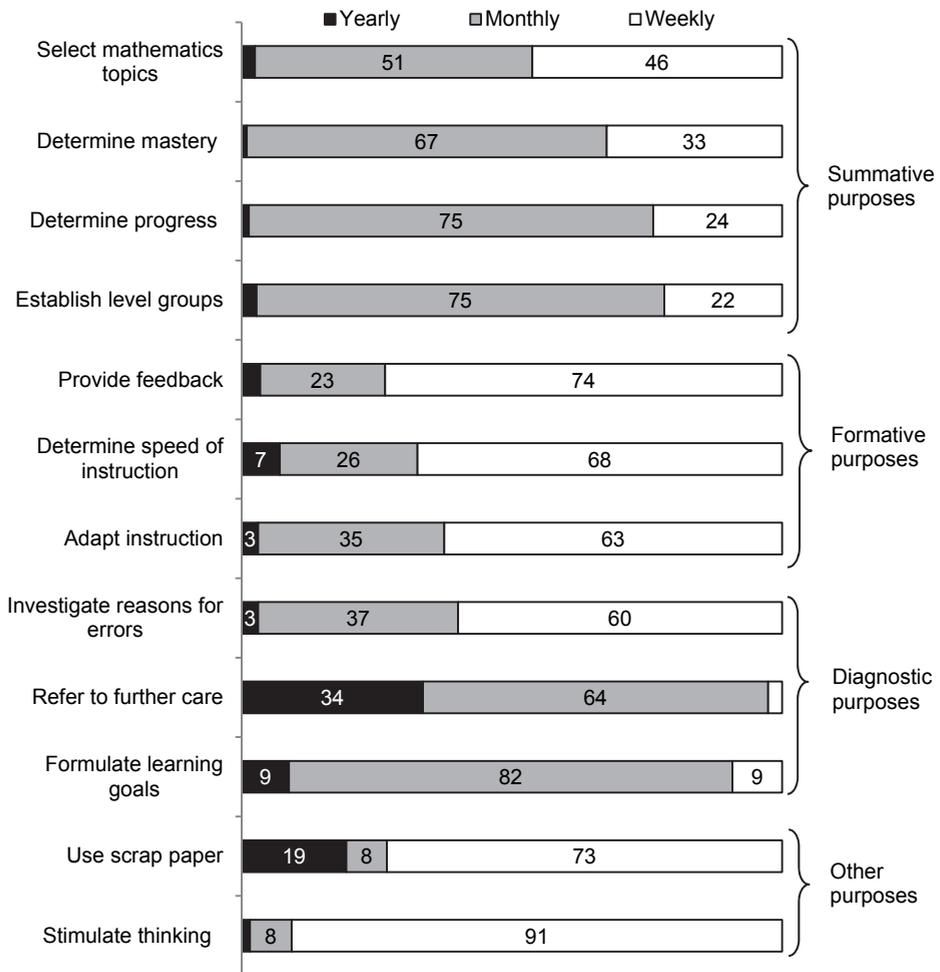


Figure 1. Percentages of frequencies of assessment methods' use ( $ns > 940$ )

#### 4.2 The use of assessment for different purposes

In Figure 2 the percentages of the frequency of using assessment for different purposes are shown for the categories *Formative*, *Diagnostic*, *Summative* and *Other*. As was done for assessment methods, the six different time periods have been merged into three main categories: *Yearly*, *Monthly* and *Weekly*.

## Teachers' current assessment practice



*Figure 2.* Percentages of frequencies of assessment methods' purposes ( $n = 930$ )

The first conclusion is that teachers generally reported using assessment more for formative purposes than for summative purposes, which stands to reason from the nature of these purposes. For summative purposes selecting specific mathematics topics was mentioned most (46% of the teachers reported doing this weekly). The summative purposes, determination of progress and mastery, as well as establishing level groups were considered to be done monthly (> 66% of the teachers reported to do this monthly). For formative purposes, more than 60% of the teachers collected data for giving feedback,

determining instructional speed, and adapting instruction. For diagnostic purposes, 60% of the teachers reported that they investigated reasons for errors weekly. Almost 64% of the teachers reported monthly collection of information to refer students to further care. Within the other purposes, we found that teachers used weekly the stimulation of using scrap paper (73% of the teachers reported to do this weekly) and of thinking (91% of the teachers reported to do this weekly). However, formulating learning goals was less sought as a purpose (82% of teachers reported to do this monthly).

### 4.3 Perceived relevance of assessment

Figure 3 shows the percentages of the assessment methods which were chosen by the teachers as the four most and least relevant methods (out of the 11 methods that were presented).

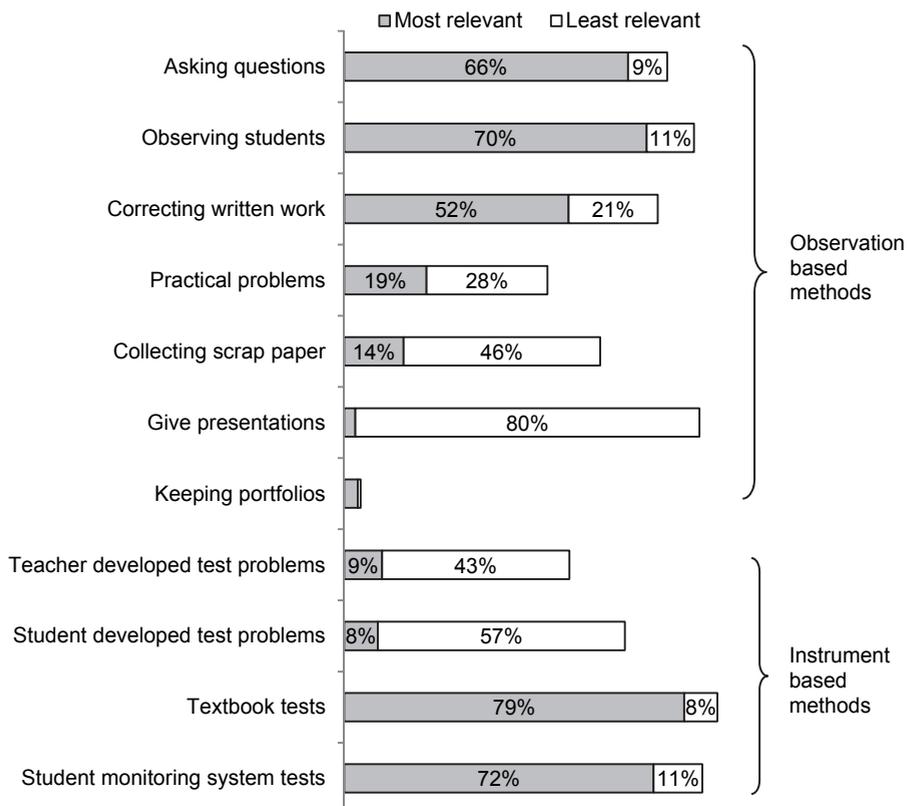


Figure 3. Percentages in which teachers chose assessment methods as the four most relevant and least relevant ( $n = 884$ )

As can be seen, teachers mostly agreed on the four most relevant assessment methods, with more than 65% choosing observation-based methods such as asking questions and observing students and more than 70% choosing the instrument-based methods, textbook and tests from a student monitoring system. The teachers were more divided on what they considered the least relevant methods. Only on giving presentations (80%) and using student-developed test problems (57%) did more than half of the teachers agree on their irrelevance. Collecting scrap paper (46%) and using teacher-developed test problems (43%) were also seen as irrelevant. It is remarkable that keeping portfolios was generally not chosen as either most (3%) or least (1%) relevant assessment method.

#### **4.4 Perceived usefulness of assessment**

Figure 4 shows the responses to the question on the perceived usefulness of assessment. Generally, teachers disagreed with negative statements on assessment, such as that assessment has little influence on teaching, and assessment does not tell what students can do. In contrast, teachers commonly agreed to the positive statements, for example, that assessment is useful, provides information, and improves the classroom climate. The only exception was the positive statement that assessment predicts student performance. Almost one third (30%) of teachers disagreed with this statement. For all other statements, more than 90% of the teachers agreed that assessment was useful for a variety of ends.

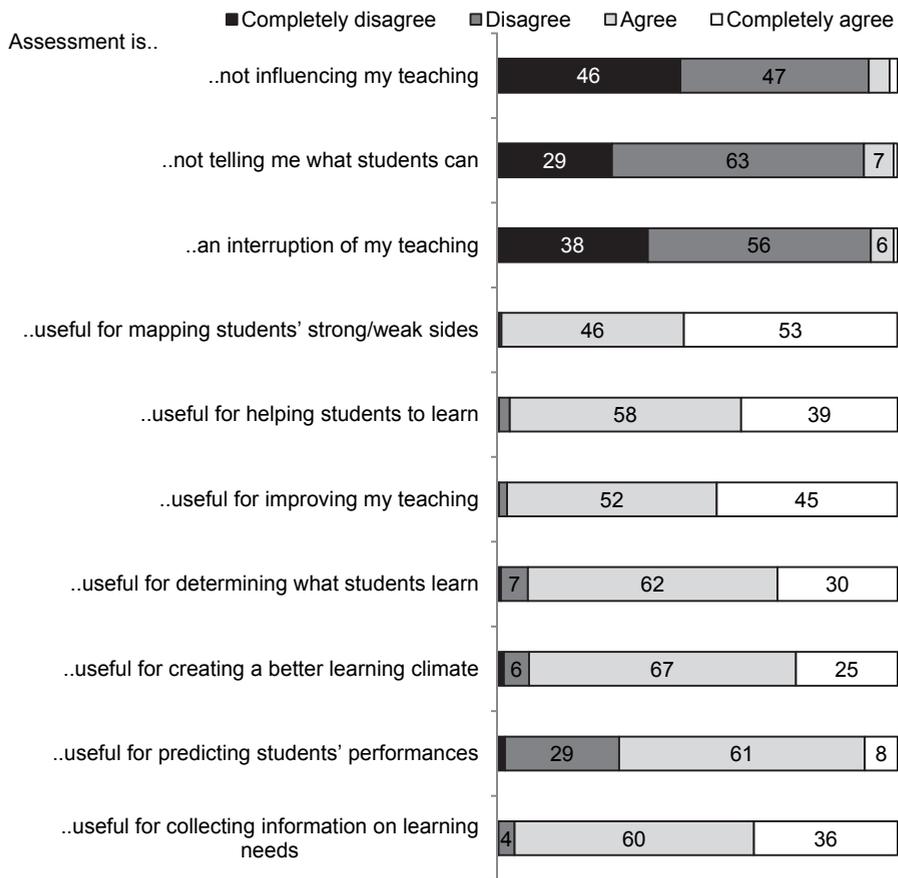


Figure 4. Percentages of the frequencies of teachers' perceived usefulness of assessment

#### 4.5 Relations between assessment use, purposes, and perceived usefulness

The relations between the frequency of the different assessment methods and that of the purposes were quite straightforward. The observation-based method of observing students was used weekly (78%), similar to the formative purposes of providing feedback, determining speed of instruction, and adapting instruction (63–74%), and to the diagnostic purpose of investigating reasons for errors (60%). The other frequently used observation-based methods of asking questions and correcting written work were used weekly (91%), which also had the purpose of stimulating thinking (91%). Of the less frequently used observation-based methods, collecting scrap paper stood out. Remarkably, teachers reported almost never (56% yearly or less) to collect scrap paper, but

used assessment in 73% of the cases weekly to encourage the use of scrap paper. The instrument-based methods of using tests from a textbook or student monitoring system were used monthly (> 85%), corresponding to the summative purposes of establishing level groups, determining progress, and determining mastery (> 66%), that were mainly used with the same frequency.

When comparing perceived relevance of assessment methods with frequency of use, the most used observation-based methods, such as asking questions and observing students were generally those teachers found most relevant too. Textbook tests and student monitoring system tests, instrument-based methods, were seen as relevant, and were broadly used. Reciprocally, the least used methods, giving presentations and keeping portfolios, were seen as least relevant. Interestingly, student-developed and teacher-developed test problems were reported as irrelevant for approximately 50% of the time, whereas teachers reported using them at least monthly for 65-76% of the time.

In Table 2, the Spearman's rank correlations between the frequencies of assessment use, purposes of assessment, and perceived usefulness of assessment are shown. The total frequency of assessment use and the number of purposes of assessment were positively related ( $r = .416$ ,  $n = 927$ ,  $p < .001$ ). This indicates that the more often a teacher mentioned assessment purposes; the more often they reported to use assessment methods as well. The use of instrument-based assessment was slightly stronger related to the number of summative ( $r = .308$ ,  $n = 920$ ,  $p < .001$ ) than to the number of formative purposes ( $r = .225$ ,  $n = 920$ ,  $p < .001$ ). However, the use of observation-based assessment correlated almost equally with the number of summative purposes ( $r = .269$ ,  $n = 913$ ,  $p < .001$ ) and the number of formative purposes ( $r = .262$ ,  $n = 913$ ,  $p < .001$ ). Finally, the perceived usefulness of assessment was positively related to the frequency of assessment use and all purposes of assessment. It was slightly higher correlated with observation-based ( $r = .202$ ,  $n = 920$ ,  $p < .001$ ) than with instrument-based assessment methods ( $r = .143$ ,  $n = 920$ ,  $p < .001$ ).

Table 2

*Correlations Between the Frequencies of Assessment Use, Purposes of Assessment, and Perceived Usefulness of Assessment (ns > 918) <sup>a</sup>*

	I	I.a	I.b	II	II.a	II.b	II.c
I. Total frequency of assessment use	–						
I.a. Frequency of instrument-based ass.	.766 <sup>b</sup>	–					
I.b. Frequency of observation-based ass.	.895 <sup>b</sup>	.430	–				
II. Total number of assessment purposes	.416	.354	.359	–			
II.a. Number of summative purposes	.335	.308	.269	.866 <sup>b</sup>	–		
II.b. Number of formative purposes	.288	.225	.262	.657 <sup>b</sup>	.444	–	
II.c. Number of diagnostic purposes	.260	.208	.230	.594 <sup>b</sup>	.426	.321	–
III. Perceived usefulness of assessment	.212	.143	.202	.175	.160	.147	.067

<sup>a</sup> All Spearman's rank correlations significant at  $p < .001$  level

<sup>b</sup> These correlations are inflated as they are between the whole (e.g. Total frequency..) and its parts (e.g. Instrument-based assessment, Observation-based assessment)

## 5. Conclusions and discussion

Our study's primary goal was to close the knowledge gap on the current assessment practice of primary school teachers in mathematics education. We found that teachers in primary mathematics education in the Netherlands use a variety of assessment methods, use instrument-based and observation-based assessment methods just as frequently and find assessment generally useful. This perceived usefulness is shown by the overall very positive reactions teachers gave on the different uses of assessment (see Figure 4). The two main instrument-based assessment methods, textbook tests and tests from a student monitoring system, are reported as the most relevant, with asking questions and observing students the most relevant of the observation-based assessment methods. Furthermore, the teachers' responses to the questionnaire revealed that

they used assessment for formative, summative, and diagnostic purposes. Some researchers (e.g., Birenbaum et al., 2006) have argued that summative assessments are still more prevalent, or have found that teachers do not have enough time for formative-like assessments due to the time summative assessments take (e.g., Schmidt & Brosnan, 1996). Remarkably, teachers in our sample used formative assessment, such as providing feedback and adapting instruction, more frequently than summative assessment, such as determining progress or establishing level groups. Evidently this result can in part be explained by the mere nature of formative and summative purposes. The time devoted to summative assessment is obviously less than the time devoted to formative purposes.

Of course some prudence is needed in considering these conclusions. While using an online-questionnaire has allowed us to reach a large group of respondents in a fairly short time, this teachers' self-report through a questionnaire has its limitations. Individual interviews and classroom observation would have provided more detailed information on assessment practice. Further research is needed in this respect.

Another point of concern is the sampling process of this survey. Only a relatively small proportion of schools in the Netherlands responded. In international comparative studies such as TIMSS, schools in the Netherlands are known to not be very prone to respond to requests to participate in research. Nevertheless, our survey was not specifically focused on evaluating teachers or schools, but rather on collecting information on teachers' assessment practice. This could explain the relatively large sample that gave suite to our request. However, because not all teachers filled in the questionnaire we cannot exclude that there was some further bias in our sampling. Nonetheless, through the size of our sample and a comparison of several background characteristics (cf. Section 3.1) with a national sample, we have tried to eliminate confounding variables as much as possible. Therefore, we are convinced this study gives us a quite reliable picture of current classroom assessment practice in Dutch primary school mathematics education.

What does this picture tell us? First of all, when we compare our findings with those from the last survey (Janssens, 1986) of teachers' mathematics assessment practice in primary education in the Netherlands there are clear differences.

## Chapter 2

Some 25 years ago a lack of teacher knowledge on assessment was found, resulting in incoherent evaluation practices. Our results indicate that teachers now use a variety of assessment methods for different purposes, combining observation- and instrument-based assessment methods with formative, diagnostic, and summative purposes. Apparently, in these twenty-five years, teachers in the Netherlands have acquired some knowledge on assessment.

Another finding is that the assessment methods used in primary mathematics education do not completely coincide with the educational approach of RME, though almost all teachers use a textbook based on its principles. In fact, Dutch teachers have a quite classical way of assessing their students' knowledge, in the sense that assessment methods that relate more to a reformed constructivist approach to mathematics education such as practical assignments, collecting scrap paper, and student-developed test problems, apparently still play a minor role in Dutch mathematics education in primary school. The integration of the educational approach (RME) and the methods used for assessment, as illustrated by this study, remains wanting.

Nevertheless, formative use of assessment was reported to be most frequent, which matches the recent focus on data-driven teaching (Ledoux, Blok, Boogaard, & Krüger, 2009; Timminga & Swanborn, 2010). This is an encouraging finding if one takes into account that data-driven teaching and a formative use of assessment have been related to considerable improvement in student achievement (e.g., Black & Wiliam, 1998; Crooks, 1989; Kingston & Nash, 2011; Natriello, 1987). An earlier study found (Ledoux et al., 2009) that teachers used assessment effectively to gather data, but did not sufficiently take the next step of acting on the information gained with these data. The results of our survey indicate that teachers do use assessment information for various purposes, from giving feedback via adapting instruction to stimulating thinking, to name a few. In the continuous struggle to improve education, this is yet another encouraging finding.

## References

- Airasian, P. W. (1991). *Classroom Assessment*. New York: McGraw-Hill.
- Anderson, L.W. (2003) *Classroom Assessment: Enhancing the Quality of Teacher Decision Making*. Mahwah, N.J. Lawrence Erlbaum Associates, Inc.
- Berry, R. (2011). Assessment reforms around the world. In R. Berry & B. Adamson, *Assessment reform in education – Policy and practice* (pp. 89-102). Dordrecht, the Netherlands: Springer.
- Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R., & Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review*, 1(1), 61-67.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74.
- Blok, H., Otter, M. E., & Roeleveld, J. (2001). *Het gebruik van leerlingvolgsystemen anno 2000* [The use of student monitoring systems in the year 2000]. Amsterdam, the Netherlands: SCO-Kohnstamm Institute.
- Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in Motivating Student Effort and Achievement. *Applied Measurement in Education*, 10(2), 161-180.
- Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education: Principles, Policy, & Practice*, 11(3), 301-318.
- Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation in Higher Education*, 37(1), 33-43.
- Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438–481.
- De Lange, J. (1987). *Mathematics, insight and meaning*. Utrecht, the Netherlands: OW & OC, Utrecht University.
- Dixon, H. R., Hawe, E., & Parr, J. (2011). Enacting assessment for learning: the beliefs practice nexus. *Assessment in Education: Principles, Policy, & Practice*, 18(4), 365-379.
- Gardner, J., Harlen, W., Hayward, L., & Stobart, G. (2011). Engaging and empowering teachers in innovative assessment practice. In R. Berry & B. Adamson, *Assessment reform in education – Policy and practice* (pp. 105-119). Dordrecht, the Netherlands: Springer.

- Janssens, F. (1986). *De evaluatiepraktijken van leerkrachten. Een beschrijvend onderzoek naar het evalueren tijdens het rekenen in het primair onderwijs* [The evaluation practices of teachers. Descriptive research into the evaluation of mathematics in primary education]. Arnhem, the Netherlands: Cito.
- Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment: 75 practical strategies for linking assessment, instruction, and learning*. Thousand Oaks, CA: Corwin.
- Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28-37.
- Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). *Opbrenstgericht werken; over waarde van meetgestuurd\_onderwijs* [Data-driven decision making; the value of measurement oriented education].\_SCO-Rapport 812. Amsterdam: SCO-Kohnstamm.
- Mavrommatis, Y. (1997). Understanding assessment in the classroom: Phases of the assessment process – the assessment episode. *Assessment in Education: Principles, Policy, & Practice*, 4(3), 381-399.
- Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, 22(2), 155–175.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching & Teacher Education*, 27(2), 472-482.
- Romberg, T. A., Zarinnia, E. A., & Collis, K. F. (1990). A new world view of assessment in mathematics. In G. Kulm, *Assessing higher order thinking in mathematics* (pp. 21-38). Washington, DC: AAAS.
- Scheltens, F., Hemker, B., & Vermeulen, J. (2013). *Balans van het rekenwiskundeonderwijs aan het einde van de basisschool 5* [Audit of mathematics education end of primary school]. Arnhem, the Netherlands: Cito.
- Schmidt, M., & Brosnan, P. A. (1996). Mathematics assessment: practices and reporting methods. *School Science and Mathematics*, 96(1), 17-20.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné and M. Scriven (Eds.), *AERA Monograph Series on Curriculum Evaluation Vol. 1 - Perspectives on Curriculum Evaluation* (pp. 39-83). Chicago: Rand McNally.

- Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation*, 37(1), 49-54.
- Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.
- Statistieken ArbeidsMarkt Onderwijs Sectoren* [Statistics of the educational labour market] (2010). Retrieved from <http://www.stamos.nl>
- Stiggins, R. J., & Bridgford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), 271-286.
- Suurtamm, C., Koch, M., & Arden, A. (2010). Teachers' assessment practices in mathematics: Classrooms in the context of reform. *Assessment in Education: Principles, Policy & Practice*, 17(4), 399-417.
- Timminga, E., & Swanborn, M. (2010). *Stand van zaken opbrengstgericht werken in het basisonderwijs bij rekenen-wiskunde* [The current situation of data-driven decision making in primary mathematics education]. Utrecht: Inspectie van het Onderwijs.
- Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht, the Netherlands: CD-β Press / Freudenthal Institute, Utrecht University.
- Van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic Mathematics Education. In S. Lerman (Ed.), *Encyclopedia of mathematics education*. London: Springer.
- William, 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, CT: Information Age Publishing.
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3-14.



## **Chapter 3**

# **Exploring teachers' use of classroom assessment techniques in primary mathematics education**

Michiel Veldhuis<sup>a</sup> & Marja van den Heuvel-Panhuizen<sup>a,b</sup>

*<sup>a</sup>Freudenthal Institute for Science and Mathematics Education, Utrecht University, <sup>b</sup>Department of Pedagogical and Educational Sciences, Utrecht University*

## **Exploring teachers' use of classroom assessment techniques in primary mathematics education**

### **Abstract**

For teachers to teach their students mathematics in the best possible way, they need to know students' understanding, in order to adapt their teaching accordingly. Classroom assessment provides a means to get access to students' mathematical knowledge and thinking. In the present paper we describe two studies on the feasibility of primary teachers using classroom assessment techniques and on their effectiveness on students' achievement in mathematics in the Netherlands. These classroom assessment techniques were proposed to the teachers in five (Study 1) or four (Study 2) workshops. Both studies were small-scale and made use of qualitative and quantitative data of in total 10 teachers and their 214 students. To investigate the feasibility, data from classroom observations and teacher feedback reports were used. To investigate the effects on student achievement we used a quasi-experimental pretest/posttest design. Concerning the feasibility of the classroom assessment techniques, teachers and students reported enjoying the techniques and finding them useful; in the sense that they provided them with valuable information supporting their teaching and learning. In terms of the possible effects of the classroom assessment techniques on mathematics achievement, results indicate students improving considerably more than would be expected when compared to national reference data. These results mainly produce evidence for the feasibility of classroom assessment techniques in primary school mathematics, but also provide an indication for their possible effect on students' mathematics learning.

*Keywords:* Mathematics education; primary school; classroom assessment techniques; teachers' experiences; student achievement.

## **1. Introduction: Classroom assessment**

Developing and keeping track of students' mathematical abilities are important parts of every mathematics teacher's daily practice. In order for teachers to gather knowledge of their students' learning, assessment plays a pivotal role (Cizek, 2010). When this assessment is truly in the hands of teachers it is called *classroom assessment* (e.g., Shepard, 2000). The main purpose for teachers using classroom assessment is to assess students' skills and understanding in order to make informed instructional decisions. Such use of assessment for formative purposes differs from summative assessment that is aimed at making decisions about certification, and from diagnostic assessment that is focused on identifying reasons for individual students' difficulties (see, Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015).

Knowing what students know is the sine qua non for teaching (e.g., Pellegrino, Chudowsky, & Glaser, 2001). Knowledge of students' comprehension is important for finding out whether a particular instruction has had its desired effect and generating ideas for how to proceed in subsequent mathematics lessons (e.g., Jacobs, Lamps, & Philip, 2010). Based on assessment information on students' mathematical understanding teachers can align their teaching to their students' needs (Schoenfeld, 2014), which in turn can result into adapting their teaching, but can of course also mean not changing anything and continuing with what was planned before.

Classroom assessment in this sense has been related to achievement gains in students. An important study that brought the potential of classroom assessment to the fore is the review by Black and Wiliam (1998) that reported the different practical expressions of classroom assessment to be the most effective interventions for teachers to improve student learning. Although several researchers have questioned the size of the effect of formative assessment on student learning in recent reviews or meta-analyses of existing studies (e.g., Briggs, Ruiz-Primo, Furtak, Shepard, & Lin, 2012; McMillan, Venable, & Varier, 2013), they do not contest its positive effect. Moreover, classroom assessment can provide teachers with "a record of far better reliability than any external test can achieve" (Black, 1990, p. 25). As a result of these findings, policymakers have urged the educational community, and in particular teachers, to embrace classroom assessment in their practice. For example, the United

States National Council of Teachers of Mathematics (NCTM, 2013) took the following position on formative assessment in mathematics education: “The use of formative assessment has been shown to result in higher achievement. The National Council of Teachers of Mathematics strongly endorses the integration of formative assessment strategies into daily instruction” (p. 1). Teachers are the only ones that can actively integrate these formative assessment strategies into their practice. Similarly, Wiliam (2007) earlier emphasized that for formative assessment “[t]o successfully raise student achievement, we must improve the quality of the teachers working in the schools” (p. 184).

This was also the focus of a number of projects carried out by Black and Wiliam and colleagues in the United States (Leahy, Lyon, Thompson, & Wiliam, 2005) and the United Kingdom (Wiliam, Lee, Harrison, & Black, 2004) to provide teachers with various assessment techniques to improve, and integrate in, their classroom assessment practice. What these assessment techniques have in common is that they blur the divide between assessment and instruction, require only small, low-tech and low-cost, changes in practice, and are suitable for teachers to implement in their classroom practice. An additional facet of these ‘classroom assessment techniques’, as we call them, is that they are predominantly activities that are familiar to teachers but that are now used with a clear assessment focus. We use classroom assessment technique to describe the entire activity, from the proposal by the teacher of a series of questions or mathematical problems and the practical reactions of the students, to the teacher’s continual gathering of information in different modes. This reflects Bennett’s (2011) characterisation of formative assessment, as that it “might be best conceived as neither a test nor a process but some thoughtful integration of process *and* purposefully designed methodology or instrumentation” (p. 7).

Various types of classroom assessment techniques have been proposed and used in international research (e.g., Torrance & Pryor, 2001; Van den Heuvel-Panhuizen & Becker, 2003; Wiliam, Lee, Harrison, & Black, 2004), or in more practice-oriented work (e.g., Keeley & Tobey, 2011; Leahy, Lyon, Thompson, & Wiliam, 2005; Wiliam, 2011). An example of a form of a classroom assessment technique is a question to which all students respond individually by holding up a card: an all-students response system (‘ABCD’ cards, Wiliam, 2011). Teachers can use the information gathered in this way to decide to go

over a particular explanation or subject again, or instead move onto the next; an instructional decision teachers make on a day-to-day basis (William, 2011). If teachers are better aware of their students' mathematical capabilities and understanding, through using classroom assessment techniques, then they can undoubtedly better adapt their teaching to the needs of the students. In doing this and providing explicit, and implicit, feedback students also become more aware of their own knowledge and ways of problem solving, and the circle is complete: students and teacher simultaneously advance. To be able to make a sound instructional decision on the basis of useful information, teachers have first to generate opportunities through posing particular questions for students to show their thinking. The implementation of such classroom assessment techniques and providing effective professional development for these were the challenges of the Improving Classroom Assessment (ICA) project in the Netherlands.

## **2. Classroom assessment in the Netherlands**

The starting point of the ICA project was to get an overview of the current assessment practice in mathematics of primary school teachers in the Netherlands. By means of a nationwide survey through an online questionnaire it was investigated which particular assessment methods the teachers used, the purposes they assessed for, and teachers' perceived usefulness of these assessment methods in mathematics (Veldhuis, Van den Heuvel-Panhuizen, Vermeulen, & Eggen, 2013 – cf. Chapter 2 of this thesis). The responses of this sample of 960 teachers revealed that to find out students' skills and comprehension level, teachers used methods ranging from externally developed standardized tests and tests that come with a textbook, to asking questions and observing students while they were doing exercises to adapt their instruction. It appeared that teachers generally used classical assessment methods, such as questioning and standardized tests, and even though teachers used assessment for various formative purposes, only very few teachers used more authentic assessment methods, such as classroom assessment techniques. Other studies suggest that this is also the case for secondary education (e.g., Segers & Tillema, 2011) where summative assessments are abundant. The Dutch Inspectorate (Inspectie van het onderwijs, 2013) pointed out that many primary schools (40%) and secondary schools (33%) do not systematically use assessments to monitor their students' progress. As such, the Inspectorate strongly advised schools and teachers to make more use of

classroom assessment in mathematics and language. Clearly there is room for improvement in the use of classroom assessment by teachers in primary mathematics education in the Netherlands. This situation is not exclusive to the Netherlands, the same matters on the use of classroom assessment have been signalled internationally, also in the countries where nationwide attention to assessment for learning is given (for example in New Zealand, Absolum et al., 2009; or the United Kingdom, James, 2011). In both these countries classroom assessment practice has been receiving attention for several decades and is advocated by the governing bodies. Nonetheless it is found that teachers also struggle in their practice to use the kinds of classroom assessment techniques such as described, for example, by Leahy et al. (2005). So, in the Netherlands and internationally, there is still a need for making teachers familiar with using classroom assessment techniques.

The purpose of the studies described in the current paper was to investigate the implementation of the use of classroom assessment techniques by primary school teachers in the Netherlands. As such our endeavor was to develop a kind of blueprint for the implementation of classroom assessment in primary mathematics education. To do this we first strived to generate more knowledge about the practical applications of classroom assessment techniques for mathematics in primary school, and secondly we aimed at establishing possible effects on student learning. More particularly our research questions were:

1. In what ways do primary school teachers take up the classroom assessment techniques in mathematics?
2. Could the use of these assessment techniques contribute to an achievement gain for the students?

### **3. Method**

#### **3.1 General design**

To investigate these research questions we performed two consecutive studies with groups of third-grade teachers in the Netherlands. The studies were carried out from October 2011 to June 2013. The teachers participated in monthly workshops, consisting of three or four teachers and the first author, during the

second semester of Grade 3. In these workshops classroom assessment techniques were presented, discussed, and evaluated.

The first research question was investigated by conducting regular classroom observations of every teacher in between workshops. These observations were intertwined with short informal interviews. The teachers were also asked to register their evaluation of the used assessment techniques. These different sources of information were used to determine how the teachers implemented the classroom assessment techniques in practice, how their students reacted to them, and what the students and teachers alike thought of the classroom assessment techniques.

To answer the second research question, we used a pre-/post-test evaluation of students' mathematics achievement. The pre-test data consisted of the results from the midyear student monitoring system test for Grade 3 and the results from the end of year student monitoring system test for Grade 3 served as post-test data (Cito LOVS; Janssen, Scheltens, & Kraemer, 2006). These biannual student monitoring system tests are used in virtually all primary schools in the Netherlands to monitor students' development in mathematical ability over the primary school years. The teachers administered the tests in their own classes as is common in educational practice in the Netherlands. The scores on these tests are mathematical ability scores calculated through item response theory models. By using these test results as pre-test/post-test data we could evaluate firstly whether the students progressed in their mathematics ability, and secondly whether students of teachers that had participated in the workshops improved more than the national norm sample of students of teachers (that of course had not participated in the workshops).

### **3.2 Participants**

The two studies were carried out in the second semester of two consecutive school years. Schools from our network were approached by e-mail and eventually ten teachers volunteered to participate. The teachers taught 214 Grade 3 students (14 to 29 students per class; see Table 1 for some background characteristics on these teachers). In the first study four female teachers participated in five workshops. The mean age of these teachers was 38.5 years ( $SD = 15.1$ ). In the second study two male teachers and four female

teachers participated in four workshops and the mean age of these six teachers was 52.5 years ( $SD = 10.9$ ). The teachers used four different textbooks that were all based on the principles of Realistic Mathematics Education (see Van den Heuvel-Panhuizen & Drijvers, 2014), as is common in the Netherlands. All schools were situated in urbanized areas, with highly mixed student populations. As one of the teachers put it: “We have the dentist’s son and the cleaning lady’s daughter.”

Table 1

*Background Characteristics of the Classes of Participating Teachers*

	Number of students in Grade 3	Teacher age (years)	Teacher gender	Region	Grades in class	Text- book
<b>Sub-study 1</b>						
Class I	15	46	F	ZH	2 <sup>nd</sup> /3 <sup>rd</sup>	RR
Class II	13	56	F	ZH	2 <sup>nd</sup> /3 <sup>rd</sup>	WR
Class III	14	28	F	ZH	2 <sup>nd</sup> /3 <sup>rd</sup>	PP
Class IV	24	24	F	ZH	3 <sup>rd</sup>	PP
<i>Total</i>	<i>66</i>					
<b>Sub-study 2</b>						
Class V	22	62	F	ZH	3 <sup>rd</sup>	WIG
Class VI	17	59	M	ZH	3 <sup>rd</sup> /4 <sup>th</sup>	WIG
Class VII	26	50	M	NH	3 <sup>rd</sup>	WIG
Class VIII	27	54	F	ZH	3 <sup>rd</sup>	WIG
Class IX	29	32	F	NH	3 <sup>rd</sup>	WIG
Class X	27	58	F	NH	3 <sup>rd</sup>	WIG
<i>Total</i>	<i>148</i>					

*Note.* F = female, M = male, NH = Noord-Holland (region of Amsterdam); ZH = Zuid-Holland (region of The Hague); RR = Rekenrijk; PP = Pluspunt; WR = Wis en Reken; WIG = De Wereld in Getallen

### 3.3 Material: Classroom assessment techniques

In this study we proposed classroom assessment techniques consisting of short activities of less than 10 minutes to the teachers (cf. Figure 1). The techniques were intended to help teachers to quickly find out something about their students’ mathematical skills and understanding, provide teachers with

indications for further instruction, and focus on particular mathematics content. Each technique had a particular format and focused on part of the mathematics curriculum of the second semester of Grade 3.

<b>Characteristics of classroom assessment techniques</b>
<ul style="list-style-type: none"> <li>• Every technique has a particular <i>format</i> and is focused on particular <i>content</i></li> <li>• Helps teachers to <i>quickly</i> find out something about their students' mathematics comprehension</li> <li>• Provides indications for <i>further teaching</i></li> <li>• Consists of a <i>short activity</i> (&lt;10 minutes)</li> </ul>

*Figure 1.* Characteristics of the classroom assessment techniques

We used two different formats for the classroom assessment techniques: (1) whole-classroom immediate response systems and (2) worksheets that the teacher/students had to check later. The mathematics content on which most assessment techniques were centred was number knowledge, mainly in the context of addition and subtraction, but the techniques could also be used to assess multiplication and division tables. In using the techniques we strived for teachers to enlarge and develop their assessment repertoire, as such enabling them to in the future continue to use, and elaborate on, the techniques. In all workshops attention was paid to reasons for using these 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 feedback to students about the findings of the assessments, they could become explicitly aware of their own understanding. We also underlined that assessment is a discursive process, in the sense that teachers assess students not only through their written but also through their spoken interactions with students (Reis & Barwell, 2013). Out of the nine assessment techniques we used in these sub-studies we illustrate three in detail in the following.

### *3.3.1 Example 1: CAT Crossing ten and more*

This classroom assessment technique can help teachers to quickly find out whether students have a particular mathematical insight and is inspired by several existing assessment techniques (cf. Wiliam, 2011). The teacher asks all students a series of questions that can be answered quickly with Yes or No. Students all have a red and a green card to show their answers. By inspecting

the waving red and green cards the teacher gets an immediate overview of all students' responses. The CAT *Crossing ten and more* (Figure 2) with the red and green cards is used for assessing whether students have ready knowledge about when a total of two numbers, e.g., 7 and 4, is under ten or over ten. This instant number fact knowledge is needed to perform numerical operations. Knowing whether the digits are crossing ten is imperative when students have to carry out mental additions and subtractions with two-digit numbers. For solving these problems it is crucial that students can instantaneously identify whether crossing is the case, because this has consequences for the strategy to be applied. After *crossing ten* the teacher can continue to use this classroom assessment technique with *crossing one hundred* and *one thousand*. In this way the teacher can also assess whether the students have a clue about the analogy between different number domains. For some students 70 and 40 will be a new problem whereas others knew immediately that what applied to 7 and 4 also applies to 70 and 40. Analogously, 700 and 400 will be new to some, but a piece of cake to those that understand the analogy.

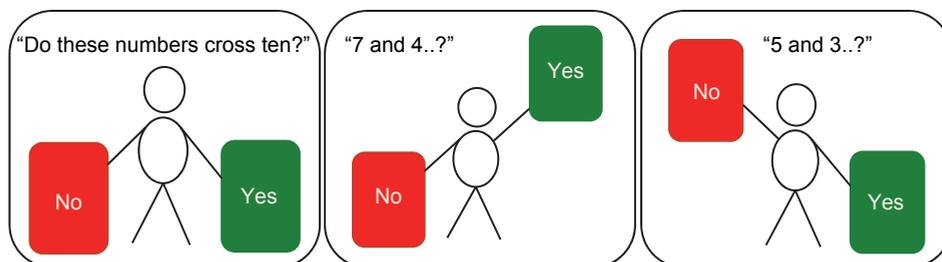


Figure 2. CAT *Crossing ten and more* with the red and green cards

The CAT *Crossing ten and more* can easily be adapted to investigate subtraction: "Is the difference between 15 and 7 bigger than 5?" or multiplication and division tables: "Is 44 in the table of 8?" By inspecting students' reactions to these series of questions, teachers quickly gather information on their students' number knowledge, while students are practicing and developing further understanding of the relations between numbers and the parallels between ones, tens, and hundreds.

### 3.3.2 Example 2: CAT *Easy or difficult*

Another classroom assessment technique, also related to students' knowledge of number facts and number operations, is the CAT *Easy or difficult*. Here the

main issue for the teacher is to get to know whether their students are aware of the difficulties some number operations have and whether they can reflect on these difficulties. In this classroom assessment technique students have to indicate of problems whether they are, according to them, easy or difficult (cf. Van den Heuvel-Panhuizen, Middleton, & Streefland 1995). All students are given a worksheet containing two columns of similar problems that differ on particular aspects. These aspects pertain to whether the two numbers presented cross ten or not, and this in the context of addition (e.g.,  $11 + 2$  or  $13 + 12$ ) and subtraction (e.g.,  $26 - 7$  or  $35 - 4$ ). Other aspects are for instance the size of the numbers (the number of digits: e.g.,  $20 + 40$  or  $200 + 40$ ) or the order in which the numbers are presented (larger or smaller number first: e.g.,  $54 + 20$  or  $19 + 54$ ). Of a pair of problems the student has to circle the one of which he or she thinks is the easiest, without calculating the answer. After completing the series of problems (about 15 problems) the students exchange their worksheets and discuss their reasoning with each other, explaining differences or commonalities. All the while the teacher listens in on the students' discussions, as such clearly using this assessment technique as a discursive process (e.g., Reis & Barwell, 2013), while gathering information about their level of understanding and reasoning when solving problems. In the whole-class discussion the teacher can further use this information. Of course the teacher can also collect the worksheets to be able to inspect them in more detail.

### 3.3.3 Example 3: CAT Word problem experiment

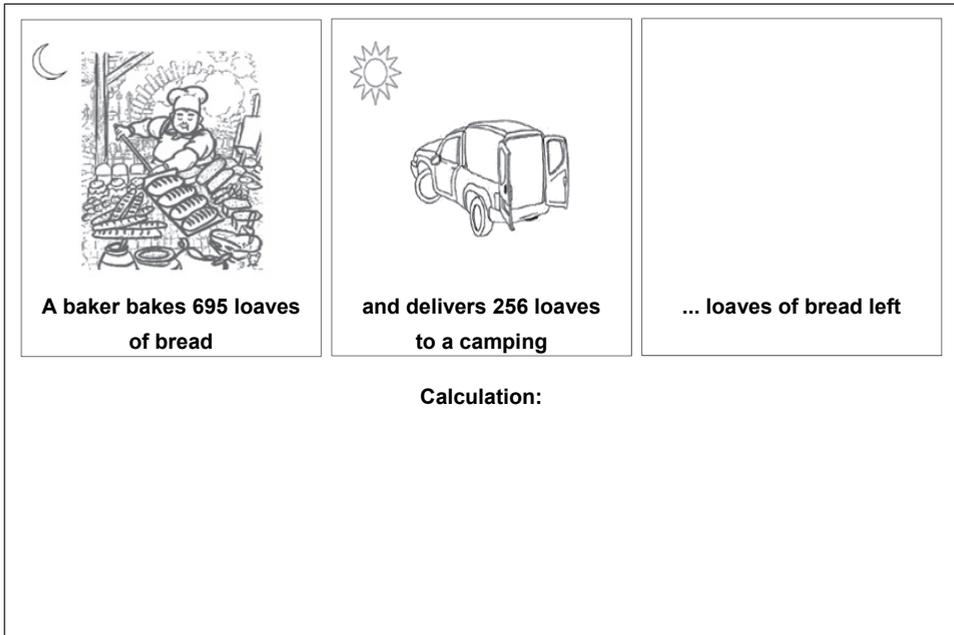
The *CAT Word problem experiment* consists of setting up an experiment by the teacher to find out what students can or cannot do. In fact, the teacher behaves as a researcher investigating some phenomenon by creating different conditions for solving problems. In case of the classroom assessment technique shown in Figure 3 the focus is on word problems. Despite the high value attached to teaching students the ability to use mathematics to solve context problems, students often struggle with solving word problems. These difficulties with word problems can be due to a variety of reasons, for instance, 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 (e.g., Tolar, Fuchs, Cirino, Fuchs, & Hamlett 2012). To find out where the problem lies for individual students this assessment technique works as follows. The teacher lets the students solve a

series of problems as word problems, and later gives them the same problems in a different format, namely as bare number problems. 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.

<p><b>At night a baker bakes 695 loaves of bread. He delivers 256 loaves to a camping. How many loaves of bread does he have left?</b></p> <p>... loaves</p> <hr/> <p><b>Calculation:</b></p>	<p><b>695 – 256 = ...</b></p> <hr/> <p><b>Calculation:</b></p>
---	--

*Figure 3. Two worksheets of the CAT Word problem experiment, presented as a word problem (on the left) and as a bare number problem (on the right)*

When checking students' work the teacher can find that solving the problem in one format or the other does not trigger the use of the same solution strategies nor results in the same outcome for all students. Some students can have difficulties transforming the problem situation into a mathematical problem, others with carrying out the calculation, and some with a combination of these. With this knowledge the teacher can adapt his or her instruction and explicitly address the comprehension of the text and the transformation as fundamental abilities for solving word problems. For example, one way to help students with comprehending the text is offering them the problem in a format that is in between the word problem and the bare number problem: a schematic representation of the word problem (Figure 4). When this intermediate form helps the students they can, for example, be stimulated to make a drawing by themselves next time. Moreover, discussing parallels between the different representations of the same problem, whether it is presented as a word problem, bare number problem, or schematically may also help them to get a better grip on word problems.



*Figure 4.* Schematic presentation of the word problem of Figure 3.

### 3.4 Procedure

The teachers used the classroom assessment techniques for mathematics at moments when they fitted in their schedule. In doing so they enlarged and reinterpreted their repertoire of assessment techniques. In both studies, during the second semester of Grade 3, teachers and the first author convened in monthly workshops. In these workshops each assessment technique was explicitly introduced as modifiable: teachers could vary the content and the form of the classroom assessment technique in such a way that it fitted their practice. The teachers could also adapt the classroom assessment techniques together during the workshops (e.g., Lin, 2006). All this is in line with Wilson and Sloane (2000) who noted that:

[T]eachers must be: Involved in the process of collecting and selecting student work, able to score and use the results immediately, able to interpret the results in instructional terms, [and] able to have a creative role in the way that the assessment system is realized in their classrooms. (p. 191)

The workshops were organized following the principle of ‘practice what you preach’ and could be considered teacher learner communities (cf., Lave & Wenger, 1991). According to Wiliam (2007) “five principles are particularly

important [in establishing and sustaining teacher learning communities]: gradualism, flexibility, choice, accountability, and support” (p. 197); we strived to incorporate all of these in our workshops. As such teachers could “adopt and integrate these techniques and others into their own practice, find a new synergy and see their own practice in new ways, which in turn leads to new thinking” (ibid, p. 195).

Teachers and researchers worked together in a teacher learning community in order to determine the important mathematics content in the weeks between the workshops and ways to find out whether students had learned the prerequisites or not. The researcher visited every teacher at least once between two consecutive workshops. In these visits he observed the teachers during mathematics instruction and the assessment techniques. As such the implementation of the techniques in the classrooms could be inspected and the researcher could reflect upon what he had seen and heard in the following workshop.

The order of business of every workshop was that the teachers shared their experiences with the assessment activities of the preceding weeks: which assessment techniques they used, why they used them, in what form, how the students reacted, what they thought of the activities, and what information they collected by the classroom assessment techniques and what they did as a follow-up with this information. These same questions were also on a feedback form the teachers were asked to fill out directly following the use of an assessment technique. During the workshop, after every teacher had told how their weeks had been, the researcher shared some observations he had made in the classrooms. All the while the teachers reacted to each other’s stories, they could suggest different approaches or ask for more details; generally these workshops were very lively and informative. Then the focus switched to the future weeks: the mathematics content and accompanying assessment possibilities. These were discussed, but after some discussion eventually the researcher proposed several ideas of which the teachers could select some. Then the researcher explained and sometimes showed how the assessment technique worked, and in particular what could be investigated with them. Finally, there was some more discussion about the activities and the researcher distributed the discussed techniques on paper so that the teachers could reflect upon them in preparing their lessons.

## 4. Results

The presentation of the data here is not intended to give a comprehensive overview of everything that went on in the classrooms of the participating teachers during the use of the classroom assessment techniques, but is focused on showing how they took up the techniques in their classrooms and what effects this seemed to have on student achievement. This means that we start with providing some illustrations of their practice with the techniques (for example what adaptations they made) and indications of what they thought of the techniques (for example whether they were useful and how students reacted to them). Hereafter we go into the apparent effects teachers' use of the classroom assessment techniques showed on student achievement.

### 4.1 Feasibility of the classroom assessment techniques

The CAT *Crossing ten and more* with the red and green cards seems quite straightforward on paper; nonetheless there was great variation in how teachers used this format in their respective classrooms. A teacher noticed that some students waited to see which card other students held up before choosing their own. She considered this a problem 'as it was a testing situation' and decided that students had to be in 'testing positions' (students seated at separated desks) and close their eyes so that they could not 'cheat'. Another teacher spent quite some time to ensure that all students were clear about what the colors green and red were, and in which hand they held each color. Yet another teacher interpreted the use of this classroom assessment technique more as a game and adapted it to his own practice. He considered it to be 'nonsense to be the only one doing the work' and let a student 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.

After using the CAT *Crossing ten and more* the teachers integrated their observations in their further instruction. One teacher had identified a type of problems most students struggled with and she wrote two examples of these on the blackboard to refer to in further instruction. Several teachers struggled with the information density of the CAT *Crossing ten and more*. They found it difficult to direct their attention to all relevant aspects of students' behavior,

including the color of the card the students showed and the speed with which they showed it, and then this for, for example, more than ten number pairs of which the students had to say whether they crossed ten or not. A possible solution that was suggested for this overload of information was formulating a hypothesis about specific students and paying special attention to these aspects of these students, without neglecting the overall overview. Many teachers found this solution helpful.

The CAT *Easy or difficult* gave the teachers insight into the characteristics of problems that students found more or less difficult. As this classroom assessment technique uses the worksheet format teachers could easily register the choices students made and elaborate on the knowledge of students' comprehension of crossing ten they had gathered with the CAT *Crossing ten and more*. Generally students found problems where the numbers did not cross ten easier and those where the numbers 7 and 8 appeared difficult. Students were taken in by this activity: this was the first time they had to do an assignment in the mathematics lesson in which they had to reflect upon what they found easy or difficult, without performing any calculations. When the assignment was discussed afterwards in class and the students had to explain why and how they had decided whether a problem was easy or difficult, many students identified crossing the ten as the difficulty.

For the CAT *Word problem experiment* all teachers agreed that it provided them with very valuable information on their students' strategies and difficulties when solving word problems. Most teachers expected all students to struggle more with the word problems than the bare number problems, before performing the experiment. They did not take into account the particularities of certain word problems or individual differences between students. Afterwards they remarked that students' performance on either type of problem format depended on the student and the type of operation and wording that was used. Using the results of this assessment technique, teachers adapted their further instruction to the specific needs of their students. Most teachers 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 three 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

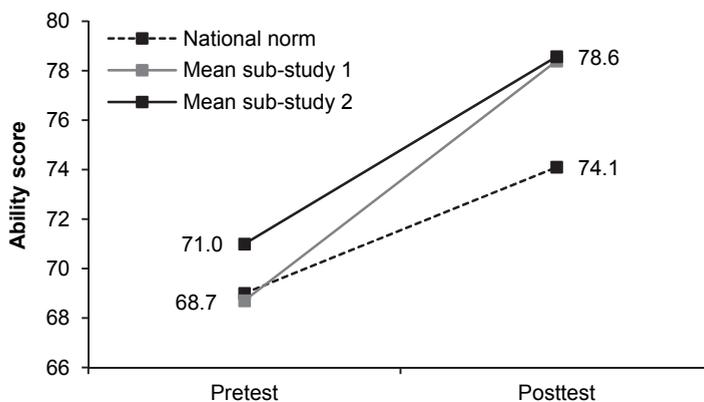
was, that the way of presenting them had changed; the mathematics needed to solve the problem remained exactly the same.

An overall finding from the classroom observations, the interviews, and the discussions in the workshops was that every teacher liked to use the classroom assessment techniques and that all teachers interpreted them in their own way and adapted them to their own practice. Even though teachers operated diversely in their classrooms and flexibly organized the implementation of different assessment techniques in their practice, this adaptive use did not counteract the usefulness of the techniques. Another general finding was that the students referred to the classroom assessment techniques as ‘mathematics games’. They were very motivated to participate in atypical classroom activities such as the CAT *Crossing ten or more* with the red and green cards and the CAT *Easy or difficult*, but also in the CAT *Word problem experiment*, which contains exercises that are not very different from those they normally have to do; nonetheless they happily worked on the worksheets. Afterwards students related that working on mathematics in this way made it much more fun.

#### 4.2 Effects associated to the classroom assessment techniques

For the second research question, about the effectiveness, we collected pre-/post-test data on students’ performance on the Cito student monitoring system tests. Evidently, without a control group the direct attribution of any effect of the intervention on student learning is not warranted. However, in analysing the pre-test/post-test data in comparison to the national reference data on the same student monitoring system tests, we can identify the size of the improvement, working as *proof of principle* for the effectiveness of the classroom assessment techniques. In both sub-studies the mean ability of students increased from midyear to end-of-year testing (see Figure 5). It can of course be expected that students advance in their mathematical ability, whether teachers perform specific assessment activities or not; the scores of the national norm sample also showed this direction. However, the mean difference between pre-test and post-test of the classes of participating teachers and its effect size were notably larger than those of the national norm sample (Study 1:  $M_{pre} = 68.7$ ,  $SD_{pre} = 14.0$ ,  $M_{post} = 78.4$ ,  $SD_{post} = 14.1$ , gain score ( $M_{post} - M_{pre}$ ) = 9.7,  $d = 0.81$ ; Study 2:  $M_{pre} = 71.0$ ,  $SD_{pre} = 14.2$ ,  $M_{post} = 78.6$ ,  $SD_{post} = 15.1$ , gain score = 7.6,  $d = 0.55$ ; National norm:  $M_{pre} = 69.5$ ,  $SD_{pre} = 14$ ,  $M_{post} = 74.1$ ,  $SD_{post} = 14$ ;

gain score = 5.1,  $d = 0.36$ ). Comparing these effect sizes with each other shows that students of teachers that participated in our studies advance between 53% ( $= (0.55 - 0.36) / 0.36$ ) and 125% ( $= (0.81 - 0.36) / 0.36$ ) more than students in the reference sample. This result provides an indication for the effectiveness of the use of these assessment techniques; students appear to advance more from the midyear to the end of the year testing than expected (gain scores of +9.7 (Study 1) and +7.6 (Study 2) instead of the expected +5.1 (National norm)).



*Figure 5.* The score gains for Study 1 (dark gray) and 2 (black) and the national norm (reference) sample (dotted line)

## 5. Discussion

The feasibility of the classroom assessment techniques combined with the improvement of students' mathematics achievement are the main results of these studies. Teachers found the classroom assessment techniques useful for getting insight in students' understanding and noticed students becoming more fluent in voicing their solution strategies. The gain scores in the two small-scale studies indicated that students improved considerably more when teachers made use of classroom assessment, than students from the norm sample. This relative gain was quite large taking into account that the professional development only took four or five meetings of about an hour.

## **5.1 Limitations**

These promising results have, however, to be interpreted carefully, for the design of our study had some inherent flaws. Because our primary focus was on finding out whether it was feasible for teachers and students to use the classroom assessment techniques in mathematics, we did not include a proper control group. In this way we were able to use all participating teachers' experiences to investigate their implementation of the assessment techniques. Due to the lacking of a control group, it is possible that other factors influenced teachers' behaviour and students' performance. The direct attribution of the learning effects to the sole use of classroom assessment techniques would thus be too simple. However, we were able to compare the score gains to a reference sample, as such providing an indication for the relative size of students' improvement. Our sampling was convenience based –not random– what could possibly have negatively influenced the generalizability of the results. Also, the mere fact that the teachers voluntarily participated in meetings focused on teaching and learning mathematics could have had an influence on the results. These teachers could have been teachers with a particular interest in mathematics education. The same can be said for the extra attention the teachers got from the researcher, which could have caused the so-called Hawthorne effect. To avoid this noise in the design and get more robust findings about the effects of the classroom assessment techniques it is necessary to carry out a more controlled experiment with a larger sample of teachers. In further research into classroom assessment techniques for mathematics in primary education the research design should include a control group and strive for random sampling of participating teachers.

## **5.2 Conclusions**

Bearing these limitations in mind, what remains of our results? Firstly we note that our studies were only exploratory of nature and in this sense provide food for thought and further investigation. Nevertheless, they also gave at least some indications of what is important for the effective use and the implementation of classroom assessment techniques. For instance, there is some evidence that the frequency and number of workshops in which teachers participated might have had an influence on the student mathematics achievement results: in the second study, with four instead of five meetings, the achievement gain was smaller than in the first study. Furthermore, the adaptability of the classroom assessment

techniques could very well have contributed to the enthusiasm with which the teachers used the techniques (cf. Wilson & Sloane, 2000). Yet, the openness in the suggested classroom assessment techniques also made it impossible to compare the effectiveness of the exact use of these specific techniques. Another thing that was clearly experienced in working with the teachers was that using the classroom assessment techniques brought about a certain change in teachers' perspective on assessment and teaching. This was evident in how teachers came to see the techniques more and more as valuable sources of information. Students also had an added benefit, not only that their teachers were more aware of their understanding and thus could give explicit feedback, but students also received implicit feedback during the assessment activities and were all the while developing their mathematical skills.

The classroom assessment techniques focused on ways to promote learning and engage students in their learning process. This permitted teachers to develop new insights about their students' understanding and mastery of mathematics all the while letting students practice and develop insight about their own problem solving and understanding. This echoes the good teaching practice, which Ginsburg (2009) voiced when discussing formative assessment in relation to mathematics education. An important distinction has to be made between assessment techniques and (pure) instruction. Evidently the two are intrinsically linked, as an important aspect of the classroom assessment techniques is that they direct teachers' and students' thinking related to a particular mathematical topic. However, the most important feature of an assessment technique is whether it provides the teacher with valuable and useful information about his or her students (e.g., Erickson, 2007). In that sense the classroom assessment techniques, mainly focusing on revealing information about students' learning; clearly differ from instruction, mainly focusing on creating a setting in which students can come to learning.

Teachers liked the fact that they could easily adapt the assessment techniques, allowing them to reflect on their own practice. Through these adaptations teachers also developed ownership of the classroom assessment techniques; they became an integral part of the teachers' own educational practice. Teachers' ownership of the assessment techniques gives a strong indication for the sustainability of these: teachers are likely to continue using them in the future. While using the assessment techniques, such as those described in this

article, teachers became, by looking analytically at students' learning process, more aware of student learning. The underlying goal of letting teachers work with these different techniques was also to serve as an eye-opener in two ways. Each assessment technique could reveal to the teachers the possibilities and capabilities of their students, but could also give teachers a view on what aspects of mathematics are of importance to teach their students. Using the assessment techniques was the first step; the next step was for teachers to integrate the ideas behind the assessment techniques into their practice. Upon their first encounter with the classroom assessment techniques teachers generally thought them just to be 'interesting mathematics activities' but were far from sure whether using them would have any effect on their students or their own instruction. When they returned for the next meetings they were very enthusiastic about the techniques themselves, underlining the fun students had while participating and the amount of information they were gathering in the meantime. Towards the end of the series of meetings with the teachers they voiced how from using the techniques as 'just another mathematics activity' in the beginning they had come to realize that they could use the techniques to gather valuable information about their students. They underscored the fact that in the following years it would be easier and to further integrate the assessment techniques in their daily practice. This 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. As a conclusion for our current studies, we can say that these primary school teachers had integrated a fresh and new perspective on the assessment and the teaching of mathematics in their classrooms, which holds promise for the further dissemination of classroom assessment techniques in mathematics education.

## References

- Absolum, M., Flockton, L., Hattie, J., Hipkins, R., & Reid, I. (2009). *Directions for Assessment in New Zealand (DANZ): Developing students' assessment capabilities*. Retrieved from: <http://assessment.tki.org.nz/Assessment-in-the-classroom/DANZ-report>
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5-25.
- Black, P. J. (1990). APU science: The past and the future. *School Science Review*, 72(258), 13–28.
- 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.
- Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3-17). Abingdon, UK: Routledge.
- Clark, I. (2012). Formative assessment: assessment is for self-regulated learning. *Educational Psychology Review*, 24(2), 205-249.
- 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.
- Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: children's minds, teachers' minds. *Human Development*, 52(2), 109-128.
- Inspectie van het onderwijs (2013). *De staat van het onderwijs* [The current status of education]. Utrecht, Netherlands: Author.
- Jacobs, V. R., Lamb, L. L. C., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, 41(2), 169-202.
- James, M. (2011). Assessment for learning: research and policy in the (dis)United Kingdom. In R. Berry & B. Adamson (eds.), *Assessment Reform in Education, Policy and Practice* (pp. 15-32). London: Springer.

- 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, Netherlands: Cito.
- Keeley, P., & Tobey, C. R. (2011). *Mathematics formative assessment: 75 practical strategies for linking assessment, instruction, and learning*. Thousand Oaks, CA: Corwin.
- Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.
- Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: minute-by minute and day by day. *Educational leadership*, 63(3), 18-24.
- 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.
- 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). Retrieved from <http://pareonline.net/getvn.asp?v=18&n=2>
- National Council of Teachers of Mathematics (2013). *Formative assessment: a position of the National Council of Teachers of Mathematics*. Position document, NCTM. Retrieved from <http://www.nctm.org/about/content.aspx?id=37990>
- Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing What Students Know. The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Reis, G., & Barwell, R. (2013). The interactional accomplishment of not knowing in elementary school science and mathematics: Implications for classroom performance assessment practices. *International Journal for Science and Mathematics Education*, 11, 1067-1085.
- Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? A story of research and practice, productively intertwined. *Educational Researcher*, 43(8), 404-412.
- Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation*, 37(1), 49-54.

- 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.
- Tolar, T. D., Fuchs, L., Cirino, P. T., Fuchs, D., & Hamlett, C. L. (2012). Predicting development of mathematical word problem solving across the intermediate grades. *Journal of Educational Psychology*, 104(4), 1083-1093.
- 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.
- 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*. London: Springer.
- 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.
- Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K. & Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*. Advance online publication doi: 10.1080/0969594X.2014.999024
- Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen J. A., & Eggen, T. J. H. M. (2013). Teachers' use of classroom assessment in primary school mathematics education in the Netherlands. *CADMO*, 21(2), 35-53.
- William, 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, CT: Information Age Publishing.

- William, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.
- William, 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.



## **Chapter 4**

# **Supporting primary school teachers to improve their assessment practice in mathematics: Effects on students' learning**

Michiel Veldhuis<sup>a</sup> & Marja van den Heuvel-Panhuizen<sup>a,b</sup>

*<sup>a</sup>Freudenthal Institute for Science and Mathematics Education, Utrecht University, <sup>b</sup>Department of Pedagogical and Educational Sciences, Utrecht University*

## **Supporting primary school teachers to improve their assessment practice in mathematics: effects on students' learning**

### **Abstract**

In a large-scale quasi-experimental study we investigated the effectiveness of teachers' use of classroom assessment in mathematics education. The leading research question was: Do primary school students improve their mathematics achievement from teachers using classroom assessment? Participants were 616 students in 30 third-grade classrooms. Thirty teachers were randomly assigned to four conditions: participating in 1, 2, or 3 one-hour workshops for the experimental conditions and no workshops in the business-as-usual control condition. These workshops were focused on the use of classroom assessment techniques to reveal students' learning in the domain of number and enable teachers to adapt their teaching to the students' needs. Before the workshops started, students had completed the standardized midyear mathematics student-monitoring test, which was used as pre-test data. The results of the end-of-year test were used as post-test data. Contrary to our expectations students from teachers that participated in one or two workshops did not improve more than students from teachers in the control group, however, students from teachers who had three workshops on the use of classroom assessment techniques did improve significantly more than students from teachers in all other conditions ( $d = 0.26$ ). This result suggests that a short intervention (of three hours) on the use of classroom assessment techniques for mathematics led to increased student achievement.

*Keywords:* Classroom assessment; student learning; mathematics education; teachers; professional development

## **1. Introduction**

Improving students' proficiency in all school subjects is an omnipresent topic in educational and learning sciences. One of the key school subjects for which researchers strive to raise students' proficiency is mathematics. To reach high achievement levels in mathematics various angles of approach can be, and have been, chosen. Approaches range from improving the mathematics curriculum or textbooks, developing better instruction methods and learning materials, to making teachers better in teaching mathematics by enhancing mathematics teacher education, or creating more challenging and attractive school settings and learning environments. Of course, all these factors are relevant in developing students' mathematical knowledge, skills, and understanding, but the factor that is often considered to be most important for students' achievement in mathematics is the teacher (e.g., Slavin & Lake, 2008). The teacher not only initiates students' learning process by means of instruction and activities, but also provides guidance throughout the process, for instance through providing meaningful exercises and accompanying feedback. As the teacher is crucial for students' study success, it makes sense to zoom in on the role of the teacher in an effort to improve mathematics learning.

### **1.1 Classroom assessment**

To be able to provide pertinent guidance during the learning process, teachers need to have profound knowledge of their students' learning progress (Moreland, Jones, & Northover, 2001). Not only this knowledge is necessary, it is even impossible to teach without it, because the teaching should build on and link to what the students already know. In other words, mathematics teachers need to have insight into students' mathematical thinking (Gearhart & Saxe, 2005).

The guidance teachers provide in their mathematics classes can be more or less effective for stimulating students' understanding, depending on whether their instruction is attuned to the students' needs and possibilities for further development. In a continual striving for providing the best possible explanations to students, teachers need to know at practically every particular moment in their classes of every single student where they are in their understanding (cf. Wiliam, 2011). This was recently echoed by Schoenfeld (2014) when he wrote, that "[p]owerful instruction 'meets students where they are' and gives them opportunities to move forward" (p. 407). Without knowing 'where students are'

teachers will frequently be out of tune with their students, because “[o]nly as we appraise a student’s achievement and as we get a comprehensive description of his growth and development are we in a position to give him sound guidance” (Tyler, 1941, p. 493).

Collecting information about students’ performance is ubiquitous in education, as is, for example, shown by the overwhelming number of tests that students encounter during their educational career (as for example, is reflected in the Common Core State Standards Mathematics; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). However, administering standardized tests is but one way for a teacher to know about students’ proficiency in mathematics. Teachers can also acquire insight in students’ mathematical abilities by more qualitative and holistic assessments; for example, observing students in class and giving them open-ended tasks can provide teachers a far more reliable window for knowing their students’ progress (cf. Black, 2014). This is an old insight already asserted by the philosopher John Locke (1693; as cited in Baldwin, 1911) in the 17th century: “[w]e must observe children carefully for ‘favorable seasons of aptitude and inclination’ and teach the child when he is in tune” and by William James (1899), in his *Talks to Teachers*:

Experimental tests [...] can give us useful information only when we combine them with observations made without brass instruments, upon the total demeanor of the measured individual, by teachers with eyes in their heads and common sense, and some feeling for the concrete facts of human nature in their hearts. (p. 136)

Taking this perspective in assessment offers teachers the possibility to adequately assess students’ understanding in such a way that it informs their teaching. Teachers can use the results of these assessments to take instructional decisions; for example, to decide whether they need to adapt instruction to fit to students’ current understanding, repeat a particular exercise or explanation, or, if students have reached a satisfactory insight, continue with further instruction. As Wiliam (2007) put it: “evidence about student learning is used to adjust instruction to better meet students’ needs – in other words, teaching is adaptive to the students’ learning needs” (p. 1065). Assessing with the purpose of making informed decisions about how instruction should be continued is called formative assessment and differs from summative assessment – a distinction first made by Scriven (1967). Summative assessment is meant for making decisions about

certification, namely to decide whether students have met the requirements for obtaining a particular qualification. Although formative assessment can also be carried out with externally developed assessment instruments, this assessment is almost completely in the hands of teachers. Such assessment that is interwoven with instruction and fully integrated in the teachers' daily teaching practice is called classroom assessment<sup>1</sup> (e.g., Black & Wiliam, 1998; Brookhart, 2004; Shepard, 2000; Stiggins & Chappuis, 2005; Wiliam, 2007).

## **1.2 Previous research on the effect of classroom assessment**

In earlier educational research –often large– positive effects of teachers' use of classroom assessment on student achievement have been reported (reviewed in Black & Wiliam, 1998, or more recently in Briggs, Ruiz-Primo, Furtak, Shepard, & Lin, 2012; Kingston & Nash, 2011). Common to the research projects and interventions reviewed is that they were generally small-scale of size and more focused on finding principles for practice (e.g., Wiliam, Lee, Harrison, & Black, 2004) than on clearly establishing the effectiveness of (the different conceptualizations of) classroom assessment. Due to this, in recent reviews (e.g., Bennett, 2011) the comparability of the different approaches used in the reviewed studies has been criticized. Notwithstanding the fact that many authors were referring to classroom or formative assessment when discussing their studies, 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 (see also Veldhuis & Van den Heuvel-Panhuizen, 2014a). What strings these studies together, however, –in addition to the terminology used– is that the interventions are based on teachers' specific subject knowledge and on the frequent use of assessment, allowing the teacher to provide formative feedback to students, meaning “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). So even though these studies' specificities differ, their results point to the effectiveness of the implementation of some forms of classroom assessment.

---

<sup>1</sup> We employ the term *classroom assessment* where other authors prefer formative assessment or assessment for learning. Essentially these terms refer to the same; with that difference that classroom assessment in our view represents the teacher-initiated variety of formative assessment.

This possible relation between the effect of classroom assessment and the frequency of assessment is supported by findings within cognitive psychology, where the use of frequent testing has been found to greatly improve retention of tested knowledge. In their review of this ‘testing effect’, also called ‘retrieval practice’, Roediger and Karpicke (2006) discuss the history and development of evidence for its existence, mainly from laboratory studies, and raise the question of why such a powerful and simple instrument is not more generally applied in education (and educational research). Since they raised this question the testing effect has been investigated in some real classrooms in undergraduate courses (e.g., Butler & Roediger, 2008), middle school classrooms (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011), and more recently in primary school classes (e.g., Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014), however, to our knowledge it has not as such been investigated in primary mathematics education.

To fill this knowledge gap pertaining to the integration of these results from educational and cognitive psychology research on the beneficial effects of regular classroom assessment and feedback, a project was set up investigating the influence on student learning of teachers using classroom assessment in real primary classrooms in the subject domain of mathematics. This project started with two consecutive small-scale studies (Veldhuis & Van den Heuvel-Panhuizen, 2014b, Chapter 3 of this thesis) in which the use and effect of classroom assessment in primary mathematics education was investigated. These studies, in which in total 10 third-grade teachers and 214 students were involved, yielded positive results: teachers were glad to use classroom assessment and found it useful for their teaching practice. The effectiveness was explored with a pretest/posttest design, albeit without a proper control group. Effect sizes for students’ improvement from pretest to posttest were  $d = 0.81$  for the sub-study with five meetings and  $d = 0.55$  for the sub-study with four meetings, as measured by regular standardized student monitoring system tests of mathematics. A control condition was lacking, but the national norm data of the same tests were used for comparison: the effect size of the improvement in this reference sample was  $d = 0.36$ . Apparently students with teachers using classroom assessment improved their scores by between  $(0.55 - 0.36)/0.36 = 53\%$  and  $(0.81 - 0.36)/0.36 = 125\%$  more than would usually be expected. The positive reactions of the teachers about the use of classroom assessment and the relatively strong indication for the effectiveness of it were enough to set up a larger scale study to

further investigate the effectiveness of classroom assessment when compared to a proper control group.

## **2. Research questions and hypotheses**

This led us to formulate this main research question for the current study:

What are the effects of supporting teachers' use of classroom assessment in mathematics on students' mathematics achievement?

As in earlier research the number of professional development sessions in which the classroom teachers participated seemed to play a role in the effectiveness of their use of classroom assessment we included a secondary question: Does the number of workshops on the use of classroom assessment in primary mathematics matter for student achievement? Exploratively, we wanted to also investigate the potential differential effects of the use of classroom assessment: so whether students of different achievement levels had different learning gains following their teachers' use of classroom assessment?

More in particular, the current study aimed to find out whether a professional development program for teachers had an effect on students' mathematics achievement. In this study, the use of classroom assessment is conceived as the use of *classroom assessment techniques* consisting of short teacher-initiated assessment activities that teachers can use in their daily practice to reveal their students' understanding of a particular mathematical concept or skill. These classroom assessment techniques have been used in earlier research in the Netherlands (Veldhuis & Van den Heuvel-Panhuizen, 2014b) and China (Zhao, Van den Heuvel-Panhuizen, & Veldhuis, 2015). Because in the earlier study in the Netherlands it was found that the more attention, exchanges with other teachers, and feedback teachers got during the professional development sessions, the more their students improved their mathematics achievement (see Veldhuis & Van den Heuvel-Panhuizen, 2014b), we expected to find a larger effect on students' mathematics achievement when the number of workshops on the use of classroom assessment was larger.

### 3. Method

#### 3.1 General design

We investigated the research questions in a quasi-experiment with pretest/posttest and control group with third-grade teachers in the Netherlands. There were four different conditions. Apart from a control, business as usual, condition, in which teachers did not partake in any professional development sessions, we had three experimental conditions in which the number of professional development sessions (workshops) varied from one to three (see Table 1 for an overview of the planning of the study). As the influence of the amount of professional development on teachers' classroom practice is closely related to teachers' feeling of ownership of the activities, the involvement, supervision, and discussions of the classroom assessment techniques of teachers in the experimental conditions were expected to lead their students to outperform those in the control condition. For this reason, we experimentally varied the number of professional development sessions between the conditions. Teachers were randomly distributed over the four conditions, with the exception of the few teachers that worked at the same schools; these were grouped in the same condition, so as to not cause interference between conditions.

Table 1  
*Scheme of the General Design of the Study*

Condition	January	February	March	April	June
Control	Pretest				Posttest
1 <sup>st</sup> Experimental	Pretest	Workshop			Posttest
2 <sup>nd</sup> Experimental	Pretest	Workshop		Workshop	Posttest
3 <sup>rd</sup> Experimental	Pretest	Workshop	Workshop	Workshop	Posttest

*Note.* Pretest refers to the regular student-monitoring test mid-Grade 3; Posttest refers to the regular student-monitoring test end-Grade 3

#### 3.2 Participants

##### 3.2.1 Recruitment of teachers

To obtain a sizable and representative sample of primary school teachers and students in the Netherlands we employed the following procedure. Based on our experience in previous studies in this research project, in which we had encountered quite some difficulty in finding teachers willing to participate in

the study, we recruited teachers in a stepwise procedure: from providing little information about the study to finally giving all information (cf. foot-in-the-door technique; Freedman & Fraser, 1966). As a first step we sent an e-mail to all, approximately 7000, primary school directors in the Netherlands (we obtained their e-mail addresses from the Ministry of Education) in March 2013. In this e-mail we simply asked teachers, through their school directors, whether they were interested in receiving more information on a research project on improving classroom assessment in primary mathematics education. This led to 387 positive responses. A few months later, in June 2013, we sent an e-mail to these teachers/directors explaining that teachers would be invited to participate in a number of workshops. For this 79 teachers signed up. To them we sent a questionnaire to find out their availability during the experiment, the textbook they used, and some more background information. In the end 33 classroom teachers reacted on this questionnaire, were available on the dates of the workshops, and all used the textbook *De wereld in getallen* [The world in numbers] (Huitema et al., 2009); these teachers originally formed our sample (i.e. about 8% of teachers who initially reacted). Three of these teachers dropped out during the study: one for health reasons, and the remaining two due to logistical concerns. In the end the final sample contained 30 teachers.

The participating teachers taught 616 third-grade students and worked on 25 different primary schools from all over the Netherlands; just as well from rural parts as from densely populated areas (see Appendix A for their geographical distribution). Of the students there were 286 boys (46.4%) and 268 girls (43.5%) (see Table 2), of the remaining 62 students (10.1%) the gender was unknown as some teachers could not communicate this for privacy reasons.

Table 2

*The Number of Teachers and Students in Four Conditions (Control and Three Experimental) and Some Background Characteristics on Class Composition*

Condition	Number of teachers	Number of 3 <sup>rd</sup> graders	Mean number of 3 <sup>rd</sup> graders	Percentage of combi-classes <sup>a</sup>	Percentages male/female students <sup>b</sup>
Control	6	99	18.8	50.5%	39.4%/37.4%
1 <sup>st</sup> Experimental	6	138	25.3	23.9%	42.8%/55.8%
2 <sup>nd</sup> Experimental	8	172	25.1	5.2%	41.3%/37.2%
3 <sup>rd</sup> Experimental	10	207	23.0	17.4%	56.5%/43.5%
Total	30	616	23.4	20.8%	46.4%/43.5%

<sup>a</sup> Combi-classes are classes with students of more than one grade level, e.g. 3<sup>rd</sup> and 4<sup>th</sup> Grade.

<sup>b</sup> As some teachers did not communicate their students' gender –due to privacy concerns– the sum of these percentages can be less than 100%

### 3.2.1 Recruitment of teachers

Teachers and the complete classes of students of these teachers were randomly distributed over the conditions. Considering that our main focus was on the third experimental group (with three workshops) and since we expected that some teachers that were supposed to participate in three workshops would miss one or more of the sessions, the number of teachers and students in this condition was the largest by far. A few teachers indeed did not attend all workshops that they were supposed to, in the analyses these teachers switched from one experimental condition to another. Six teachers attended two workshops instead of three and three teachers attended just one workshop instead of three. From each experimental condition one teacher dropped out of the study altogether, i.e. three teachers in total. See Table 2 for the final distribution of the thirty teachers and their 616 students over the four conditions, with some background characteristics on class composition. The students were approximately evenly distributed over the conditions, with relatively more students in the third experimental condition (33.6% of the students) than in the control condition (16.7% of the students). After their conditions were determined we made groupings for the workshops based on schools' geographical location, so that teachers from schools that were near each other would participate in the same workshops. The teachers in the different experimental conditions were ignorant of the fact that there were other conditions with more or less workshops.

### **3.3 Material**

#### *3.3.1 Measures*

As pretest data we used students' results on the midyear student monitoring system test for Grade 3 (Cito LVS; Janssen, Scheltens, & Kraemer, 2010). Data from the end of year test for Grade 3 from this same student monitoring system were used as posttest data. In the Netherlands, these tests are used in virtually all primary schools to monitor students' development in mathematical ability over the years. Teachers administered the tests, as per usual, in their own classes. The midyear test contains 58 items (administered in two parts), 6 of which are multiple-choice questions and the remaining 52, closed student-constructed response items. The end of year test has 80 items (administered in three parts) and contains 1 multiple-choice item. The items are on the different subject domains 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 tests' scores are mathematical ability scores calculated through item response theory models. In addition to the ability scores also student performance level indicators are provided. Students' ability scores are divided over five performance levels; these levels each contain 20% of the norm sample students. As such, the lowest performing 20% are in level V, the average performing students in level III and the top-20% in level I. These level indicators allow teachers to categorize students' mathematics ability scores. At the end of the study teachers sent their students mathematics ability scores to the first author.

In addition to these quantitative data the first author also observed one mathematics lesson of each participating teacher. This was done for two main reasons: firstly to assist teachers in implementing the classroom assessment techniques in their own classroom, and secondly to ensure that teachers really used the classroom assessment techniques and whether they did this as intended. Of course only one instance of observation per teacher is not enough to guarantee the fidelity of the implementation, but at least it gave an impression of how the teachers implemented the classroom assessment techniques in their teaching practice. Teachers also provided feedback forms for every classroom assessment technique they used. The comments on these forms were generally comparable to those of the teachers in Veldhuis and Van den Heuvel-Panhuizen (2014b).

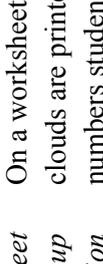
*3.3.2 Classroom assessment techniques*

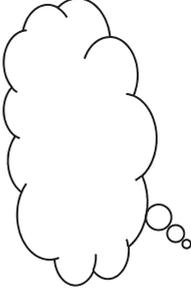
In total nine classroom assessment techniques were suggested to the teachers. These assessment techniques typically blur the divide between instruction and assessment, are low-tech and low-cost, and can be feasibly implemented by teachers. Every technique has a particular format and is focused on particular content, helps teachers to quickly find out something about their students' mathematics comprehension, provides indications for further teaching, and consists of a short activity (less than 10 minutes). The focus of these assessment techniques was on the mathematics curriculum in the second half of Grade 3, meaning that they were mostly centred on the domain of number and the ability to understand word problems. The content, format, and goal of the different classroom assessment techniques are described in detail in Table 3.

Table 3

*Descriptions of the Classroom Assessment Techniques for Mathematics in Grade 3*

<b>Title</b>	<b>Format</b>	<b>Description</b>	<b>Example</b>	<b>Goal</b>
<b>1a</b>	<i>Red and green cards</i>	The teacher calls a series of number pairs and asks: “Do <i>these numbers cross</i> 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.
<b>1b</b>	“	The teacher asks: “ <i>Is the difference bigger than 5/10/50/100?</i> ” 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.

Title	Format	Description	Example	Goal
1c	"	The teacher asks: "Is [a series of numbers] in the table of 4?". This question is asked for various multiplication tables.	<p>"Is 32 in the table of 8?" *Cards*</p> <p>"Is 44 in the table of 8" *Cards*</p>	Assess whether students have ready knowledge of the multiplication/division tables.
2	<i>Worksheet and group discussion</i>	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.	<p><i>Which is easiest to you?</i></p> <p>"11 + 2 or 13 + 12"</p> <p>"26 – 7 or 35 – 4"</p>	Assess whether students are aware of the difficulties some number operations contain and whether they can reflect on these difficulties.
3a	<i>Worksheet and group discussion</i>	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 equalling 10/100/1000 and see the analogy with other numbers.

Title	Format	Description	Example	Goal
3b	<p><b>Make your own clouds</b></p> <p><i>Empty worksheet for own production</i></p>	<p>On a worksheet on which empty clouds are printed students can fill in pairs or another number of numbers that add up to 10, 100 or 1000, or a different number. Exchange work with peer and discuss differences in approach.</p>		<p>Assess the insight of students in the combinations of numbers equalling 10/100/1000 and creating exercises.</p>
4	<p><b>Word problem problems</b></p> <p><i>Classroom experiment</i></p>	<p>Students solve a series of problems first in word problem format, then the same problem but as a bare number problem. After class teacher compares student work on the different presentational formats.</p>	<p><i>"Charly saved 680 euro, a computer costs 1000; how much does he still need?"</i></p> <p><i>"1000 – 680 = "</i></p>	<p>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.</p>

<b>Title</b>	<b>Format</b>	<b>Description</b>	<b>Example</b>	<b>Goal</b>
<b>5</b>	<b>What could have been the question?</b> <i>Worksheet (for own production)</i>	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 x 16" 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.
<b>6</b>	<b>Find your errors and correct them</b>	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.

### **3.4 Procedure**

The teachers in the experimental conditions discussed the same classroom assessment techniques in the workshops that were led by the first author. Teachers in the conditions with two or three workshops also reflected upon the use of various classroom assessment techniques in consecutive workshops, whereas teachers in the condition with just one workshop did not. Except for this, the content and procedure of the different workshops were identical: first the important mathematics content of the following weeks (or months, depending on the condition) was discussed. Teachers and the first author would discuss this for some time in order to clearly articulate the turning points in mathematics of this period of time. These discussions would generally revolve around the domains of mental arithmetic, such as crossing ten in addition or subtraction, knowledge of multiplication or division tables, and understanding word problems. Teachers in all the workshops identified these as the main obstacles in the second half of Grade 3. After having identified and discussed these subject domains, the discussion would turn to their assessment. Particularly, the ways in which they can be assessed providing teachers with specific knowledge about students' understanding in these domains, were discussed. This led to the first author providing examples of the classroom assessment techniques and the illustration of their use. The teachers used the classroom assessment techniques in their classes at moments that they considered them to be useful, which generally came down to using every technique once or twice. The workshops were held at the schools of the participating teachers.

### **3.5 Analyses**

The pretest and posttest ability score data (and level scores) were analyzed descriptively ( $M$ ,  $SD$ , and correlations) generally and per condition. Then we investigated whether there was an intervention effect in the different conditions using an analysis of covariance (ANCOVA). In this ANCOVA we included the pretest as covariate, the posttest as dependent variable, and the four different conditions as fixed factors (results were analogous whether using gain scores or analysis of covariance). In order to control for the higher relatedness of

students' scores<sup>1</sup> from the same classes and teachers, we used a procedure based on clustered robust standard errors (e.g., Angrist & Pischke, 2008; Cameron & Miller, in press). This approach adapts the standard errors that are estimated in the normal linear model (e.g., ANCOVA) after the estimation is completed, to account for cluster-specific variance. This is necessary because with data from clustered sampling an analysis approach neglecting the clustering of the data provides underestimations of the standard errors; using cluster robust standard errors corrects for this underestimation.

## 4. Results

### 4.1 Preliminary analyses

We used chi-squared tests (for percentage data) and univariate analyses of variance (to compare means) to check whether students in the different conditions were comparable regarding student or class characteristics, such as student gender or number of students per class, and their pretest scores. The conditions differed significantly on the percentage of combination classes ( $p < .001$ ; Control: 50.5%, 1<sup>st</sup> Experimental: 23.9%, 2<sup>nd</sup> Experimental: 5.2%, and 3<sup>rd</sup> Experimental: 17.4%), with all pairwise comparisons being significant, except for the comparison of first experimental and third experimental condition. The conditions did not differ significantly on gender distribution ( $p = .125$ ; Control: 51.3%, 1<sup>st</sup> Experimental: 43.4%, 2<sup>nd</sup> Experimental: 52.6%, and 3<sup>rd</sup> Experimental: 56.5% male). Also, students' pretest scores did not differ significantly between the conditions ( $F(3,611) = 2.196, p = .087$ ). The number of third-grade students per class was significantly different between the conditions ( $F(3,616) = 24.512, p < .001$ ; Control 18.8, 1<sup>st</sup> Experimental: 25.3, 2<sup>nd</sup> Experimental: 25.1, and 3<sup>rd</sup> Experimental: 23.0 students), post hoc tests

---

<sup>1</sup> Another approach is hierarchical linear modeling or multilevel modeling (see e.g. Hox, 2010) that allows one to take into account the nested structure of the data. However, as our hypotheses are on the individual students' level and not on class level – we expect students to benefit from their teachers' use of classroom assessment techniques – multilevel modeling is not necessarily called for. Notwithstanding this, we included the results of multilevel modeling of our data in Appendix B. The conclusions are comparable, as the students from the third experimental group significantly outperform the students from the other conditions. The main effect for condition is only marginally significant; however, the direction and approximate size of the effects remain the same.

showed that this number differed between all conditions except for the pairwise comparison between the first and third experimental condition. A descriptive overview of the scores on pretest and posttest of the participating students is shown in Table 4.

Table 4

*Descriptive Statistics of Students' Mathematics Ability Scores per Condition for Pretest and Posttest*

Condition	Pretest ability score			Posttest ability score			Gain <sup>a</sup>	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>		
Control	70.0	15.3	97	75.9	14.8	98	+5.9	0.39
1 <sup>st</sup> Experimental	74.5	11.5	137	79.9	11.9	136	+5.4	0.47
2 <sup>nd</sup> Experimental	72.1	14.4	170	77.5	13.1	171	+5.4	0.38
3 <sup>rd</sup> Experimental	71.4	14.6	207	79.5	13.2	206	+8.1	0.55
Total	72.0	14.1	611	78.4	13.3	611	+6.4	0.45
National norm	69.2			74.8			+5.6	0.36

<sup>a</sup> Gain is the difference between the score on the pretest and the posttest (i.e. posttest score–pretest score)

From the descriptive statistics in Table 4, it becomes clear that students in all conditions improved their mathematics achievement scores from pretest to posttest (overall gain = +6.4,  $d = 0.45$ ) and that this is slightly more than the expected improvement, as found in the original norm sample (+5.6,  $d = 0.36$ ). Students in the control condition seemed to improve slightly more (+5.9 points) than those in the first (+5.4) and second (+5.4) experimental conditions, although students in the third experimental condition improved the most by far (+8.1,  $d = 0.55$ ). The Pearson's correlation between pre- and posttest scores was  $r = .820$  ( $p < .001$ ) for all students. For the different conditions these correlations were also high: Control:  $r = .731$ , 1<sup>st</sup> Experimental:  $r = .811$ , 2<sup>nd</sup> Experimental:  $r = .873$ , and 3<sup>rd</sup> Experimental:  $r = .842$  (all  $ps < .001$ ).

If we look at another indicator for student performance the monitoring system tests provide, the level indicator, an even clearer picture arises. The level indicators for students in the four conditions are displayed in Table 5. The mean performance level of students in a class is by definition not expected to change over time, nonetheless, in all conditions students improved slightly from pretest

to posttest (average level gain = -0.09); as I is the highest level and V the lowest, an improvement in level score is reflected in a negative gain. Between the conditions a remarkable difference appears, students in the third experimental condition improved more than twice as much as the others (average level gain 3<sup>rd</sup> Experimental = -0.24).

Table 5

*Performance Level Indicators per Condition for Pretest and Posttest*

Condition	Pretest level			Posttest level			Gain <sup>a</sup>
	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	
Control	2.87	1.33	97	2.80	1.38	96	-0.07
1 <sup>st</sup> Experimental	2.44	1.25	137	2.37	1.22	135	-0.07
2 <sup>nd</sup> Experimental	2.57	1.38	170	2.61	1.32	169	-0.06
3 <sup>rd</sup> Experimental	2.71	1.34	207	2.47	1.31	206	-0.24
Total	2.63	1.33	611	2.54	1.31	606	-0.09
National norm	3			3			+/- 0

<sup>a</sup> Gain is the difference in level on the pretest and the posttest

## 4.2 Explorative analyses

To explore possible differential effects of the use of the classroom assessment techniques between the students of different mathematics performance levels we identified high- and low-achieving students at the pretest. We identified high-achieving students as those with level scores indicating they had scores comparable to the top 40% and low-achieving students as the bottom 40%. Regarding the score gains of these different groups (Table 6), one can easily spot that the only condition in which students of all levels of performance improved, was the third experimental condition. Remarkably, the level gain of students from the third experimental condition was the smaller than that of the students in the control condition for the average achieving students. Furthermore, students that we identified as low-achieving at pretest improved most; it seems that a particular regression-to-the-mean effect was happening, where low-achieving students on pretest make the biggest achievement gains (score +10.25, level -0.48) and some high-achieving students regress or do not improve (score +4.85, level +0.09); in the third experimental condition, only high-achieving students improved their level score (-0.10).

Table 6  
*Score and Performance Level Gains for the Different Performance Groups Categorized by Pretest Performance*

Condition	Score gain		Level gain	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
High performers (top 40%, <i>n</i> = 333)	4.85	7.73	0.09	0.73
Control	2.98	11.01	0.23	1.02
1 <sup>st</sup> Experimental	4.45	7.50	0.14	0.69
2 <sup>nd</sup> Experimental	3.94	6.16	0.19	0.72
3 <sup>rd</sup> Experimental	6.72	7.37	-0.10	0.59
Average performers (mid 20 %, <i>n</i> = 105)	6.15	7.02	-0.10	0.86
Control	8.14	9.39	-0.27	0.94
1 <sup>st</sup> Experimental	6.31	5.07	-0.12	0.77
2 <sup>nd</sup> Experimental	5.13	6.06	0.00	0.74
3 <sup>rd</sup> Experimental	5.44	7.18	-0.03	0.97
Low performers (low 40%, <i>n</i> = 173)	10.25	8.56	-0.49	0.82
Control	10.61	10.90	-0.48	1.03
1 <sup>st</sup> Experimental	7.03	7.89	-0.52	0.78
2 <sup>nd</sup> Experimental	9.49	7.74	-0.32	0.66
3 <sup>rd</sup> Experimental	12.08	7.81	-0.59	0.83

*Note.* Improvement in performance level is indicated by a negative value for the level gain, as Level I is the highest and Level V the lowest level.

### 4.3 Main analyses

In addition to these descriptive analyses, we performed an analysis of covariance (ANCOVA) in R (R Core Team, 2014). In this ANCOVA the pretest score was entered as covariate and condition as fixed factor. This model explained 67.9% of the overall variance in the posttest scores ( $R^2 = .679$ ). We used Arai's (2011) function to calculate the cluster robust standard errors. We built the model step-by-step including explanatory variables one-by-one in the analysis of covariance. All models showed significant main effects for pretest

score ( $ps < .001$ ) and condition ( $ps < .01$ ); all other predictors, such as student gender, number of students per class, combination class, and teacher gender were, however, not significant. Including the (non-significant) explanatory variables only very slightly improved the amount of explained variance of the model, from 68% to 71%. Therefore we selected the simplest model, as shown in Table 7; with posttest score as outcome variable, pretest score as covariate, and condition as fixed factor with four levels (Control, 1<sup>st</sup> Experimental, 2<sup>nd</sup> Experimental, and 3<sup>rd</sup> Experimental).

Table 7

*Results of the ANCOVA with Clustered Robust Standard Errors to Estimate Effect of Condition ( $R^2 = .68$ )*

Variables	Estimates				
	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>	95% CI
Intercept	22.81	2.22	10.26	< .001	[18.37, 27.25]
Pretest	0.77	0.02	32.50	< .001	[ 0.73, 0.81]
1 <sup>st</sup> Experimental	0.69	1.14	0.60	.547	[-1.59, 2.97]
2 <sup>nd</sup> Experimental	-0.74	0.71	-1.03	.302	[-2.16, 0.68]
3 <sup>rd</sup> Experimental	1.57	0.68	2.30	.022	[ 0.21, 2.93]

*Note.* The baseline condition was the Control condition, CI = Confidence interval

We found a small significant main effect for condition ( $F(3, 601) = 3.8, p = .01, \eta_p^2 = 0.02$ ) and a large significant effect for pretest score ( $F(1, 601) = 1261.2, p < .001, \eta_p^2 = 0.676$ ). Post-hoc comparison showed that students in the third experimental condition scored significantly higher on the posttest corrected for pretest differences ( $M_{corr} = 80.02, d = 0.255$ ) than students in all other conditions ( $Ms_{corr} < 78.13, ps < .04$ ; see Figure 1, Control = 78.1, 1<sup>st</sup> Experimental = 78.0, 2<sup>nd</sup> Experimental = 77.7).

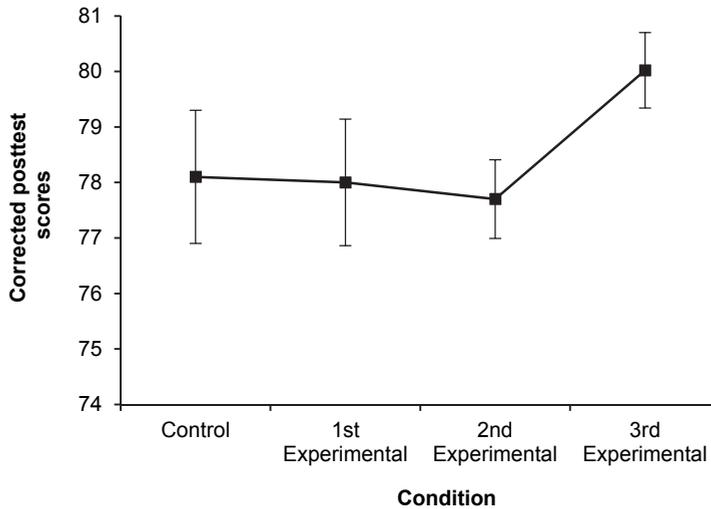


Figure 1. Results of the ANCOVA: Posttest scores corrected for pretest differences for the four conditions with error bars.

## 5. Discussion

### 5.1 Main conclusions

In this study we investigated the effects of supporting teachers' use of classroom assessment, more particularly whether their participation in workshops on the use of classroom assessment techniques led to a significant larger improvement of their students' mathematics achievement. Students of which the teachers participated in three one-hour workshops on classroom assessment techniques had a significantly higher score gain than students from the business-as-usual control condition as measured by a standardized mathematics ability test. Contrary to our expectations, students from the two other experimental conditions had comparable score gains to the students in the control condition; so the changes in these teachers' practice did not alter students' improvement from usual maturation and schooling. Taking into account the clustered nature of the data, by adjusting the standard errors or carrying out a multilevel analysis, only confirmed these conclusions. Apparently teachers need to have followed at least three workshops for their students to clearly benefit from teachers' use of classroom assessment techniques. In this condition, 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 & Koch, 2014, p. 283), which was the biggest difference with the other conditions. In earlier research on the classroom assessment techniques it was found that the effect size of the improvement, compared to the pretest, remained considerable even as it decreased from five ( $d = 0.81$ ) to four ( $d = 0.55$ ) meetings (cf. Veldhuis & Van den Heuvel-Panhuizen, 2014b). The effect size of three meetings as found in the current study ( $d = 0.55$ ) was comparable to that of four meetings and provides support for that at least three teacher workshops on the use of classroom assessment techniques are necessary to find an effect on student achievement.

Our exploratory descriptive analysis of students with different performance levels provided us with insight on the differential influence of teachers’ use of the classroom assessment techniques. Since the classroom assessment techniques were focused on quite elementary mathematical topics (number knowledge, multiplication tables, understanding of word problems), it could be expected that teachers’ use of the techniques would mainly benefit their lower achieving students, since the higher achieving students had already mastered these skills. A first glance at Table 5 with the performance level changes from pretest to posttest of students with the different levels confirms this expectation. Low-achieving students improved the most by far whereas the levels of average and high-achieving students slightly decreased. This also contradicts results from an earlier study on a formative assessment intervention for algebra of which students with higher pretest scores tended to benefit more (Phelan, Choi, Vendlinski, Baker, & Herman, 2011). However, upon closer inspection of the pattern we found in our study, it was quite different for students of teachers from the third experimental condition: in this condition high-achieving students also improved their performance level. This counters the possible expectation about the classroom assessment techniques being mainly useful to gather information about the lower achieving students. A possible explanation for this result would be that the high-achieving students generally are not bothered with questions about, for example, number knowledge, because teachers assign them the more difficult or challenging exercises. Through the use of the classroom assessment techniques focusing on specific and important key understandings in number knowledge, the teachers and their high-achieving students could have

become aware of some weaknesses these students had, that had otherwise been masked by their high achievements or good performance on other tasks.

When reflecting upon the mechanisms that might have caused the effectiveness of teachers' use of the classroom assessment techniques, parallels emerged between the functioning of classroom assessment and the testing effect. Both make use of frequent assessments (such as *quizzing*, see McDaniel et al., 2011 or Wiliam, 2011) that require students to actively process knowledge, making use of retrieval processes, generally followed, or directly accompanied, by feedback. All these assessment processes have been demonstrated to positively influence learning, directly through the retrieval process (e.g., Roediger & Pyc, 2012) or indirectly through the provided feedback focusing on closing the gap between the current level and the desired level of performance (cf. Ramaprasad, 1984) as such improving students' metacognitive judgments and teachers' instructional decisions (e.g., Hattie & Timperley, 2007). Essentially, the classroom assessment techniques in our study combine the benefits of the testing effect with that of providing feedback and the more detailed knowledge the teacher has about students, allowing to adapt his or her instruction to students' needs. This could be an explanation for the beneficial effect on student learning.

## **5.2 Limitations**

Naturally, there are some limitations to our study. Relating to the testing effect mentioned just before, it has been found to be most effective when multiple tests on the same subject were used, whereas, in our study, most of the teachers admitted to only using the same technique once or twice. Additionally the focus of the classroom assessment techniques used in our study, mostly on one subdomain of mathematics at a time, might have negatively attenuated their influence on student performance. Since studies on practicing have shown that variation in subject domains can have a positive effect on student achievement (e.g., Rohrer, Dedrick, & Stershic, 2014), this could also apply to assessment. Integrating different domains in the used assessment techniques, might give teachers more comprehensive understanding of the students' learning process. To give the teacher a broad scope on students' learning, in at least one domain, the assessment techniques did contain variation in aspects of number knowledge. For example, simple addition and subtraction problems were mixed

up in a technique (cf. Table 3, technique 2) in an effort to identify students' insight in degrees of difficulty of particular problems (cf. Van den Heuvel-Panhuizen, Middleton, & Streefland, 1995).

Another point of contention of the reported effects could be teachers' motivation in the project, as they all volunteered to participate. Only 8% of the teachers that expressed interest in the earliest stage participated in the final experiment. To control for this self-selection of the participating teachers they were randomly distributed over the conditions as such levelling out this potential pre-existing motivational stance and isolating the effect of workshops on the use of classroom assessment techniques. In this way it could still be that our sample of teachers was not representative of the Dutch population of teachers, and for example, overly motivated, but it ensures that the different effects we found between the conditions were probably due to the intervention. It could of course be that the few teachers, who participated in fewer workshops than they were supposed to and as such switched experimental conditions, 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 as such helps to explain their students' improvement being comparable to usual. However, most occurrences of nonattendance were due to external reasons, such as sick children or an administrative matter to attend to, so the motivational explanation for the smaller effect on students in these conditions does not hold ground.

A related issue for which we tried to control in our design was the so-called Hawthorne-effect. The only way to completely exclude the possibility of this effect is the use of a pseudo-intervention and this was not deemed practically or ethically feasible in our study. Nonetheless, the experimentally varied frequency of the workshops in the different experimental conditions was a way to remediate possible tendencies of teachers and students to change their behavior due to merely participating in a research project. If just knowing that they were participating in research had effectively an influence than one would expect this influence to gradually increase hand-in-hand with the increasing number of workshops. As we only found an effect in the third experimental group, apparently the Hawthorne-effect was not in play in our study.

Also, 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. Surely between the pretest and the posttest five months had passed, but it would be interesting to see if these beneficial effects of teachers' use of classroom assessment techniques continue to show in student achievement in the ensuing school years. 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.

Of note as well was that the effect of students' pretest scores on their posttest scores was much larger than that of the teacher's condition. This stands to reason given that the consecutive student monitoring tests are supposedly measured on a common continuous latent ability scale (e.g., Janssen, Verhelst, Engelen, & Scheltens, 2010). Students' mathematics proficiency improves quite logically throughout their schooling, so those that are in the higher regions of proficiency in January are most likely to score high on the test in June as well (cf. the Matthew effect, e.g., Merton, 1968).

### **5.3 Implications**

Openness and adaptability of the implementation of an educational intervention have been found to improve teachers' feeling of ownership of and involvement in the intervention (cf. Suurtamm & Koch, 2014). Teachers could freely adapt the assessment techniques to fit their practice and use them at opportune –for them– moments in their classrooms. The quality of the implementation of the classroom assessment techniques by the participating teachers could not reliably be verified, as we had only one instance of classroom observation per teacher. In earlier research it was 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., Black & Wiliam, 1998). In our study, the fact that, at least the teachers in the third experimental condition had clearly become owners of the classroom assessment techniques, shown by their reflections in the consecutive workshops on what they had learned from them, brings us to hypothesize that the teachers faithfully implemented the techniques in their practice. Here, faithfully does not mean 'exactly as they had been proposed to the teachers' but conveys the spirit of the assessment techniques: that they are used to gather worthwhile information on students' mathematics

skills and knowledge and in the hands of the teachers become part of their educational practice. Still, for further research and theory development, it would be valuable to investigate the specific factors that influenced teachers' implementation of these classroom assessment techniques in their mathematics teaching practice, and their relation with student performance.

Another direction for future research would be to investigate the use of such classroom assessment techniques in different grade levels of primary school. Most of the techniques can quite easily be adapted to fit mathematics skills and understanding students are acquiring, or supposed to acquire, in earlier or subsequent grades. The adaptation of the classroom assessment techniques used in the current study for use in a different cultural and educational context is also a road to pursue. Some first experiences with this in Chinese primary school mathematics classrooms (cf. Zhao et al., 2015), showed that a mere translation of the techniques in a different language was not possible; a careful analysis of the curriculum, mathematics teachers' practices, and the used textbooks were first required. Even after these analyses and subsequent adaptations, the use of the classroom assessment techniques demanded an important change in teachers' and students' learning culture (cf. Shepard, 2000), which was not easily accomplished. Investigations into the implementation of the classroom assessment techniques in other contexts or cultures can provide us with more information on their functioning in general.

Bearing the results of this study in mind a probable chain of events to explain the effects would be that teachers participating in at least three workshops pay more attention to the use of the techniques, through the mere fact of attending three meetings with other teachers and being asked to reflect on their use of the techniques. By having these reflective discussions teachers developed more ownership of the classroom assessment techniques and as such were positively inclined to use them and the information gathered by them in their further teaching. This form of teacher learning community is far from new (e.g., Lave & Wenger, 1991) and has also been advocated to use with teachers developing their classroom assessment skills (e.g., Suurtamm & Koch, 2014; Wiliam et al., 2004). However, in its current form, with merely three meetings on the use of classroom assessment techniques, and the relatively high effect these have on student mathematics achievement; clearly suggest continuing investigating

these techniques and assist teachers in using them. The awareness of students' mathematics skills and knowledge teachers develop while using the classroom assessment techniques 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 & Boerst, 2012, p. 1039). Let us finish this discussion by reminding that knowing what students know is quintessential in teachers' practice, as already illustrated by the following citation of Dewey (1904):

Only in this way can the most essential trait of the mental habit of the teacher be secured--that habit which looks upon the internal, not upon the external; which sees that the important function of the teacher is direction of the mental movement of the student, and that *the mental movement must be known before it can be directed*. (p. 262, emphasis added)

The classroom assessment techniques used in the current study in 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.

## References

- Anderson, J. R., Reder, L. M., & Simon, H. A. (2000). Applications and misapplications of cognitive psychology to mathematics education. *Texas Educational Review*.  
Retrieved from <http://act-r.psy.cmu.edu/papers/misapplied.html>
- Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press
- Arai, M. (2011, January 31). *Cluster-robust standard errors using R*.  
Retrieved from <http://people.su.se/~ma/clustering.pdf>
- Baldwin, B. T. (1911). William James' contributions to education. *Journal of Educational Psychology*, 2(7), 369-382.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5-25.
- Black, P. (2014). Assessment and the aims of the curriculum: an explorer's journey. *Prospects*, 44(4), 487-501.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education: Principles, Policy & Practice*, 5(1), 7-74.
- Butler, A. C., & Roediger, H. L., III (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, 36(3), 604-616.
- 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.
- Brookhart, S. (2004). Classroom assessment: tensions and intersections in theory and practice. *Teachers College Record*, 106(3), 429-458.
- Cameron, A. C. & Miller, D. L. (in press). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*. Retrieved from [http://cameron.econ.ucdavis.edu/research/Cameron\\_Miller\\_JHR\\_2014\\_July\\_09.pdf](http://cameron.econ.ucdavis.edu/research/Cameron_Miller_JHR_2014_July_09.pdf)
- 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.
- Freedman, J. L. & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4(2), 195-202.

- Gearhart, M., & Saxe, G. B. (2005). When teachers know what students know. *Theory Into Practice*, 43(4), 304-313.
- Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, 3(3), 177-182.
- Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.
- Hox, J. (2010). *Multilevel analysis. Techniques and applications*, 2<sup>nd</sup> Edition. New York: Routledge.
- Huitema, S., Erich, L., Hijum, R. van, Wetering, M. van de, et al. (2009). *De wereld in getallen – vierde editie* [The world in numbers - fourth edition]. Den Bosch: Malmberg.
- James, W. (1899). *Talks to teachers on psychology: and to students on some of life's ideals*. New York: Henry Holt.
- 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, Netherlands: 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, Netherlands: Cito.
- 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 University Press.
- Locke, J. (1693). *Some thoughts concerning education*. London: Churchill.
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103(2), 399-414.
- Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810), 56-63.

- Moreland, J., Jones A., & Northover, A. (2001). Enhancing teachers' technological knowledge and assessment practices to enhance student learning in technology: A two-year classroom study. *Research in Science Education, 31*(1), 155–176.
- National Governors Association Center for Best Practices & Council of Chief State School Officers (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors.
- Phelan, J., Choi, K., Vendlinski, T., Baker, E., & Herman, J. (2011). Differential improvement in student understanding of mathematical principles following formative assessment intervention. *The Journal of Educational Research, 104*(5), 330-339
- R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4-13.
- Roediger, H. L., III & Karpicke, J. (2006). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.
- Roediger, H. L., III & Pyc, M. A. (2012). Inexpensive techniques to improve education: applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248.
- Rohrer, D., Dedrick, R. F., & Stershic, S. (2014). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*. Advance online publication, <http://psycnet.apa.org/doi/10.1037/edu0000001>
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, M. Scriven (Eds.), *AERA monograph series on curriculum evaluation Vol. 1 -Perspectives on curriculum evaluation* (pp. 39-83). Chicago: Rand McNally.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.
- Shute, V. J. (2007). *Focus on formative feedback*. Princeton, NJ: Educational Testing Service.
- Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427-515.

- 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.
- 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.
- Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research, 35*(7), 492-501.
- 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. (2014a). Teachers' assessment profiles in primary school mathematics education. *PLoS ONE, 9*(1), e86817.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014b). Exploring the feasibility and effectiveness of assessment techniques to improve student learning in primary mathematics education. In C. Nicol, S. Oesterle, P. Liljedahl & D. Allan (Eds.), *Proceedings of the joint meeting of PME 38 and PME-NA 36, Vol. 5* (pp. 329-336). Vancouver, BC: PME.
- William, 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, CT: Information Age Publishing.
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3-14.
- William, 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.
- Zhao, X., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (2015). *Chinese teachers' use of classroom assessment techniques in primary mathematics education*. Manuscript submitted for publication.

**Appendix A: Geographical distribution of the participants**



*Figure A.* Geographical distribution of the participants over the Netherlands (every cross represents a school)

**Appendix B: Results from the multilevel analyses**

We used R (R Core Team, 2014) package lme4 (Bates, Maechler, Bolker & Walker, 2014) to perform a linear mixed effects analysis of the relationship between posttest mathematics achievement scores and the professional development in classroom assessment techniques. We first estimated an unconditional model in order to check the intraclass coefficient ICC(1) and decide whether multilevel modelling was called for. If the ICC(1) of the unconditional model is sufficiently close to zero, it would imply that students' posttest scores are independent of the classes they are in, if however it is considerably larger than zero, students' score variation can be attributed to their nesting in classes (cf. Krull & MacKinnon, 2001). In Table A the results of the different models we estimated are displayed. The unconditional model gave an ICC(1) of 0.153 (calculated as the variance due to clustering on the class level divided by the total variance, i.e. the clustering variance plus the residual variance). This implies that 15.3% of the variance in the posttest scores is accounted for by clustering of students in classes (i.e., between-classes differences). Model fit statistics for this unconditional model were: loglikelihood = -2421.6, AIC = 4849.3, and BIC = 4862.5. To this model we added, analogously to the ANCOVA, pretest score as a student level predictor/covariate, and condition as a fixed factor. This model (loglikelihood = -2071.0, AIC = 4156.0, and BIC = 4186.8) fitted significantly better than the unconditional model ( $\chi^2(df = 4) = 701.28, p < .001$ ). However, this model, with the condition as fixed factor, did not fit significantly (albeit 'marginally') better ( $\chi^2(df = 3) = 7.00, p = .072$ ) than the model that only included students' pretest score as predictor. The pretest-model (without condition as fixed factor) had the following model fit statistics: loglikelihood = -2074.5, AIC = 4157.0, and BIC = 4174.6. This can be seen as indicating that there was not a significant effect for condition (albeit a 'marginal' one).

Table A  
*Multilevel Modelling Results*

Model	Random factors		Fixed factors						R <sup>2</sup>
	Icpt	Res	Intercept		Condition		Pretest		
			B	SE	B	SE	B	SE	
Unconditional	5.22	12.28	78.14	1.09					0.19
Pretest-only	1.58	7.30	25.27	1.65			0.75	0.02	0.69
Pretest and condition	1.39	7.31	23.08	1.76	0.62	0.36	0.76	0.02	0.69

### References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7, <http://CRAN.R-project.org/package=lme4>.
- Krull, J.L., & MacKinnon, D.P. (2001) Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36(2), 249-277.

## Chapter 5

# Primary school teachers' assessment profiles in mathematics education

Michiel Veldhuis<sup>a</sup> & Marja van den Heuvel-Panhuizen<sup>a,b</sup>

*<sup>a</sup>Freudenthal Institute for Science and Mathematics Education, Utrecht University, <sup>b</sup>Department of Pedagogical and Educational Sciences, Utrecht University*

Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). Primary school teachers' assessment profiles in mathematics education. *PLoS ONE*, 9(1), e86817.

## **Primary school teachers' assessment profiles in mathematics education**

### **Abstract**

The aim of this study was to contribute to knowledge about classroom assessment by identifying profiles of teachers' assessment of their students' understanding of mathematics. For carrying out this study we used data of a nationwide teacher survey ( $N = 960$ ) in the Netherlands. The data were collected by an online questionnaire. Through exploratory factor analyses the underlying structure of what is measured by this questionnaire was uncovered as consisting of five factors: Goal centeredness of assessment, Authentic nature of assessment, Perceived usefulness of assessment, Diversity of assessment problem format, and Allocated importance of assessing skills and knowledge. By using a latent class analysis four different assessment profiles of teachers were identified: Enthusiastic assessors, Mainstream assessors, Non-enthusiastic assessors, and Alternative assessors. The findings suggest that teachers with particular assessment profiles have qualitatively different assessment practices. The paper concludes with discussing theoretical implications of these assessment profiles and indications these profiles can offer both for designing material for professional development in classroom assessment and for evaluating changes in teachers' classroom assessment practice.

*Keywords:* Classroom assessment; mathematics education; questionnaire research; latent class analysis; assessment profiles.

## 1. Introduction

Classroom assessment is crucial for students' learning (Cizek, 2010). A main reason for this is that through classroom assessment teachers can gather information on their students' skills and level of understanding to make decisions about further instruction. Based on this information teachers can adapt their teaching to their students' needs and create an ideal learning environment for them in their classroom. Therefore, the use of classroom assessment as an integrative part of education has been named as one of the most important activities for teachers to improve student achievement (e.g., Black & Wiliam, 1998a).

Consequently, gaining knowledge about classroom assessment has high priority in educational research. The better we know how the individual teacher carries out the collection of data on students' learning, the more we are able to optimize this process. Contributing to this knowledge was the aim of this study. Our focus was on classroom assessment in primary school mathematics education.

To realize this aim we built on a previous study which investigated how primary school teachers in the Netherlands collect information about their students' progress in mathematics (see Veldhuis, Van den Heuvel-Panhuizen, Vermeulen, & Eggen, 2013 – or Chapter 2 in this thesis). The data for this earlier study were collected by means of an online questionnaire. The prior analysis of these data gave a general overview of how often Dutch primary school teachers are using particular assessment methods, the purposes they are assessing for, and the teachers' perceived usefulness of these assessment methods, and the relations between assessment methods, purposes, and perceived usefulness. In addition to these overall findings, the present study was aimed at gaining knowledge of how the assessment practices of individual teachers can be characterized within the universe of assessment skills and activities. In fact, in this study, we wanted to understand assessment from the conglomerate of choices a single teacher is making when collecting information about his or her students' learning process. To achieve this we performed a secondary analysis of the earlier gathered questionnaire data to identify a profile characterization of every teacher's assessment practice. The rationale for distinguishing assessment profiles of teachers is that these can contribute to our theoretical understanding of assessment as it is carried out by teachers. In addition, knowledge about these

assessment profiles can help us in a practical sense with designing tailor-made courses for professional development that fit the teachers' needs. Furthermore, these assessment profiles can provide us with a tool to measure changes in teachers' classroom assessment practice.

### **1.1 Theoretical background: A classroom assessment theory?**

A scientific theory of any given process generally consists of a description of the constituting components, the causal mechanisms that govern these components, information about factors influencing all of these, and implications for practice. In the end, for further theory building, it is necessary that observational consequences of a theory are tested.

With respect to classroom assessment in mathematics education, many scholars have proposed tentative theories of classroom assessment. As such a variety of conceptualizations exists of what assessment in mathematics education is, and entails, which have abundantly been investigated and discussed. Generally, the skills teachers need to have in order to perform various assessment activities are part of these conceptualizations. Some go a bit further and also describe conceptual models integrating theoretical concepts and practices. However, the descriptions rarely surpass a mere listing of concepts related to assessment. In any case, testing a proposed theory about assessment is certainly not something that is frequently done.

To illustrate the great variety of approaches and methods describing teachers' specific assessment skills and activities, and, more generally, models of assessment, we give a brief sketch of the available research (also strikingly labelled as a "patchwork" of research, Brookhart, 2004). We start by describing research into the *assessment skills* of teachers (also called assessment literacy, e.g., Stiggins, 1995), then we focus on inventories of teachers' *assessment activities*, and finally we set out some *conceptual models of assessment* that outline relations between concepts, skills, and activities.

This sketch is structured following the recent change in focus in research and theories about classroom assessment: from descriptions of assessment skills teachers should have to teachers' actual assessment activities. These two aspects of classroom assessment are evidently related, in the sense that the assessment skills a teacher has (or does not have) influence the assessment activities he or she

actually uses in the classroom. Quite logically one could expect that there is a temporal, and maybe even a causal, link between assessment skills and assessment activities: if a teacher is not knowledgeable about assessment, he or she will probably not use assessment in the proper way. Both assessment skills and assessment activities have quite extensively been studied, and are used as a basis for concepts and conceptual models in theory on classroom assessment.

### *1.1.1 Assessment skills of teachers*

In the early 1990s the assessment skills of teachers became the main focus of assessment-related research. Ever since the publication of the first version of the standards for teacher competence in educational assessment of students (American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), National Educational Association (NEA), 1990), assessment skills have regularly been investigated (Mertler, 2003; Mertler & Campbell, 2005; Plake, Impara, & Fager, 1993; Popham, 2009; Stiggins, 1995). These standards were developed by an expert group based on a review of research literature focused on improving and defining the assessment skills teachers should have. The particular skills teachers were supposed to have according to these standards were (i) choosing and developing assessment methods, (ii) using assessment results for decision making and grading, (iii) communicating assessment results, and (iv) recognizing unethical assessment practices. These standards were clearly centered on teachers' assessment competence, i.e. assessment skills, but made no mention at all of their actual assessment activities.

Brookhart (2011) recently updated the standards for assessment, taking into account the recent surge the use of formative assessment has taken, especially after the influential work of the Assessment Reform Group (1999) and the famous review study by Black and Wiliam (1998a, 1998b). In the updated standards some assessment skills are still mentioned but the assessment activities of teachers such as setting goals, communicating learning intentions, and interpreting assessment results are given much more importance (Brookhart, 2011). The same trend can be observed in the writings of the American National Board for Professional Teaching Standards (NBPTS, 2010), where assessment practice is one of the certification standards:

Accomplished mathematics teachers *integrate a range of assessment methods* into their instruction to promote the learning of all students by designing, selecting, and ethically *employing assessments* that align with educational goals.

They provide opportunities for students to reflect on their strengths and weaknesses in order to revise, support, and extend their individual performance. (p. 61) [*emphasis added*]

A combination of assessment skills and assessment activities is clearly advocated in the recent standards of both Brookhart (2011) and the NBPTS (2010). The focus in the original version of their standards from over 20 years ago was exclusively on the *assessment skills* teachers should have, whereas in their more recent standards the *assessment activities* of teachers have become the focal point. This transfer can be seen as a parallel to the move from teacher-centered to student-centered education, in the sense that assessment skills only address the teacher, while assessment activities immediately imply that students are involved, in the sense of an interaction between teacher and students.

### *1.1.2 Assessment activities of teachers*

Descriptions of teachers' assessment activities come in different forms and with manifold foci. Here we will outline some examples from research to illustrate the recurring types of assessment activities teachers are using. Most research on assessment activities has been done through a combination of surveys and classroom observations. For instance, McMillan and colleagues (McMillan, 2001; McMillan, Myran, & Workman, 2002) inventoried the assessment activities of primary and secondary education teachers in the U.S., focusing on the information they used to grade their students' performance. Here the assessment activity can be identified as collecting information and providing feedback through grading. Mavrommatis (1997) used a framework based on interviews and observations to describe mathematics teachers' assessment process, taking place in four phases, including evidence collection, evidence interpretation, teachers' responses, and students' reactions. For every phase the actions are described that teachers can undertake, for instance the type of questions they can use to elicit evidence of learning. Here the activities of assessment are observation and questioning to gather 'evidence' or information, and providing feedback to the students. A further example is the study by Wiliam, Lee, Harrison, and Black (2004) on the effects of a professional development track in assessment for learning, where teachers had to use among others questioning and providing feedback. From the foregoing examples of research on teachers' assessment activities (see also, Ginsburg, 2009; Torrance & Pryor, 2001) the following core activities of teachers' classroom assessment practice, emanate: questioning, observation, and providing feedback.

In addition to capturing assessment activities, research has also portrayed the beliefs teachers have about assessment. These beliefs of teachers are chiefly related to the practical (activities) side of assessment. For example, teachers can conceptualize assessment as consisting of rich questioning, and providing feedback to move learning forward (James & Pedder, 2006). Furthermore, another way researchers have looked into the matter of assessment is investigating the relation between the core assessment activities, teachers' assessment skills, and theories of learning and motivation. Then, we come close to what can be considered conceptual models of assessment.

### *1.1.3 Conceptual models of assessment*

As Brookhart (2004) described in a review of research literature on classroom assessment, there are different approaches to study this topic:

Theory relevant to studying classroom assessment comes from several different areas: the study of individual differences (e.g. educational psychology, theories of learning and motivation), the study of groups (e.g. social learning theory, sociology) and the study of measurement (e.g. validity and reliability theory, formative and summative assessment). (p. 429)

This rich variety of perspectives from which assessment can be approached results in conceptual models about classroom assessment showing many different emphases (see Brookhart, 2004). Some authors mainly focus on feedback (Sadler, 1989) or motivation through self-regulation (Clark, 2012), while others concentrate on scaffolding (Van de Pol, Volman, & Beishuizen, 2010), for example.

In addition some broader models have been described that include several factors determining classroom assessment. For example, McMillan (2003) presented a model including teacher knowledge, external factors, and the realities teachers encounter in classroom as the most important influences on the instructional decision-making rationale, which in turn determine the classroom assessment practice. The classroom assessment practice ranged from quizzes and tests, to informal observation, which we can again identify as several of the assessment activities. Another broad vision of classroom assessment was provided by Watson (2000) who listed concepts ranging from theoretical, such as psychological, cognitive, and social factors, via views of mathematics, interpersonal relations, attitudes, feedback or motivation, to classroom practice such as exercises, use of

specific tasks for assessment, and homework. Similar to McMillan's (2003) model, again core assessment activities, assessment skills, and relations between them can be identified. In both models there is a whole system that exists around an individual learner and the assessor, which can be considered of great importance for assessment.

Yet, further models have been proposed as well. For example, Schneider and Gowan (2013) suggested a 'theory of action' of formative classroom assessment. Four assumptions formed the basis of this model and in these assumptions we can once more identify assessment skills as well as activities. The first assumption in this model is the gathering of accurate information about student learning, the second is the analysis of the responses and inferences about learning, the third is providing feedback or adapting instruction, and the fourth is that the student uses this feedback to move forward. Black and Wiliam (2009) proposed a framework for what they called the theory of formative assessment. This framework consisted of a description of practice for the teacher, learners, and peers during (formative) assessment. As a background for this framework they sketched relations between formative assessment and instruction-related issues such as cognitive acceleration, dynamic assessment, and models of self-regulated learning and classroom discourse. Finally, Pellegrino, Chudowsky, and Glaser (2001) have also proposed a model of assessment that can be used to make the relations between different concepts more insightful. They used a triangle with on one end, the assessment activity of observation, the way to elicit evidence of students' competences, and on the two other ends the assessment skills of interpretation, which refer to the process of making sense of the evidence, and the teacher's model of students' cognition or learning in the assessed domain.

A common denominator in all the foregoing models, frameworks, or attempts at theory building, is that they consider assessment to be an interactive process between students and teacher, where the teacher actively searches for information about students' abilities and understanding (assessment activities), and communicates this with the students, as such giving them cognitive and motivational support (assessment skills) to offer learning opportunities. In the end, most studies focus on the purpose of assessment being the improvement of learning (Torrance, 2012; Wiliam, 2011). Some researchers (Sadler, 1989) have called the identification of the gap between the actual current level of performance and the aimed-for level the main goal of assessment. Furthermore,

what can be concluded from these theoretical considerations on classroom assessment is that most are made up of a flat description of the relations between the core assessment activities and theoretical factors influencing assessment. Core assessment activities of questioning, observation, and feedback that could be considered as part of contingent teaching (Van de Pol, Volman, & Beishuizen, 2011) and links to psychological theories on motivation through feedback or self-regulated learning, recur in these considerations. Questioning is considered a mix of questions aimed at revealing what a student knows and questions that help a student to learn (Torrance & Pryor, 2001). Similarly, the feedback teachers provide is generally formatively used and aimed at helping students acquire more knowledge, confidence, and understanding (Hattie & Timperley, 2007). Although the aforementioned lists of assessment skills, activities, and conceptual models cannot be considered a fully-fledged, crystallized theory about assessment, they clearly illustrate that classroom assessment is a complex, all-encompassing process that fulfills a central role in instruction.

## **2. Present study**

In our current investigation we followed the described trend from *assessment skills* teachers have, to *assessment activities*, focusing on what teachers report doing in their classrooms. The goal of the present study was the identification of teachers' assessment profiles on the basis of questionnaire data on teachers' reported assessment practice. Via these profiles we intended to characterize individual teachers' assessment practice. Moreover we strived for a contribution to a better theoretical understanding of the assessment by teachers through the detection of relevant concepts in classroom assessment in mathematics education. We did not have the pretention to propose a new theory or model of assessment, but merely tried to identify clusters of factors in classroom assessment that are important for determining teachers' assessment practice. The idea in this study was to go one small step further than just list concepts and their interrelations, and describe the factors that lie in between. The aim of the study was offering teachers and researchers of assessment in mathematics education a characterization of assessment practice through the determination of teacher profiles. The research question that guided this endeavor was:

Can teachers' current practice of assessment in primary school mathematics education be described by means of assessment profiles?

### **3. Method**

#### **3.1 Ethics statement**

Before starting to fill in the questionnaire teachers were provided with information on the researchers, on the purpose, and on the content of the research. Teachers were also given the choice to participate by agreeing to this information, or to not participate, and could quit the questionnaire at any moment. As all participants voluntarily subscribed to the study and data were analyzed anonymously, we did not formally ask teachers for written consent. Our research was on normal educational practice and we did not consult with an institutional review board (our institute which only focuses on educational research does not have such a board). All this is in line with section 3.4.1 of the VSNU (Dutch Association of Universities) regulations on the use of personal information in scientific research in the Netherlands, the Federal Policy for the Protection of Human Subjects of the National Science Foundation in the USA and section 8.05 “Dispensing with Informed Consent for Research” of the APA ethical standards.

#### **3.2 Online teacher questionnaire**

An expert group consisting of researchers, test developers, education developers, measurement specialists, and didactical experts developed an online questionnaire to collect information on primary school teachers’ assessment practices and beliefs about assessment in mathematics (Veldhuis et al., 2013). This questionnaire contained 40 items (see Table 1 to 5), pertaining to the teachers’ (i) background characteristics, (ii) mathematics teaching practice, (iii) assessment practice, and (iv) perceived usefulness of assessment. Questions with different formats were included: fixed-response and items with a rating scale, but also some open-ended items. Lists of possible assessment methods, and purposes of assessment, were deduced from literature on classroom assessment (Black & Wiliam, 1998a; 1998b; Mavrommatis, 1997; Suurtamm, Koch, & Arden, 2010).

#### **3.3 Procedure of data collection**

The sample of participating teachers was obtained through an open invitation by e-mail, which was sent successfully to 5094 primary schools for regular education in the Netherlands. Teachers who were willing to respond to the online questionnaire were promised a set of digital mathematical exercise

material as a reward. In February 2012, we sent a renewed request to all teachers that did not fill in the questionnaire after the first request. The final sample included 960 teachers from 557 different schools, who filled in at least one question about their assessment practice. Of the sample of teachers 83.7% were female, and the mean age was 41.4 years ( $SD = 11.6$ ).

To investigate the representativeness of the sample we compared background characteristics with available national statistics (Statistieken ArbeidsMarkt Onderwijs Sectoren, 2010). Almost all variables, including age, gender, geographical location of the school, urbanization level of the school, textbook use, education, religious denomination of the school, and the size of the appointment of the teacher followed approximately the same distribution as the national statistics. See Appendix for more details.

### **3.4 Data analyses**

We analyzed the data in two steps. First, we looked into the factorial structure of the questionnaire and the underlying classes of teachers. Then, we investigated the differences between different classes of teachers on the factors of the assessment questionnaire. To identify the latent structure of what was measured by the questionnaire and be able to construct assessment profiles of teachers we used a combination of latent variable modeling techniques. In this approach it is important to be knowledgeable of the fact that every model is an oversimplification of reality, and can thus never be a perfect fit to the data. Additionally, no golden rules for deciding upon the fit of the model to the data exist; therefore we have to investigate the relative fit of the model in comparison to other, comparable, models. Then, to decide which model is most appropriate in describing the data it is advised to use substantive as well as statistical model fit checking (Muthén, 2003). Substantive model checking concerns checking whether the model's predictions and constituents are in line with theoretical and practical expectations. Statistical model fit checking can be done in a variety of ways. There exists a multitude of statistical methods to compare the statistical merit of different models that can generally be divided in two categories. One is a statistical test of model fit, where the model of interest is compared via a likelihood ratio test or a  $\chi^2$ -test to neighboring models. The other is to compare statistical indicators such as information criteria or entropy between different nested models (Clark et al., 2009; Nylund, Asparouhov, & Muthén, 2007).

In evaluating the different latent variable models in this study both the aforementioned statistical and the substantive model fit checking methods have been used. To explore the underlying structure of the items that measure teachers' mathematics assessment practice, we performed several exploratory factor analyses, which was deemed most appropriate (Fabrigar, Wegener, MacCallum, & Strahan, 1999), because the questionnaire was constructed to measure assessment practice in mathematics education in a rather open way and no specific theoretical ideas about the factorial structure were proposed in advance. The technique of exploratory factor analysis was used to understand the structure of variation on measured variables by estimating the correlations between latent factors and these measured variables. Experts in factor analytical research have different opinions about which statistics to include to evaluate statistical model-data fit, but they generally agree that at least a  $\chi^2$ -statistic, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI) should be reported (Barrett, 2007; Bentler, 2007). To indicate acceptable to good model fit, the conventions are that the RMSEA should be around 0.06 (Hu & Bentler, 1999) and the CFI more than 0.96 (Yu, 2002). Using Mplus 5.21 (Muthén & Muthén, 2007) we performed exploratory factor analyses with weighted least squares method (WLSM) estimation and geomin oblique rotation. Finally, we took into consideration whether the items making up the factors had sufficiently in common and whether the factors theoretically made sense, which provided us with substantive reasons to decide upon fit and allowing us to name the factors accordingly.

Furthermore, to investigate whether these latent factors could also be used to interpret classes of teachers, we performed a latent class analysis. This is a statistical technique permitting the identification of underlying classes of individuals based on differences in their responses on items in a questionnaire or test. The underlying classes are identified on a discrete latent variable and permit the division of the sample in qualitatively differing subgroups (Magidson & Vermunt, 2002). As input for this analysis the item scores on the part of the questionnaire related to teachers' assessment practice were used. The teachers in our sample were assigned to the different latent classes – that we will call assessment profiles – through modal assignment, i.e. they were assigned to the latent class to which they had the highest probability of belonging.

The differences between teachers with different assessment profiles on several

background variables were investigated with analyses of variance, Kruskal Wallis and  $\chi^2$ -difference tests. Through these analyses we could determine the characterizing elements for every profile. All inferential analyses were performed in SPSS 20 (IBM Corp, 2011) and the latent variable modeling was done in Mplus 5.21 (Muthén & Muthén, 2007).

## 4. Results

### 4.1 Teachers' assessment practice

The earlier study in which we carried out a descriptive analysis of the questionnaire data [3] revealed that the Dutch primary school teachers involved in the survey used a mix of observation- and instrument-based methods in mathematics education. The most used observation-based methods were questioning, observing, and correcting written work (> 77% weekly). The main instrument-based methods were textbook and pupil monitoring system tests (> 85% several times a year). Teachers also used these methods for a mix of summative, formative, and diagnostic purposes. The most used purposes were: of the summative type, selecting what mathematics subjects should be taught (42% weekly); of the formative type, providing feedback, determining the speed of teaching, and adapting instruction (> 62% weekly); of the diagnostic type, investigating reasons for errors (60% weekly).

### 4.2 Teachers' assessment profiles

#### 4.2.1 Factor analyses

After comparing one- to seven-factor solutions and eliminating items with cross loadings over |0.4|, an exploratory factor analysis delivered a five-factor solution that had a good enough fit ( $\chi^2(1076, N = 960) = 5494.1, p < .001, RMSEA = .064, CFI = .961$ ). Also, these five factors all had eigenvalues over 2.5 and the scree plot showed a clear "elbow" after the fifth factor. The  $\chi^2$ -statistic of the overall model fit was significant, indicating a less than optimal fitting model. Nevertheless, this nested five-factor solution fitted significantly better than the four-factor solution, as illustrated by the Satorra-Bentler scaled  $\chi^2$ -test, which is unaffected by non-normality ( $TRd(df = 48) = 952.68, p < .0001$ ). The different subscales used in the questionnaire loaded on these latent factors (see Tables 1-5 for the items constituting the factors and the corresponding scale's Cronbach's alpha), providing substantive evidence for this five-factor solution.

Table 1

*Factor Loadings of the Items on Goal Centeredness of Assessment ( $\alpha = .804$ )*

Questionnaire item	Factor loading
Assessment purpose: Determine mastery	.793
Assessment purpose: Adapt instruction	.778
Assessment purpose: Determine progress	.734
Assessment purpose: Tune the speed of instruction	.728
Assessment purpose: Select mathematics subjects	.636
Assessment purpose: Investigate reasons for errors	.592
Assessment purpose: Formulate learning goals	.520
Assessment purpose: Provide feedback	.512
Assessment purpose: Establish level groups	.489
Assessment purpose: Stimulate thinking	.487
Assessment method: Textbook tests	.401
Assessment purpose: Stimulate use of scrap paper	.381
Frequency of need for assessment information	.374
Setting of clear goals for students	.363
Assessment method: Correct written work	.339
Assessment method: Questioning	.328
Assessment method: Observation	.301

Table 2

*Factor Loadings of the Items on Diversity of Assessment Problem Format ( $\alpha = .770$ )*

Questionnaire item	Factor loading
Mathematical problems in context	.930
Bare mathematical problems	.887
Mathematical problems, students explain their calculations	.875
Mathematical problems with more than one correct answer	.699

Regarding the items that constitute these factors we decided on the following names: (1) *Goal centeredness of assessment*, (2) *Authentic nature of assessment*, (3) *Perceived usefulness of assessment*, (4) *Diversity of assessment problem format*, and (5) *Allocated importance of assessing skills and knowledge*. Among the items in the factor *Goal centeredness of assessment*

were whether teachers set goals for students and in particular the types of purposes their assessments served. The items relating to the type of exercises teachers included in mathematics tests made up the *Diversity of assessment problem format* factor. The *Allocated importance of assessing skills and knowledge* factor constituted of items measuring the importance of assessing different skills and types of knowledge. The *Perceived usefulness of assessment* factor comprised the items with statements about assessment such as: assessment helps to improve my teaching. The *Authentic nature of assessment methods* factor consisted of items measuring the frequency of the use of authentic assessment methods, such as practical assignments, student- or teacher-developed tests, and items loading negatively on this factor, such as the use of student monitoring system tests or textbook tests, that are the opposite of authentic assessment methods.

Table 3

*Factor Loadings of the Items on Authentic Nature of Assessment ( $\alpha = .456$ )*

Questionnaire item	Factor loading
Assessment method: Practical assignments	.706
Assessment method: Teacher-developed tests	.643
Assessment method: Student-developed tests	.382
Importance of assessing: Students' design skills	.322
Importance of assessing: Students' memory skills	-.246
Assessment purpose: Assessing use of scrap paper	-.334
Importance of assessing: Students' factual knowledge	-.353
Assessment method: Student monitoring tests	-.361
Assessment method: Correcting written work	-.378
Assessment method: Textbook tests	-.483

Table 4

*Factor Loadings of the Items on Allocated Importance of Assessing Types of Skills and Knowledge ( $\alpha = .823$ )*

Questionnaire item	Factor loading
Importance of assessing procedural knowledge	.709
Importance of assessing factual knowledge	.707
Importance of assessing conceptual knowledge	.701
Importance of assessing memory skills	.684
Importance of assessing understanding skills	.675
Importance of assessing applying skills	.640
Importance of assessing analyzing skills	.631
Importance of assessing evaluation skills	.520
Importance of assessing self-knowledge	.473
Importance of assessing design skills	.425

Table 5

*Factor Loadings of the Items on the Perceived Usefulness of Assessment ( $\alpha = .803$ )*

Questionnaire item	Factor loading
Assessment can determine what students have learned	.880
Assessment results predict students' performances	.851
Assessment helps to improve my teaching	.838
Assessment helps students to learn	.837
Assessment provides information about learning needs	.833
Assessment can be used to map strong/weak sides	.817
Assessment has much influence on my teaching <sup>a</sup>	.816
Assessment creates a better learning climate	.813
Assessment is an interruption of my teaching <sup>a</sup>	.800
Assessment informs what students can <sup>a</sup>	.760

<sup>a</sup>These statements were originally phrased negatively in the questionnaire, e.g. "Assessment has little influence on my teaching", and have been recoded

#### 4.2.2 Correlations

Correlations between the five factors are displayed in Table 6. Inspecting these correlations shows that *Authentic nature of assessment* was moderately negatively correlated with all factors ( $-.301 < r < -.127$ , all  $ps < .01$ ) except for

*Perceived usefulness of assessment* with which it is uncorrelated. This indicates that the *Authentic nature of assessment* factor is quite different from the other factors, which stands to reason if one inspects the items belonging to this factor and its reliability: The items in this factor are very diverse (cf. Table 4) and the reliability was low of  $\alpha = .456$ ; whereas the items in the other factors were much more uniform with high internal reliabilities of  $\alpha > .77$ . All other factors were weakly to moderately positively correlated with each other ( $.069 < r < .425$ , all  $ps < .05$ ).

Table 6

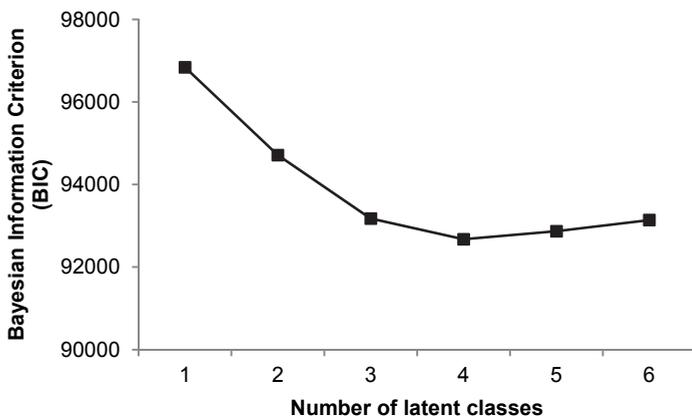
*Correlations Among the Factors from the Exploratory Factor Analysis (Ns > 857)*

Factors	GC	DAF	AA	IASK	PUA
GC	-				
DAF	.154**	-			
AA	-.301**	-.127**	-		
IASK	.346**	.102**	-.148**	-	
PUA	.262**	.069*	.025	.425**	-

*Note.* These are Pearson's  $r$  coefficients. GC = Goal centeredness of assessment, DAF = Diversity of assessment problem format, AA = Authentic nature of assessment, IASK = Allocated importance of assessing, PUA = Perceived usefulness of assessment  
 \*\* $p < .01$ . \* $p < .05$ .

#### 4.2.3 Latent class analysis

The analysis carried out thus far gave us an approximation of the underlying structure of the questionnaire, but not yet information on teachers that could be used in practice. To be able to characterize teachers' assessment practice and assign them to different assessment profiles we performed a latent class analysis using all variable scores as input. As such we were able to check whether we would be able to show differences between the latent classes on the factors we found separately. We used the Bayesian Information Criterion (BIC) and entropy to select the number of latent classes that best summarizes the variation data. As shown in Figure 1 the value of the BIC decreased until four latent classes and increased subsequently.



*Figure 1.* The value of the Bayesian Information Criterion for one to six latent classes.

This indicates that four latent classes provided the best fitting solution, as a lower value of the BIC indicates a better fit. The relative entropy of .93 (measuring the uncertainty of the classification, from 0 = high uncertainty to 1 = low uncertainty, Dias & Vermunt, 2006) of the latent class model was high; indicating that the four classes were clearly separated (Nylund, Asparouhov, & Muthèn, 2007). Including age, gender, grade, or textbook use as covariates did not improve the fit of the model. Having four latent classes provided the best solution.

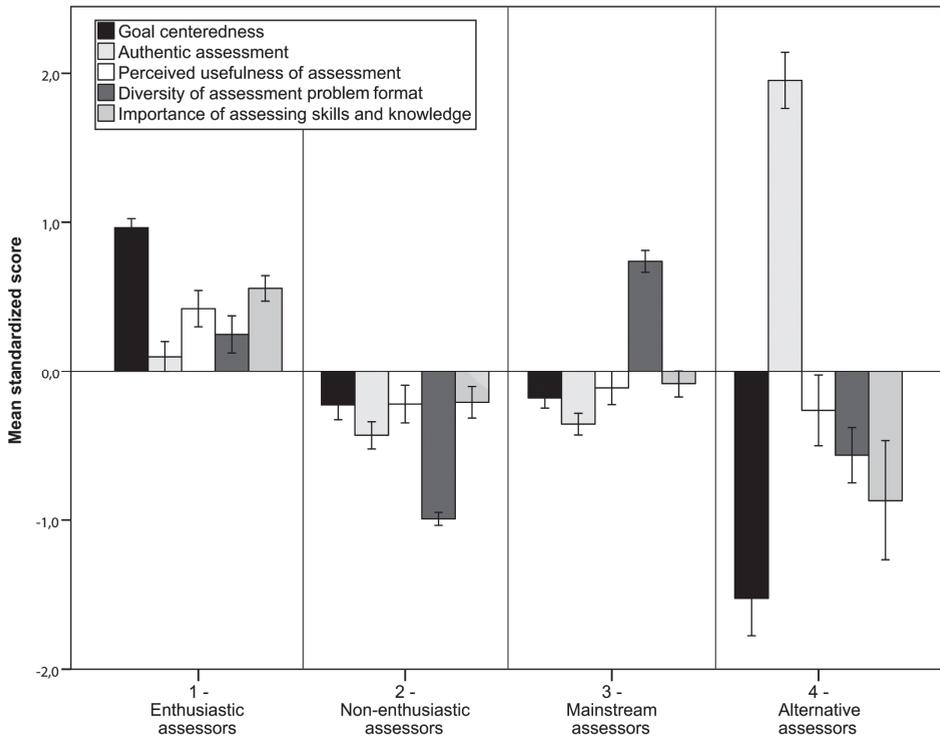


Figure 2. Mean standardized scores on factors for teachers in the four latent classes. Whiskers indicate 95% confidence interval.

Figure 2 shows the profiles of teachers from the four different classes in relation to the five standardized measures of teachers' mathematics assessment practice. To find out whether teachers thus assigned to the four latent classes differed on the five factors of assessment identified before, we performed several analyses of variance. The results showed that teachers from one latent class to another differed significantly from each other. We found large effects for *Goal centeredness of assessment* ( $F(3, 852) = 324.2, p < .001, \eta_p^2 = 0.533$ ) and *Diversity of assessment problem format* ( $F(3, 852) = 275.2, p < .0001, \eta_p^2 = 0.492$ ), and medium to small effects for *Authentic nature of assessment* ( $F(3, 852) = 258.0, p < .001, \eta_p^2 = 0.476$ ), *Allocated importance of assessing skills and knowledge* ( $F(3, 852) = 60.3, p < .001, \eta_p^2 = 0.175$ ), and *Perceived usefulness of assessment* ( $F(3, 852) = 22.8, p < .001, \eta_p^2 = 0.074$ ).

Post-hoc tests using Bonferroni correction showed that the differences between all four latent classes were significant for *Diversity of assessment problem format* (all  $ps < .0001$ ; see Figure 2 for the directions of these differences). Concerning the scores on *Goal centeredness of assessment* ( $p = 1.00$ ), *Authentic nature of assessment* ( $p = 1.00$ ), and *Allocated importance of assessing skills and knowledge* ( $p = .724$ ), teachers in the second and third latent classes did not differ significantly from each other; differences between teachers in the first and fourth latent classes, however, were significant (all  $ps < .001$ ). On *Perceived usefulness of assessment* teachers in the first latent class scored significantly higher than teachers from the other three classes ( $p < .001$ ).

Based on the results of these profile analyses we interpreted the different profiles as follows. In the first class, the teachers (28.5%) had above average scores on all assessment practice measures, with particularly high scores on *Goal centeredness of assessment*, *Perceived usefulness of assessment*, and *Allocated importance of assessing skills and knowledge*: they were aware of the different possibilities assessment offers them, reported using them likewise, and did this for a variety of goals. As such we considered these teachers to be *enthusiastic assessors*. Teachers in the second latent class were labelled as *non-enthusiastic assessors*. These teachers (25.8%) had scores below average on all measures, particularly on *Diversity of assessment problem format*. They viewed assessment more often in a negative way and used it accordingly less and in a less diverse way. Teachers in the third latent class were considered *mainstream assessors*. On four measures these teachers scored slightly below average, with the exception of the high score for the *Diversity of assessment problem format*. We called them *mainstream assessors*, because they scored generally close to average and most teachers belonged to this profile: 35.3% of our sample. Finally, the teachers from the fourth latent class (10.3%) were named *alternative assessors*. Teachers in this profile had an ambiguous view of assessment. On the one hand they reported a lot of *Authentic nature of assessment* use; for example, they devised their own tasks and tests. On the other hand they had scores below average on the remaining measures, with particularly low scores on *Goal centeredness* and *Allocated importance of assessing skills and knowledge*, clearly reflecting that they do not find assessment important, necessary, or helping them to reach certain goals.

### 4.3 Teacher characteristics and assessment profiles

To investigate which background characteristics are related to teachers' attribution to one of the latent classes, we compared the scores for teachers with different profiles. In Table 7 the standardized means per profile for the five factors of the questionnaire, as well as the means on background variables, are displayed. Using an analysis of variance we found that *non-enthusiastic assessors* ( $M = 44.3$ ,  $SD = 11.5$ ;  $F(3, 952) = 8.176$ ,  $p < .001$ ) were significantly older than *enthusiastic assessors* ( $M = 40.8$ ,  $SD = 12.0$ ;  $p = .003$ ,  $d = 0.30$  (95% CI: [-0.71, 1.31])) and *mainstream assessors* ( $M = 39.7$ ,  $SD = 11.1$ ;  $p < .001$ ,  $d = 0.41$  (95% CI: [-0.50, 1.32])). The number of years of teaching experience showed the same pattern ( $F(3, 952) = 6.705$ ,  $p < .001$ ); which seems logical, as age and teaching experience correlate highly  $r = .830$ . *Enthusiastic assessors* ( $M = 3.8$ ,  $SD = 1.2$ ) worked significantly more days than *non-enthusiastic assessors* ( $M = 3.5$ ,  $SD = 1.2$ ;  $F(3, 949) = 2.873$ ,  $p = .035$ ,  $d = 0.25$  (95% CI: [0.15, 0.35])). Belonging to an assessment profile was significantly related to whether teachers obtained their professional qualification from a teacher education college for primary school teachers ( $\chi^2(3, N = 960) = 18.97$ ,  $p < .001$ ); proportionally few *alternative assessors* attended such a college (only 69% against 77-87% for the other profiles). The assessment profile was not significantly related ( $\chi^2(6, N = 960) = 10.82$ ,  $p = .094$ ) to the type of pedagogical-didactical approach of the primary schools where the teachers were working –including regular schools and schools with a specific organization or teaching method such as Montessori and Dalton schools.

Table 7  
*Mean Values of Factors Constituting the Profiles (above Dotted Line) and of Related Variables, and the Significant Profile Differences*

Variables	Total	Assessment profiles				Sig. diff.
		1. Enthu- siastic	2. Non- enthu- siastic	3. Main- stream	4. Alter- native	
Goal centeredness of assessment <sup>a</sup>	-	<b>0.96</b>	-0.23	-0.18	0.01	1>4>2,3
Diversity of problem format <sup>a</sup>	-	0.25	-0.99	<b>0.74</b>	-0.57	3>1>4>2
Authentic nature of assessment <sup>a</sup>	-	0.10	-0.43	-0.36	<b>1.95</b>	4>1>2,3
Allocated importance <sup>a</sup>	-	<b>0.56</b>	-0.21	-0.08	-0.87	1>2,3>4
Perceived usefulness <sup>a</sup>	-	<b>0.42</b>	-0.22	-0.11	-0.26	1>2,3,4
Age (in years)	41.4	40.8	<b>44.4</b>	39.7	41.3	2>1,3
Gender (% male)	16	13	23	18	<b>2</b>	4<1,2,3
Teaching exp. (in years)	16.2	15.3	<b>18.8</b>	15	15.8	2>1,3
Teacher trainer college (%)	80	79	76	<b>86</b>	69	3>2,4
Didactical approach (% regular)	84	81	88	84	77	n.s.
Students in class (mean number)	23.8	23.8	23.4	24	23.5	n.s.
Prof. dev. sessions attended (mean number)	10.2	11.7	9.9	8.8	11.6	n.s.

*Table continued on next page*

Variables	Total	Assessment profiles				Sig. diff.
		1. Enthu- siastic	2. Non- enthu- siastic	3. Main- stream	4. Alter- native	
Size of position (days/week)	3.7	3.8	3.5	3.7	3.6	1>2
Time for assessment (min/week)	72.2	<b>85.6</b>	67.2	68.4	59.9	1>2,3,4
Frequency revise level groups <sup>b</sup>	3.2	<b>3.5</b>	3.2	3.1	3	1>2,3,4 3>4
Frequency discuss goals <sup>b</sup>	4.3	<b>5.3</b>	3.9	4.1	3.2	1>2,3>4
Frequency need information <sup>b</sup>	4.5	<b>5.2</b>	4.3	4.2	4	1>2,3,4 2>4

Note. The significantly highest value per row is printed in **bold**.

<sup>a</sup> Mean standardized scores <sup>b</sup> Mean scores of response options: 1 = *Rarely to never*, 2 = *Yearly*, 3 = *A few times a year*, 4 = *Monthly*, 5 = *Weekly*, 6 = *A few times a week*.

Grade level and profile membership were significantly related ( $\chi^2(12, N = 941) = 576.94, p < .001$ ). *Alternative assessors* were mostly kindergarten teachers (80%), whereas there were very few (5%) in the other profiles. Proportionally, more *mainstream assessors* (53%) taught Grade 4 to Grade 6, than *enthusiastic* (45%) and *non-enthusiastic assessors* (50%). There was also a significant relation between gender and assessment profile ( $\chi^2(3, N = 956) = 28.09, p < .001$ ). Very few male teachers were *alternative assessors*; just 2%, whereas in the other profiles at least 13% of the teachers were male. The time teachers reported using to assess mathematics every week showed a pattern that reinforced the interpretation of the profiles. *Enthusiastic assessors* dedicated more time to the assessment of their students ( $M = 85.61, SD = 70.0$ ) than in all three other profiles ( $F(3, 863) = 6.378, p < .001$ ; post hoc Tukey all  $ps = .003$ ).

Analysis with a Kruskal Wallis test, followed by a post-hoc Mann-Whitney test, showed that *enthusiastic assessors* revised the level groups for their students with a higher frequency than teachers from the other profiles

( $\chi^2(3, N = 955) = 57.98, p < .001$ ), and *mainstream assessors* more often than *alternative assessors* ( $p = .03$ ). Additionally, the frequency with which they discussed goals with students was higher for *enthusiastic assessors* than for *non-enthusiastic* and *mainstream assessors*, and all these frequencies were higher than for *alternative assessors* ( $\chi^2(3, N = 951) = 104.91, p < .001$ ). The need for assessment information was higher for *enthusiastic assessors*; they needed this more often than teachers from the other profiles ( $\chi^2(3, N = 862) = 117.95, p < .001$ ). *Alternative assessors* were different concerning the assessment methods they considered to be most relevant. They found practical assignments ( $\chi^2(3, N = 883) = 170.74, p < .0001$ ) and teacher-developed exercises ( $\chi^2(3, N = 883) = 95.44, p < .001$ ) considerably more relevant than teachers from the other profiles, and textbook tests ( $\chi^2(3, N = 883) = 234.12, p < .001$ ) and student monitoring system tests ( $\chi^2(3, N = 883) = 32.47, p < .0001$ ) less relevant.

## 5. Discussion

### 5.1 Conclusions

In this study we have identified four distinct teacher profiles with clearly different scores on the five underlying factors from a mathematics assessment questionnaire. Exploratory factor analyses permitted to decide on the number and content of the underlying factors of the questionnaire, followed by a latent class analysis that determined the number of distinct latent classes to which individual teachers belonged. The assessment profile to which most teachers in our sample belonged was the *mainstream assessors* profile (see Table 8). In this profile most teachers regularly used different types of assessment, test-based and observation-based, for both summative and formative purposes. On all factors, i.e., *Goal centeredness of assessment*, *Diversity of assessment problem format*, *Summative assessment methods*, *Allocated importance of assessing skills and knowledge*, *Perceived usefulness of assessment*, and *Authentic nature of assessment methods*, teachers with this profile scored around the mean. The next biggest group was the *enthusiastic assessors*. Teachers with this profile were very aware of the different possibilities assessment offers them, and used them likewise. On all components these teachers scored above the mean, with a peak on *Goal centeredness of assessment*. An almost equally large group of teachers were the *non-enthusiastic assessors*. These teachers viewed assessment

more often in a negative way and used it accordingly less. On all factors, teachers with this profile scored below average. Finally, there were the *alternative assessors*. Teachers with this profile had an ambiguous view of assessment. Although they reported a lot of own input in assessment and devised their own tasks and tests, they did not find assessment important or necessary. We found that most teachers with this profile were actually kindergarten teachers, which might explain their divergent profile: in kindergarten standardized assessment is almost absent from the classroom and as such seen as unnecessary.

In sum, we can say that our main aim of identifying meaningful assessment profiles has been achieved, but the question that remains is: How can this characterization contribute to the existing plethora of conceptualizations and lists of assessment activities and skills? Based on our analyses, we can conclude that the factors mainly fall under the headings of assessment activities (*Authentic assessment* and *Diversity of assessment problem format*) and assessment skills (*Goal centeredness of assessment*, *Perceived usefulness of assessment*, and *Allocated importance of assessing skills and knowledge*). The relations we have found between these factors and the characteristics of the teachers have enabled us to determine profiles with clear differences between teachers. These profiles serve a double purpose. First, they permit to typify the assessment teachers perform in their classroom, and as such they can be used to propose tailor-made professional development for teachers with specific profiles. A second purpose is that this profile characterization makes a connection between the assessment activities and assessment skills of teachers, and that this connection could be used in the further development of conceptualizations and eventually a theory of classroom assessment.

Table 8

*Summary and Description of Teachers' Assessment Profiles*

Profiles	Description
Assessment profile 1: <i>Enthusiastic assessors</i> (28.5%)	Enthusiastic assessors had above average scores on all measures in the questionnaire: they were particularly goal-centered in assessment, and perceived it to be useful and important. Teachers with this profile dedicated more time to assessment than teachers with the other profiles.
Assessment profile 2: <i>Non-enthusiastic assessors</i> (25.8%)	Non-enthusiastic assessors had below average scores on all measures in the questionnaire: they did not think assessment to be important or useful, and particularly did not use a variety of problem formats to assess mathematics. Teachers with this profile were generally older than teachers with the other profiles.
Assessment profile 3: <i>Mainstream assessors</i> (35.3%)	Mainstream assessors scored slightly below average on most measures in the questionnaire: they were less goal centered, used less often authentic assessment, perceived assessment as averagely useful and important, but used more diverse problem formats to assess mathematics. Teachers with this profile were more often educated at a teacher education college for primary school teachers than teachers with the other profiles.
Assessment profile 4: <i>Alternative assessors</i> (10.3%)	Alternative assessors had very low scores on all measures, except on authentic nature of assessment: they were not goal centered, perceived assessment not as useful or important, and did not use a diversity of mathematics problems. Teachers with this profile were mostly kindergarten teachers, less often educated at a teacher education college for primary school teachers, almost exclusively female, and half of them did not use a textbook for mathematics.

**5.2 Limitations**

When using the results of our study it should be taken into account that the study is based on a rather large but local sample; all teachers came from the Netherlands. Moreover, the voluntary participation of the teachers in our study

may have resulted in some bias in the sample. Although we found the teachers in our sample quite representative of the population of primary school teachers in the Netherlands, it is still possible that participating teachers were special in other aspects; they could, for instance, have been positively biased towards assessment in their responses on the questionnaire. The purpose of our survey, however, was rather neutral and by only asking teachers to inform us anonymously about their assessment practice, we think this potential positive bias did not have a detrimental influence on the reliability of teachers' responses. The fact that we used self-report data from teachers as a basis for all analyses in this study could have led to another limitation. In the interpretation of our results it is important not to forget that we evidently cannot be entirely sure from these self-report data that teachers actually do, believe, and think what they report to be doing, believing, and thinking. Nonetheless, teachers had no reason whatsoever to misreport their behavior or opinions, because the questionnaire was anonymous. Yet, to control for this, in further research it would be interesting to compare and combine different sources of data about teachers' practice in mathematics assessment, such as observations, interviews, and student data, and integrate these into the assessment profiles.

### **5.3 Perspectives**

Taking supplementary sources into account and extending this study into classroom assessment in primary school mathematics education to other countries might lead to getting even more robust assessment profiles. Another approach would be to look into applications of the assessment profiles, for instance targeting a specific type of teachers for professional development. It would also be possible to investigate the effects of professional development on the assessment profile of teachers; teachers could move from one profile to another. Furthermore, another approach would be to link teachers through their profiles to levels of student performance; as assessment and instruction are intrinsically linked (Shepard, 2000), different types of assessment would probably be linked to different learning results. In a sense this is in line with results of research on the effects of classroom assessment (ARG, 1999; Black & Wiliam, 1998a; Black & Wiliam, 1998b; Brookhart, 2004; James & Pedder, 2006; Mavrommatis, 1997; Wiliam, Lee, Harrison, & Black, 2004); teachers that assess more and in an effective, often formative, fashion, have been shown to ensure more learning gain in their students. A tentative

## Chapter 5

hypothesis would be to expect this to come from the teachers that are *enthusiastic assessors* for example, given that they assess often and use assessment in various ways (cf. Table 8).

To conclude, through our profile characterization of teachers' assessment practice we were able to select some of the skills and activities from the universe of assessment skills and activities. In this way we have brought some structure to the many possible characterizations of assessment practice and skills that exist. These assessment profiles can contribute to a better theoretical understanding of classroom assessment and can also be useful in a practical manner as a basis for designing professional development and instruments for measuring teachers' assessment practice.

## References

- American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), & National Educational Association (NEA) (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, 9(4), 30-32.
- Assessment Reform Group (1999). *Assessment for learning: beyond the black box*. Cambridge: University of Cambridge School of Education.
- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5), 815-824.
- Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, 42(5), 825-829.
- Black, P., & Wiliam, D. (1998a). Inside the black box. Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.
- Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in education: Principles, Policy & Practice*, 5(1), 7-74.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-31.
- Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, 106(3), 429-458.
- Brookhart, S. M. (2011). Educational assessment knowledge and skills for Teachers. *Educational Measurement: Issues and Practice*, 30(1).
- Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3-17). Abingdon, UK: Routledge.
- Clark, I. (2012). Formative assessment: assessment is for self-regulated learning. *Educational Psychology Review*, 24(2), 205-249.
- Clark, S. L., Muthén, B. O., Kaprio, J., D'Onofrio, B. M., Viken, R., Rose, R. J., et al. (2009). *Models and strategies for factor mixture analysis: Two examples concerning the structure of underlying psychological disorders*. [http://www.statmodel.com/download/FMA%20Paper\\_v142.pdf](http://www.statmodel.com/download/FMA%20Paper_v142.pdf)
- Dias, J. G., & Vermunt, J. K. (2006). Bootstrap methods for measuring classification uncertainty in latent class analysis. In A. Rizzi & M Vichi (Eds.), *Proceedings in computational statistics* (pp. 31-41). Heidelberg: Springer.

- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.
- Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: Children's minds, teachers' minds. *Human Development, 52*(2), 109–128.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.
- Hu, L-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.
- IBM Corp (2011). *IBM SPSS Statistics for Windows, Version 20.0*. Armonk, NY: IBM Corp.
- James, M., & Pedder, D. (2006). Beyond method: assessment and learning practices and values. *Curriculum Journal, 17*(2).
- Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A Comparison with K-means. *Canadian Journal of Marketing Research, 20*, 37-44.
- Mavrommatis, Y. (1997). Understanding assessment in the classroom: Phases of the assessment process - The assessment episode. *Assessment in Education: Principles, Policy & Practice, 4*(3), 381-399.
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.
- McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34-43.
- McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203-213.
- Mertler, C. A. (2003). *Preservice versus inservice teachers' assessment literacy: does classroom experience make a difference?* Paper presented at the Mid-Western Educational Research Association, Columbus, Ohio, October 15-18.
- Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the American Educational Research Association, Montréal, Quebec, April 11-15.

- Muthén, B. O. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods*, 8(3), 369-377.
- Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.
- National Board for Professional Teaching Standards (2010). *Mathematics adolescence and young adulthood standards*. Retrieved from [http://www.nbpts.org/userfiles/file/ea\\_math\\_standards.pdf](http://www.nbpts.org/userfiles/file/ea_math_standards.pdf)
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4), 535-569.
- Pedder, D. (2007). Profiling teachers' professional learning practices and values: differences between and within schools. *The Curriculum, Journal*, 18(3), 231-252.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know. The science and design of educational assessment*. Washington, DC: National Academy Press.
- Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, 12(4), 10-12.
- Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental? *Theory into Practice*, 48(1), 4-11.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144.
- Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analyses. *Journal of Psychoeducational Assessment*, 29(4), 304-321.
- Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education*, 26(3), 191-204.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.
- Statistieken ArbeidsMarkt Onderwijs Sectoren [Statistics of the educational labor market, Netherlands] (2010). Retrieved from <http://www.stamos.nl>
- Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77(3), 238-246

- Suurtamm, C., Koch, M., & Arden, A. (2010). Teachers' assessment practices in mathematics: Classrooms in the context of reform. *Assessment in education: Principles, Policy & Practice*, 17(4), 399-417.
- 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.
- Torrance, H. (2012). Formative assessment at the crossroads: conformance, deformative, and transformative assessment. *Oxford Review of Education*, 38(3), 323-342.
- Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271-296.
- Van de Pol, J., Volman, M., & Beishuizen, J. (2011). Patterns of contingent teaching in teacher-student interaction. *Learning and Instruction*, 21(1), 46-57.
- Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen, J. A., & Eggen, T. H. J. M. (2013). Teachers' use of classroom assessment in primary school mathematics education in the Netherlands. *CADMO*, 21(2), 35-53.
- Watson, A. (2000). Mathematics teachers acting as informal assessors: practices, problems, and recommendations. *Educational Studies in Mathematics*, 41(1), 69-91.
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3-14.
- William, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice*, 11(1), 49-65.
- Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Dissertation, University of California, Los Angeles.

## Appendix: Representativeness

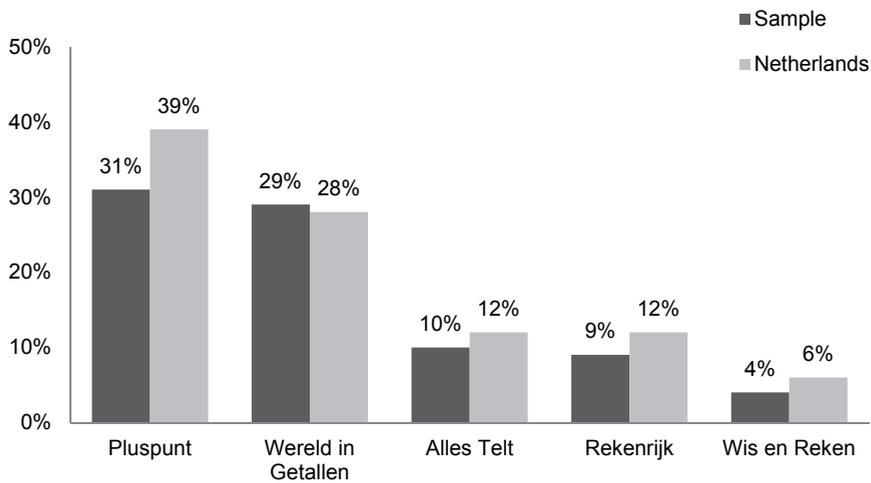
To investigate the representativeness of the sample, as teachers from schools that did respond differ could be different from those that did not; we compared their background characteristics to the available national statistics (CBS, 2012; www.stamos.nl, 2010) of all primary school teachers in the Netherlands (NL). In our sample 83.6% of the teachers were female and 15.6% male (all primary school teachers in the Netherlands in 2010, NL: 85.5% female and 14.5% male) with a mean age of 41.8 years ( $SD = 11.66$ ; NL: 43.4 years). The amount of experience in teaching years of these teachers followed the same bimodal distribution as in the total population of Dutch primary school teachers with a grand mean of 16.5 years. Of the teachers 68% worked part-time (most worked three or four days per week) and 31.9% worked fulltime. Most teachers (78.9%) were groomed at the PABO (teacher training college, see Table A1; in NL: 78.9%).

Table A1

### *Prior Education of Teachers in the Sample*

Prior education		Percent	<i>n</i>
High school	VMBO (Preparatory vocational education)	16.4	202
	HAVO (Higher general secondary education)	37.7	463
	VWO (Preparatory scientific education)	10.4	128
Vocational education	MBO (Intermediate vocational education)	2.1	26
	KLOS (Kindergarten teacher training)	5.6	69
Higher vocational education	PABO (Teacher training)	78.9	969
	HBO (Higher vocational education)	14.3	175
Higher education	HBO masters	4.6	56
	University	4.6	57

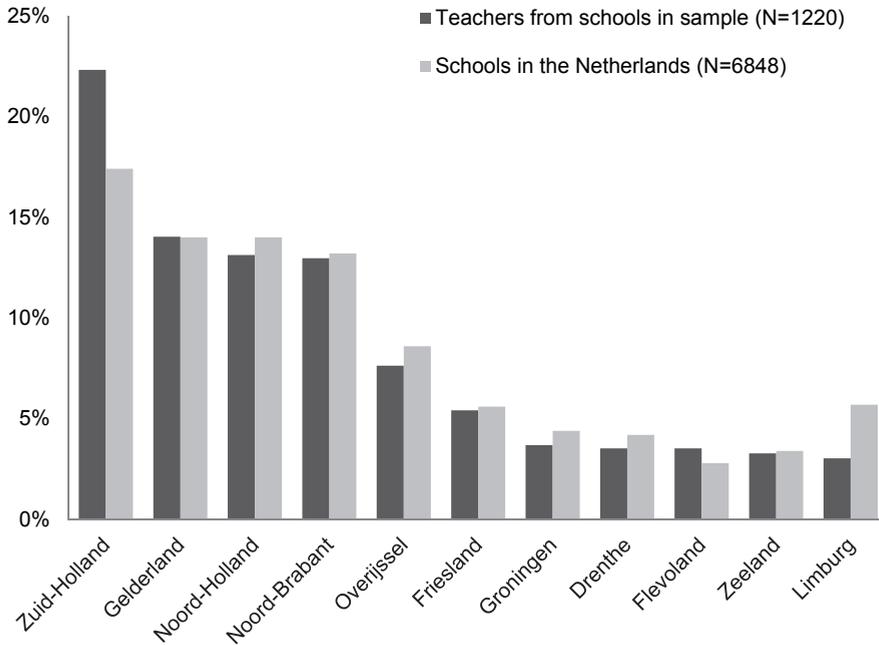
The different textbooks teachers in our sample use are displayed in Figure A1.



*Figure A1.* Textbook use in our sample (dark grey) and in the Netherlands (lighter grey)

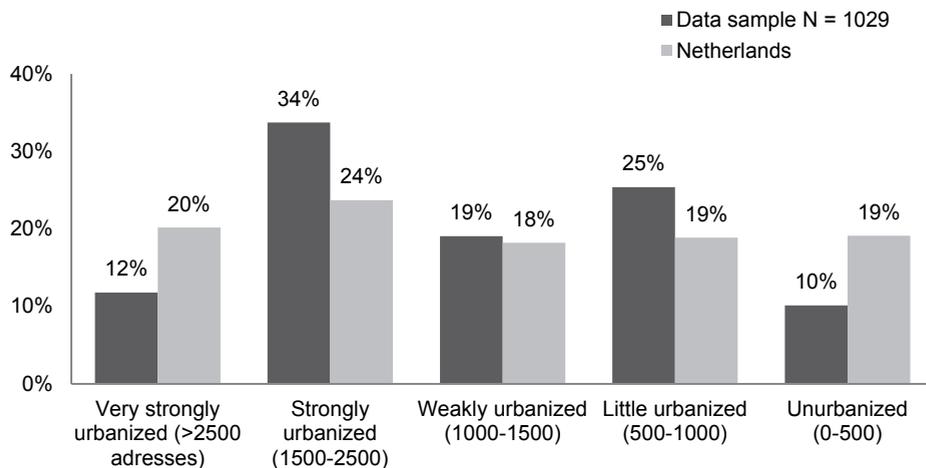
The teachers also provided some data about their schools. For example, whether they had a religious connotation (Catholic: 29.6%, or Protestant: 29.2%; in NL 29.7% and 29.8%) or not (35.8%; in NL 33.2%). Most schools (81.5%) did not have a special teaching philosophy, of the 17% that did, Dalton, Montessori, and Jenaplan were the most common ( $N_s > 20$ ).

## Teachers' assessment profiles in mathematics education



*Figure A2.* Proportion of schools per province in our sample (dark grey) and in the Netherlands (lighter grey)

The geographical distribution of the participating schools generally followed the national distribution as well (see Figure A2), nonetheless teachers from the province of Zuid-Holland were slightly overrepresented and those from Limburg underrepresented in the sample.



*Figure A3.* Percentages of urbanization levels around schools in our sample (dark grey) and in the Netherlands (lighter grey)

The urbanization level, comparable to social economic status in neighborhoods, is displayed in Figure A3. Here as well, can be seen that the distributions generally looked alike, even though there were more teachers in the sample from strongly urbanized regions and less from rural regions in comparison to the national sample. As most of these indicators are almost identical from our sample to the national percentages, our sample, even though its teachers had not been selected randomly or in a stratified way, is quite representative of the population.

## **Chapter 6**

### **Summary and discussion**

## Summary and discussion

### 1. Summary

The main goal of this PhD study was to provide insight on primary school teachers' classroom assessment practice in mathematics, with a particular focus on possibilities for improving this practice. Classroom assessment is assessment that the teacher uses to get access to students' skills and understanding in an effort to improve further instruction and move students' learning forward. Important reasons for regarding the teachers' classroom assessment practice are the often-reported effectiveness of assessment with a formative purpose and the continual striving to improve mathematics education. Assessment with a formative purpose, focused on helping the students' learning process forward, can be contrasted with other purposes of assessment (e.g., summative or evaluative), aimed at ranking or qualifying students or teachers. The usefulness of classroom assessment by the teacher to find out where students are in their understanding of mathematics in order to help them forward (i.e., a formative purpose) reposes on the assumption that this insight on students is important information to have for a teacher. This presupposes that teaching, and the ensuing learning, are not uniform one-size-fits-all concepts; the teacher needs to adapt instruction to the different students' skills and understanding at a particular moment in time. In relation to this, Von Glaserfeld (1983) pointed out that "[t]wo things are required for the teacher: on the one hand, an adequate idea of where the student is and, on the other, an adequate idea of the destination" (p. 48). To provide the teacher with such ideas classroom assessment can be an adequate tool. In this PhD thesis the results of four studies into primary school teachers' mathematics classroom assessment practice in the Netherlands are presented. The common departure point of these studies was this possibility that the use of classroom assessment to determine where students are, integrated in teachers' teaching practice, could lead to an improvement in students' mathematics performance. This appears to be of elementary logic, as a teacher who is checking students' skills and understanding, before and during the instruction of a new topic, will provide instruction that is more befitting their learning needs and as such leads to better learning results. Whether supporting teachers in the use of classroom assessment in mathematics education indeed had an incremental effect on students' mathematics achievement was the main research question guiding the studies reported on in this PhD thesis.

The results of four studies are described in the chapters constituting this thesis. The main purpose of the first study was to investigate teachers' current assessment practice in mathematics, to establish whether there was any leeway for improvement. Second, the feasibility of supporting teachers in using classroom assessment techniques in their classroom practice was regarded, for this would form the basis for enriching teachers' assessment practice. Third, it was examined whether supporting teachers to use these classroom assessment techniques led to improvement in their students' learning. Finally, meaningful profiles of teachers' assessment practice were investigated. In the following the main results per study are described in more detail.

### **1.1 Dutch mathematics teachers' current assessment practice**

As a first step towards the goal of enriching teachers' classroom assessment practice, in an effort to improve students' mathematical proficiency, teachers' current classroom assessment practice was investigated through an online questionnaire – as described in *Chapter 2*. This online questionnaire consisted of 40 questions, divided over four parts, respectively addressing teachers' background characteristics, mathematics teaching practice, assessment practice, and beliefs on assessment. In total 960 teachers from 557 Dutch primary schools responded to the questionnaire. Their answers revealed that with respect to the observation-based methods, most teachers asked questions, made observations, and corrected their students' written work weekly or more often (from 77% to 91% of the teachers), these were the most frequently applied methods by the teachers. More 'authentic' assessment methods, such as letting students present their work and keeping portfolios were rarely used (~80% of the teachers did this once a year), whereas letting students do practical assignments was a monthly activity for over half of the teachers (57%). Assessing students with either teacher-developed or student-developed problems was not very common (24%, respectively 22% of the teachers did this on a weekly basis). Concerning the instrument-based methods, almost all teachers (> 85%) used tests from the textbook and from a student monitoring system monthly, as would be expected by the directives of these tests.

The results of the survey indicated that teachers use assessment information gained from instrument-based and observation-based assessment methods for a wide range of purposes; from formative purposes, as giving feedback via

adapting instruction, to summative purposes, as determining mastery. However, the use of more authentic formative assessment methods, such as teacher-initiated whole-classroom assessment, was barely reported. In fact, Dutch teachers reported having a quite classical way of assessing their students' knowledge, in that sense that assessment methods such as practical assignments, collecting scrap paper, and teacher-developed test problems, apparently still play a minor role in mathematics education in primary school. These impressions of teachers' assessment practice showed that teachers already have formative purposes for assessment in mind to adjust their instruction to students' needs, but that the use of particular methods, namely the teacher-developed assessments, can still be encouraged.

## **1.2 The feasibility of classroom assessment techniques**

In a small-scale study the focus switched from the current assessment practice of teachers to how this practice could be improved – as reported in *Chapter 3*. In *Chapter 2* it was found that teachers rarely reported the use of teacher-developed assessment problems, therefore nine teacher-initiated classroom assessment techniques were designed on the basis of scientific literature, more practice-oriented work, and principles of assessment in Realistic Mathematics Education. Over the course of two consecutive school years different groups of in total ten teachers from ten different schools (with 214 students; 14 to 29 students per class) participated in this study. These teachers received intensive supervision in developing and execution of the classroom assessment techniques in the form of monthly workshops. Approximately every month in the second semester of Grade 3 the teachers participated in workshops of three or four teachers. The mathematics program and more specifically the key mathematics topics of the next couple of weeks were discussed in these workshops. The assessment information teachers needed in order to guide their students to their educational goals for mathematics were determined in unison. Finally, the classroom assessment techniques and didactical decisions that could be made on the basis of the assessment information were evaluated in discussions with the teachers. These assessment techniques were intended to help teachers to quickly find out something about important building blocks for mathematical skills and understandings of their students, provide teachers with indications for further instruction, and focused on some of the mathematics content of the second semester of Grade 3. The feasibility of the classroom

assessment techniques was investigated by conducting regular classroom observations of every teacher in between workshops. These observations were intertwined with short informal interviews. Additionally it was investigated whether providing teachers support in the workshops on the use of classroom assessment techniques was related to learning gains in the students. For this an explorative pre-/post-test evaluation of students' mathematics achievement was used (as measured by generally used standardized student monitoring system tests, Cito *Leerlingvolgsysteem-toetsen* LVS M5 and E5).

Teachers and students reported enjoying the techniques and finding them useful in the sense that they provided them with valuable information that supported their teaching and learning. Remarkably, students referred to the activities of the classroom assessment techniques as “mathematics games”, as such illustrating their motivation in participating in them. Teachers also mentioned that the techniques were easy to apply in their classrooms and that the workshops were an appropriate means to transmit the techniques. In terms of mathematics achievement, results indicate that these students improved considerably more than students from the national norm reference sample. Even though the treatment group was relatively small and there was no control group in this study, these results do provide an indication for the feasibility and effectiveness of the use of the classroom assessment techniques in mathematics: teachers gladly used the techniques and students appeared to advance more from the midyear to the end of the year testing than expected. This advance in student learning was related to the number of workshops in which their teacher participated. Students from teachers that participated in five workshops progressed from the middle of the year pretest to the end of the year posttest with an effect size of  $d = 0.81$  ( $M_{gain} = +9.7$  mathematics ability points), and those with a teacher that had visited four workshops showed an effect size of  $d = 0.55$  ( $M_{gain} = +7.6$  ability points). These effects are of considerable size and notably larger than that of the reference sample ( $d = 0.36$ ,  $M_{gain} = +5.1$  points), but as there was no control group the direct attribution of these effects to teachers' use of the classroom assessment techniques was not (yet) possible. To investigate whether these results would hold up in a more experimental and larger-scale design the next study was set up.

### 1.3 The effects on students' learning of supporting teachers' use of classroom assessment techniques

In this large-scale study the effectiveness of improving teachers' assessment practice was investigated – as described in *Chapter 4*. Using the results of the two previous chapters, on teachers' current assessment practice in mathematics and on the feasibility of the classroom assessment techniques, this study was set up. Since the number of workshops in which teachers participated appeared to play a role in the effectiveness and in an effort to scale-up the results from the feasibility study, in addition to further decreasing their number, the number of workshops was also experimentally varied. Thirty teachers and their 616 students participated in this experiment with pretest/posttest and control group. Teachers were randomly divided over the four conditions: a control (business-as-usual) condition, with no workshops on classroom assessment techniques, and three experimental conditions with one, two, or three workshops. In these workshops the classroom assessment techniques developed in the feasibility study were discussed with the teachers. In the conditions with two or three workshops, where teachers came back after having used the techniques, their own use of the classroom assessment techniques was also discussed and reflected upon. These workshops followed the same procedure, and the same student monitoring system tests (Cito *leerlingvolgsysteem-toetsen* LVS M5 and E5) were used as pretest and posttest, similar to the feasibility study.

Analyses of the student mathematics achievement data showed that students from teachers having participated in the condition with three workshops improved significantly more ( $M_{gain} = +8.1$  points) than students from teachers in the other conditions ( $M_{gain} \leq +5.9$ ;  $d = 0.26$ ). Students from teachers in the other experimental conditions, with one or two workshops on the assessment techniques, had comparable score gains to students in the control condition (without any workshops). Apparently a minimum of three hours of professional development was necessary for the students to benefit from their teachers' use of the classroom assessment techniques. These results indicate that supporting teachers in the use of classroom assessment techniques in mathematics can indeed improve students' mathematics achievement. This learning gain depended, however, on the number of professional development sessions the teacher attended; only in the group of teachers that participated in three one-hour sessions a significant impact on student learning was found. The finding

that students of teachers in the experimental conditions with merely one or two workshops did not progress more than students of teachers that did not have any professional development on the use of the classroom assessment techniques, was not completely unexpected. Many of the studies in which (large) positive effects for teachers' use of classroom assessment were found, were making use of very intensive prolonged professional development. So, even though the effect had slightly decreased from those found in our own earlier study with more professional development, it was still over 40% larger than the expected learning gain over this semester. The differential effect might in addition to the number of professional development sessions also have been influenced by other factors such as the teachers' motivation to use the assessment techniques or their previous assessment practice; these were not taken into account in this study.

#### **1.4 Identifying primary teachers' assessment profiles in mathematics**

To get a more comprehensive view on how teachers assess their students' learning in mathematics a secondary analysis of the questionnaire data on teachers' classroom assessment practice was carried out. As reported on in *Chapter 5*, the goal of this analysis was to understand assessment from the conglomerate of choices a single teacher is making when collecting information about his or her students' learning process. This secondary analysis was focused on identifying meaningful and useful, in a practical sense, assessment profiles of the teachers. Such assessment profiles can for one provide structure in the many characterizations that exist of assessment, but more practically, give meaningful profiles of teachers that could be used in further professional development. To identify these profiles the questionnaire data from *Chapter 2* were used. First, a number of explorative factor analyses were performed to determine the underlying (latent) factor structure of the questionnaire. A five-factor solution gave the best fit to the data. The five factors were named based on the items they contained: *Goal centeredness of assessment* (items on teachers' purposes of assessment), *Authentic nature of assessment* (items on authentic assessment methods), *Perceived usefulness of assessment* (statements on usefulness), *Diversity of assessment problem format* (items on problem formats), and *Allocated importance of assessing skills and knowledge* (items on the importance of assessing particular skills and knowledge). To be able to characterize teachers' assessment practice and assign them to different

assessment profiles a latent class analysis was performed using all variable (item) scores as input. As such differences between the latent classes of teachers on the five factors found in the separately performed factor analysis could be identified. A model with four latent classes (or assessment profiles) provided the best fitting solution.

The biggest assessment profile was that of the *mainstream assessors* (35.5% of the teachers). In this profile most teachers regularly used different types of assessment, instrument-based and observation-based, for both summative and formative purposes. On all factors teachers with this profile scored around the mean. The next biggest group (28.5%) was that of the *enthusiastic assessors*. Teachers with this profile were very aware of the different possibilities assessment offered them and used them likewise. On all components these teachers scored above the mean, with a peak on *Goal centeredness of assessment*. An almost equally large group of teachers (25.8%) were the *non-enthusiastic assessors*. These teachers viewed assessment more often in a negative way and used it accordingly less. On all factors, teachers with this profile scored below average. Finally, there were the *alternative assessors* (10.3%). Teachers with this profile had an ambiguous view of assessment. Although they reported a lot of own input in assessment and devised their own tasks and tests, they did not find assessment important or necessary. From the identification of these assessment profiles the following can be concluded: First, they permit to typify the assessments teachers perform in their classroom, and as such they can be used to propose tailor-made professional development for teachers with specific assessment profiles. Furthermore this profile characterization makes a connection between the assessment activities and assessment skills of teachers, and this connection could be used in the further development of conceptualizations and eventually theory of classroom assessment.

## **2. Conclusion and practical implications**

From the findings of the studies in this PhD thesis it becomes clear that the hypothetical teacher, as discussed in the introductory chapter, who made use of insights from the learning-teaching trajectories of primary mathematics to investigate students' skills and understanding before and during the instruction of the execution of a calculation like 77-29, is not just a hypothetical teacher.

The real teachers that participated in the studies of this PhD thesis all strived to provide the best possible instruction to their students in an effort to improve their learning, and made gladly use of the classroom assessment techniques providing them with the information enabling them to move learning forward. During a short intervention focused on instructionally worthwhile and didactically embedded classroom assessment techniques for mathematics, teachers showed to be able to improve their students' mathematics achievement.

A remarkable finding relevant for educational practice was the improvement in achievement found for students of different performance levels (cf. *Chapter 4*). It could be expected that teachers' use of the classroom assessment techniques would mainly benefit their low-performing students, as the techniques focused on revealing the understanding of building blocks of elementary mathematics topics (such as number knowledge, multiplication tables, understanding of word problems). Since low-performing students probably lack some proficiency in these, this expectation seems warranted. Indeed, low-performing students of teachers using the classroom assessment techniques showed the biggest improvement in performance by far, but in the condition with three workshops, also the high-performing students showed notable improvement. Possible explanations for this finding are that high-performing students ordinarily compensate with other strategies for a possible lack of more basic understandings and are also not often bothered with questions about basic understandings because teachers assign them the more difficult or challenging exercises. Thanks to the teachers' use of the classroom assessment techniques, it was revealed that some high-performing students did not have these insights that one would expect them to have. As such, the instruction the teachers provided subsequently was not only more adapted to the needs of low-performing students but also to the high-performing students (as evidenced in their learning gains). Through the use of the classroom assessment techniques focusing on specific and important key understandings of mathematics in Grade 3, the teachers and their high-achieving students could have become aware of some weaknesses these students had, that had otherwise been masked by their high achievements or good performance on other tasks. This indicates how teachers' use of classroom assessment techniques can not only benefit the lower performing students but also the high-performing students and this is in the context of improving the mathematics education in general an encouraging finding.

These results are directly applicable into practice; this was also one of the main purposes of this PhD thesis. Due to the step-by-step approach, from the investigation of the feasibility with a small number of teachers and regular meetings, to the scaling up with a larger group of teachers and a more limited number of meetings, the materials and professional development in the form of workshops can almost immediately be used in general educational practice. Taking account of the current assessment practice of teachers and the effects on student learning illustrated in this thesis, teacher counsellors at teacher advisory centers (*onderwijsbegeleidingsdienst*) and teacher educators at teacher education colleges (*PABO*), are already working on integrating the classroom assessment techniques developed in this PhD project in their (professional development) courses. For the concrete Dutch educational practice this means that the insights this thesis provides on the feasibility and effectiveness of teachers' use of the classroom assessment techniques can soon be put into practice on an even larger scale.

In such professional development a number of the findings of this thesis have to be taken into account. Important in this dissemination is that close attention is paid to the organization of the workshops and the adaptive nature of the discussions of the techniques. A first recommendation is to have teachers actively participate in multiple workshops on the use of classroom assessment techniques. As we found in the feasibility study (described in *Chapter 3*), teachers felt empowered by the fact that they could adapt the classroom assessment techniques to fit to their own practice and could reflect upon this in the discussions in the following workshops. By having these reflective discussions teachers developed more ownership of the classroom assessment techniques and were as such positively inclined to use them, and the information gathered by them, in their further teaching. This form of teacher learning community has been around for quite some time and has also been advocated to use with teachers developing their classroom assessment skills. It became clear that the ownership teachers felt was one of the main reasons for them to implement what they had discussed, seen, and heard in the workshops, in their practice. In large-scale implementation of professional development, the sessions and materials often become more or less rigid of design due to practical considerations of scaling up and trying to reach a level of uniformity. This would, however, most probably not lead to the satisfactory learning results we

found in this PhD study. For example, in the United Kingdom, where formative assessment was implemented nationwide, it was found that many teachers used the formative assessment activities *à contrecoeur* and barely for formative purposes, because they did not feel ownership of these, as they had not been included in the design and the decision-making process of the assessment techniques. The effective format of the professional development as realized in this PhD project with at least three meetings on the use of classroom assessment techniques and the relatively large effect this had on student mathematics achievement, clearly suggests continuing investigating the techniques and assist teachers in using them. The awareness of students' mathematics skills and knowledge teachers develop while using the classroom assessment techniques can be of use in pre-service and in-service mathematics teacher education.

### **3. Suggestions for further research**

The research in this PhD project was focused on Grade 3 of primary school mathematics education in the Netherlands. Evidently, the results on the classroom assessment techniques and teachers' assessment practice can also be used in different grades. Most of the techniques can probably quite easily be adapted to fit the teaching and learning trajectories of the other primary grades by teachers themselves. However, research into this adaptation of the techniques to other grades would be necessary to test whether this expectation holds any truth. The adaptation of the techniques for use in different cultural and educational contexts is another road to pursue. Some first experiences with this in primary school mathematics classrooms in China (cf. Zhao, Van den Heuvel-Panhuizen & Veldhuis, 2015) showed that a mere translation of the techniques in a different language was not enough. A careful analysis of the mathematics curriculum, the mathematics teachers' teaching practice, and the used textbooks was required first. Even after these analyses and subsequent adaptations to the classroom assessment techniques, their use demanded an important change in Chinese teachers' and students' learning culture, which was not easily accomplished. Investigations such as these into the implementation of the classroom assessment techniques in yet other educational or cultural contexts can provide us with more information on their functioning in general and help improving mathematics education around the world.

The teachers participating in the studies in the Netherlands clearly used the classroom assessment techniques to their and their students' advantage. However, it has not yet been investigated whether these effects persist over a prolonged period of time, with a follow-up study. Whether in subsequent school years the students having shown more improvement kept improving at a higher rate, or whether new students of teachers that had been using the classroom assessment techniques also benefitted from this, are questions that remain to be answered. We conjectured that teachers became more aware of the value of assessment for instruction, due to using classroom assessment techniques fitting their mathematics teaching practice and participating in at least three workshops with discussions and reflections on their use. However, we cannot exclude the possibility that the teachers merely used these classroom assessment techniques for the time that the intervention took place and did not develop any awareness or proneness to use them in their further practice. Without fail the participating teachers said that they would continue using the techniques and adapt them to the subjects of the first semester of Grade 3 as such providing some indirect evidence for their changed assessment practice. Many of them also clearly felt ownership of the techniques as they had adapted them to fit their own teaching practice. But such projections are very easily made at the end of some professional development, whereas after, for example, the summer holiday the contents and usefulness have lost importance in the minds of the teachers and are only present in one of the folders containing the documentation of this professional development. This folder generally has many neighboring folders containing teachers' documentation from earlier professional development. It would thus be very useful to investigate in future research whether teachers that used the classroom assessment techniques really continue with their changed assessment practice, especially since in earlier research it was found that promising results of mathematics improvement programs do not always persist from one school year to the next (e.g., Houtveen, van de Grift, & Creemers, 2004).

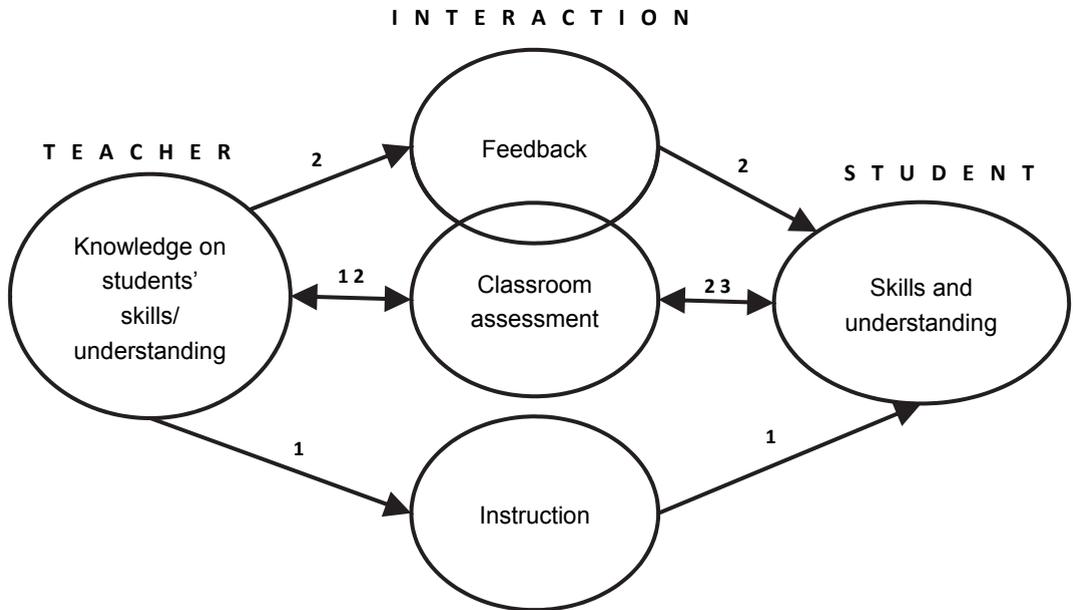
Another avenue that could be investigated as a follow-up to the studies in this PhD project is the mathe-didactical knowledge of teachers or their mathematical knowledge for teaching. This is the combined knowledge about the learning and teaching trajectories of mathematics and that of the appropriate learning situations and classroom teaching practices. The classroom assessment techniques provided to the teachers in our professional development were

explicitly designed to fit to the teaching and learning trajectory of third grade mathematics and linked to the important didactical decisions. As Moreland, Jones and Moreover (2001) concluded “[e]ffective assessment is dependent on informed assessors who are able to interpret observations and student outcomes [...] [f]ormative interactions with students become distorted if there is a lack of subject knowledge and how the subject knowledge is constructed” (p. 157-158). The mathe-didactical knowledge, i.e. subject and didactical knowledge of mathematics, of teachers was neither measured nor questioned in our studies. One of the assumptions was that teachers were experts in their own right on primary school mathematics education, but only lacked some applied didactical knowledge on the use of classroom assessment in mathematics. However, it is very well possible that teachers’ mathematical knowledge for teaching is not always sufficient to optimally interpret the assessment information they gathered through the use of the classroom assessment techniques. Investigating teachers’ mathe-didactical knowledge in relation to their classroom assessment practice can be a worthwhile extension of this research project. The assessment profiles identified in *Chapter 5* could be used for this too. A possible course of action could be to relate teachers’ assessment profiles and mathe-didactical knowledge to the effect of the use of classroom assessment techniques, as such providing evidence for – or disproving – the probable relationship between these teacher characteristics and student learning gains.

#### **4. Final remarks**

The ICA project reported on in this thesis was set up to investigate teachers’ classroom assessment practice in primary school mathematics education and, as the title of the project indicates, to *improve* this practice where possible. Concluding from the research findings it becomes clear that the short professional development on the use of classroom assessment techniques helped teachers to improve their students’ mathematics achievement. Through knowing what their students know, teachers can evidently better help their students move forward in their mathematics learning. Additionally, through the use of the classroom assessment techniques teachers become more aware of how assessment information can shape their teaching and which understandings of students are pivotal in the teaching-learning trajectory of Grade 3 in primary mathematics education. The assessment profiles that were distinguished on the basis of teachers’ assessment practice in mathematics provide viable openings for future tailor-made professional developments.

The connections between classroom assessment, instruction, teachers' knowledge of students' skills and understanding, and students' skills and understanding are illustrated in Figure 1. This figure provides an illustration for how I experienced the central role classroom assessment plays in primary teachers' practice in mathematics. The numbers 1, 2, and 3 depict the three main influences of teachers' use of classroom assessment on students' skills and understanding. The first route (1) shows the indirect influence of the information on students' skills and knowledge gained from classroom assessment that enriches the teacher's knowledge of students and as such influences his/her instruction, and consequently the students' learning. The second influence (2) shows how the classroom assessment techniques provide direct feedback to the students on their own functioning and indirectly through nourishing the teacher's knowledge and following feedback to the students. This reflects Kulhavy's (1977) assertion that "the process [of feedback] itself takes on the forms of new instruction" (p. 2). The third influence (3) is direct: in partaking in classroom assessment activities the students develop and practice their own knowledge and understanding of important topics in mathematics. The intersection of feedback and assessment practice exists because in many cases, while participating in an assessment activity, students immediately receive feedback on their own performance, whether it is from the teacher, their peers, or themselves. Manifestly this schema is not complete, for example the complete background of the teacher is missing, such as his/her mathe-didactical knowledge, but it nicely conveys how we perceive the role of classroom assessment in relation to students' skills and understanding (see Heritage, 2010, p. 11, for a more elaborate illustration of different steps in classroom assessment).



*Figure 1.* A schematic impression of the connections between classroom assessment, instruction, teachers' knowledge of students' skills and understanding, and students' skills and understanding. The numbers next to the arrows correspond to the different influences on students' skills and understanding: the influence of (1) teachers' adaptation of instruction on the basis of assessment information, (2) the provision of feedback to students, and (3) students' participation in worthwhile assessment tasks on important subdomains in mathematics education.

In conclusion I like to mention a typo I often made in the process of writing on my thesis. This typo was *assessment technique* instead of *technique*. Even though we strived for the teachers to distinguish between moments of instruction and of classroom assessment (cf. Figure 1), the term *assessment technique* nicely conveys the two joint purposes of the classroom assessment techniques: on the one hand, the assessment of building blocks of students' skills and understanding of mathematics and, on the other, the ensuing adaptation of instruction to students' learning needs. The intrinsic and complementary link between teaching and assessment is in this term beautifully

## Chapter 6

expressed. Evidently, many of the characteristics of classroom assessment are also part of good teaching practice, as Ginsburg (2009) wrote:

Good teaching [...] sometimes involves the same activities as those comprising formative assessment: understanding the mathematics, the trajectories, the child's mind, the obstacles, and using general principles of instruction to inform the teaching of a child or a group of children. (p. 126)

Based on my experiences I hypothesize that helping teachers to use these principles in their practice through providing them with classroom assessment techniques (or *techniques*), as was done in the research studies reported on in this PhD thesis, can contribute to even further improving mathematics teaching practice and at the same time student achievement in mathematics.

**References**

- Ginsburg, H.P. (2009). The challenge of formative assessment in mathematics education: children's minds, teachers' minds. *Human Development*, 52, 109-128.
- Glaserfeld, E. von (1983). *Learning as constructive activity*. Presented at the 5th annual meeting of the North American Group of Psychology in Mathematics Education, Montreal.
- Heritage, M. (2010). Assessment with and for students. In M. Heritage, *Formative Assessment: Making It Happen in the Classroom* (Chapter 2). Thousand Oaks, CA: Corwin.
- Houtveen, A. A. M., van de Grift, W. J. C. M., & Creemers, B. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 15(3-4), 337-376.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, 47(1), 211-232.
- Moreland, J., Jones A., & Northover, A. (2001). Enhancing teachers' technological knowledge and assessment practices to enhance student learning in technology: A two-year classroom study. *Research in Science Education*, 31(1), 155-176.
- Zhao, X., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (2015). *Chinese teachers' use of classroom assessment techniques in primary mathematics education*. Manuscript submitted for publication.



**Samenvatting**

**Dankwoord**

**Curriculum vitae**

**List of publications related to this thesis**

**List of presentations related to this thesis**

**FIsme Scientific Library**

**ICO Dissertation Series**

## Samenvatting

Het doel van dit promotieonderzoek was inzicht verschaffen in de toetspraktijk bij rekenen-wiskunde van leerkrachten in het basisonderwijs. In het bijzonder is hierbij gekeken naar mogelijkheden tot het verbeteren van deze toetspraktijk. De toetsen waar het hierom gaat zijn alle activiteiten die leerkrachten kunnen gebruiken om zicht te krijgen op het kennen en kunnen van leerlingen met het doel verdere instructie hierop aan te passen en zo het leren van de leerlingen te bevorderen. Zo een toets kan bijvoorbeeld bestaan uit een klassikale activiteit waar alle leerlingen tegelijkertijd aan meedoen evenals het observeren van leerlingen terwijl ze opdrachten aan het maken zijn, en is dus niet per definitie een schriftelijke overhoring. Deze vorm van toetsen is bekender onder de namen *formative* of *classroom assessment*. Zulke formatieve toetsen dienen expliciet om het leren van leerlingen te stimuleren, deze vorm van toetsen wordt vaak gecontrasteerd met summatieve toetsen, waarvan het doel juist is om leerlingen te ordenen of certificeren. Belangrijke overwegingen om het gebruik van formatieve toetsen in het reken-wiskundeonderwijs te onderzoeken zijn enerzijds de vaak gerapporteerde leerbevorderlijke effecten van het gebruik van formatieve toetsen door de leerkracht en anderzijds het voortdurende streven naar het verbeteren van het rekenwiskundeonderwijs. De centrale onderzoeksvraag van dit promotieonderzoek was dan ook: Heeft het ondersteunen van leerkrachten in het gebruik van formatieve toetstechnieken bij rekenen-wiskunde een positief effect op de leerresultaten van hun leerlingen?

In dit proefschrift zijn de resultaten van vier studies naar de toetspraktijk bij rekenen-wiskunde van leerkrachten in het basisonderwijs beschreven. De huidige toetspraktijk van basisschoolleerkrachten bij rekenen-wiskunde is allereerst beschreven, om zo te duiden of er ruimte voor verbetering was. In de tweede plaats stond het uitzoeken van de uitvoerbaarheid en indicaties voor de effectiviteit van het gebruik van speciaal ontworpen formatieve toetstechnieken voor rekenen-wiskunde centraal. Daarna is het effect van het begeleiden van leerkrachten in het gebruik van de formatieve toetstechnieken op leerprestaties van de leerlingen experimenteel getoetst. Ten slotte is uitgezocht of betekenisvolle profielen van de toetspraktijk van de leerkrachten konden worden vastgesteld.

## Hoofdstuk 2. De huidige toetspraktijk bij rekenen-wiskunde in Nederland

Als een eerste stap richting het verrijken van de toetspraktijk van leerkrachten, met het doel de leerprestaties van leerlingen bij rekenen-wiskunde te verbeteren, is de huidige toetspraktijk van de leerkrachten onderzocht. In *Hoofdstuk 2* zijn de resultaten van een online vragenlijstonderzoek beschreven. Deze vragenlijst bestond uit 40 vragen, over vier deelonderwerpen, namelijk de onderwijsachtergrond van leerkrachten, hun reken-wiskundeonderwijspraktijk, hun toetspraktijk en hun mening over toetsen. In totaal zijn de antwoorden van 960 leerkrachten van 557 verschillende Nederlandse basisscholen verkregen en geanalyseerd. De gegeven antwoorden lieten zien dat wat betreft de observationele methoden, leerkrachten het vaakst vragen stelden, observeerden en leerlingwerk corrigeerden. Deze toetsactiviteiten ondernamen de meeste leerkrachten wekelijks of vaker (van 77% tot 91% van de leerkrachten minimaal één keer per week). Meer, zogenoemde authentieke, toetstechnieken zoals leerlingen hun werk laten presenteren of het bijhouden van portfolio's werd zelden gedaan (80% van de leerkrachten op zijn hoogst jaarlijks), terwijl meer dan de helft van de leerkrachten (57%) maandelijks praktische opdrachten aan hun leerlingen gaven. Het toetsen van leerlingen met door de leerkrachten ontwikkelde toetsen werd minder vaak gerapporteerd (24% van de leerkrachten gaf aan dit wekelijks te doen). Daarnaast werden ook instrumentele technieken gebruikt, bijna alle leerkrachten (>85%) gebruikten de toetsen van de methodes of uit het leerlingvolgsysteem maandelijks tot halfjaarlijks zoals ook te verwachten valt als naar de handleidingen van deze instrumenten wordt gekeken. Leerkrachten bleken toetsen het vaakst te gebruiken voor formatieve doeleinden, zoals voor het geven van feedback, het vaststellen van de instructiesnelheid of het aanpassen van de instructie. Minder vaak gebruikten ze toetsen voor summatieve doeleinden, zoals het selecteren van specifieke onderwerpen of het vaststellen van vooruitgang. Dit was te verwachten, aangezien formatieve toetsen, gericht op het vooruithelpen van het leerproces van leerlingen, per definitie vaker worden ingezet dan summatieve toetsen, gericht op het vaststellen van het niveau van leerlingen. In hoeverre leerkrachten het gebruik van toetsen nuttig vonden werd geïllustreerd door de positieve waarderingen die ze gaven aan de verschillende vormen en doelen van toetsen. De twee meest gebruikte instrumentele toetsmethoden, methode- en LOVS-toetsen, werden als meest relevant voor hun onderwijspraktijk

beoordeeld, met daarna de observationele technieken van vragen stellen en het observeren van leerlingen.

Leerkrachten bleken dus de informatie die ze dankzij het toetsen verkrijgen voor verschillende doeleinden te gebruiken, van de meer formatieve doelen als het geven van feedback of het zorgen voor aangepaste instructie tot het summatieve doel van het vaststellen van het niveau van leerlingen. Wat opvalt, is dat door de leerkracht zelf ontwikkelde en gestuurde toetsen zelden werden gebruikt. Terwijl dit soort toetsen, en dan vooral de klassikale varianten, het type toetsen zijn waarvan de effectiviteit voor het bevorderen van leerresultaten vaak beschreven is. Feitelijk geven de onderzochte Nederlandse leerkrachten aan op een vrij klassieke wijze te toetsen; in die zin, dat toetstechnieken zoals het bekijken van kladpapier en het gebruik van door de leerkracht ontwikkelde opgaven weinig worden gebruikt. Deze bevindingen gaven des te meer aanleiding voor het ontwikkelen van speciaal op de praktijk van de leerkracht toegespitste formatieve toetstechnieken.

### **Hoofdstuk 3. De uitvoerbaarheid van formatieve toetstechnieken**

In een kleinschalige vervolgstudie verschoof de focus van de huidige praktijk van de leerkrachten naar hoe deze praktijk kon worden verrijkt. De resultaten van deze studie zijn beschreven in *Hoofdstuk 3*. Gegeven het feit dat leerkrachten zelden zelf ontwikkelde toetsen gebruiken, zijn er negen leerkracht-gestuurde toetstechnieken ontwikkeld op basis van wetenschappelijke literatuur, praktijkgerichte publicaties en principes van het toetsen in het realistisch rekenonderwijs. Gedurende twee schooljaren is samengewerkt met verschillende groepen leerkrachten, met in totaal tien leerkrachten van groep 5 van tien verschillende scholen (met 214 leerlingen, tussen de 14 en 29 per klas). Deze leerkrachten kregen begeleiding in het gebruik van de ontwikkelde toetstechnieken in de vorm van maandelijke workshops. In het eerste schooljaar waren er vijf workshops waaraan vier leerkrachten deelnamen, in het tweede schooljaar waren er twee groepen van drie leerkrachten die aan vier workshops meededen. In deze workshops stond steeds het reken-wiskundeprogramma van de daaropvolgende weken centraal, in het bijzonder de belangrijkste struikelblokken die de leerkrachten voorzagen bij hun leerlingen. Zo werd in samenspraak de benodigde informatie die de leerkrachten over hun leerlingen dienden te vergaren om hun leerproces richting

de leerdoelen te bevorderen vastgesteld. Ten slotte kregen de leerkrachten dan toelichting over een aantal voorbeelden van op deze onderwerpen toegespitste toetstechnieken. Deze toetstechnieken waren er steeds op gericht om de leerkrachten te helpen snel uit te vinden hoe het met het begrip of beheersing van een bepaald reken-wiskundig onderdeel van leerlingen staat, waarbij dit meteen aangrijpingspunten verschaft voor verdere instructie. Elke toetstechniek werd ook expliciet als aanpasbaar aangereikt, leerkrachten werd op het hart gedrukt om de precieze uitvoering van de technieken vooral aan te passen aan de eigen onderwijspraktijk. In de daaropvolgende workshop werd het gebruik van de technieken gezamenlijk geëvalueerd en bediscussieerd alvorens het vizier op de volgende periode te richten. De uitvoerbaarheid van de toetstechnieken werd onderzocht met behulp van regelmatige observaties van de klassen tussen de verschillende workshops. Deze observaties werden gecompleteerd met korte informele interviews over de keuzes van leerkracht bij het gebruik van de toetstechnieken. Om daarnaast ook een indicatie te verkrijgen van de effecten van het gebruik van de toetstechnieken op leerprestaties werd een voor- en natoets gebruikt; hiertoe zijn gegevens van het Cito leerlingvolgstelsel rekenen LVS M5 en E5 met elkaar vergeleken.

Leerkrachten en leerlingen bleken de toetstechnieken leuk te vinden, leerlingen hadden het vaak over “rekenspelletjes”. De leerkrachten vonden het ook nuttige activiteiten die waardevolle informatie gaven voor verdere instructie. Daarbij vonden ze de technieken gemakkelijk in te passen in hun onderwijspraktijk en dat de workshops een goede wijze was om ze zich eigen te maken. Ook bleken hun leerlingen aanmerkelijk meer vooruitgang te boeken in hun reken-wiskundevaardigheid dan leerlingen van de nationale normsteekproef. Ondanks het feit dat de groep deelnemende leerkrachten relatief klein was en er geen controlegroep was, gaven deze leerresultaten ook een indicatie voor de effectiviteit van het gebruik van deze formatieve toetstechnieken. Opvallend was dat leerlingen van leerkrachten die aan vijf workshops hadden meegedaan, meer vooruitgang lieten zien van voor- naar natoets (gemiddelde toename = +9.7 vaardigheidspunten,  $d = 0.81^1$ ) dan leerlingen van leerkrachten die vier

---

<sup>1</sup>  $d$  is een maat voor de grootte van een effect; deze maat is op de volgende wijze berekend:  $d = (\text{gemiddelde natoets} - \text{gemiddelde voortoets}) / \text{gepoolde standaarddeviatie}$ . Dit betekent dat het verschil tussen voortoets en natoets is uitgedrukt in het aantal gemiddelde standaarddeviaties dat dit bedraagt.

workshops hadden bijgewoond (gemiddelde toename = +7.6 vaardigheidspunten,  $d = 0.55$ ). Dit zijn grote leereffecten en daarbij ook beduidend groter dan die in de referentiesteekproef waren gevonden (gemiddelde toename = +5.1 vaardigheidspunten,  $d = 0.36$ ). Door het ontbreken van een controlegroep is het direct toewijzen van deze leerwinst aan het gebruik van de toetstechnieken door de leerkrachten niet geoorloofd, echter sprak het wel tot de verbeelding om dit meer gecontroleerd verder uit te zoeken, hiertoe is de volgende studie opgezet.

#### **Hoofdstuk 4. Leereffecten van de begeleiding bij formatieve toetstechnieken**

In een grootschalig onderzoek, beschreven in *Hoofdstuk 4*, is een meer controleerde opzet gebruikt om het leereffect van het gebruik van formatieve toetstechnieken vast te stellen. De resultaten van het kleinschalige onderzoek hierboven beschreven over de uitvoerbaarheid van de toetstechnieken in aanmerking nemende, werd de volgende quasi-experimentele onderzoeksopzet gebruikt. Aangezien het aantal bijgewoonde workshops leek uit te maken voor de gevonden mate van vooruitgang bij de leerlingen en om te vinden of dit aantal lager zou kunnen, werd het aantal workshops experimenteel gevarieerd in deze studie. Dertig leerkrachten en hun 616 leerlingen namen deel aan dit onderwijsexperiment met voortoets, natoets en controlegroep. De leerkrachten werden willekeurig verdeeld over de vier condities: een controle conditie, zonder workshops over toetstechnieken, en drie experimentele condities, waarin één, twee of drie workshops over de toetstechnieken werden aangeboden. In deze workshops werden dezelfde toetstechnieken als in *Hoofdstuk 3* besproken. In de condities met twee of drie workshops, waar de leerkrachten terug kwamen na de technieken te hebben gebruikt, werd in de volgende workshop(s) ook gereflecteerd op het gebruik van de technieken; dit gebeurde logischerwijs niet in de eerste experimentele conditie met slechts één workshop. Verder was de opzet van deze workshops identiek aan die van de workshops in *Hoofdstuk 3*. Ook hier werden de resultaten van hun leerlingen op toetsen van het leerlingvolgsysteem gebruikt als voor en nameting (LVS M5 en E5).

De analyses van de gegevens van de leerlingen op de voor en natoets duiden op een grotere vooruitgang van leerlingen van leerkrachten uit de conditie met drie

workshops (gemiddelde vaardigheidstoename = +8.1 punten) dan leerlingen van leerkrachten uit de overige condities (gemiddelde vaardigheidstoename  $\leq$  +5.9 punten,  $d = 0.26^2$ ). Deze leerlingen boekten dus 0.26 standaarddeviatie meer vooruitgang dan leerlingen uit de overige condities (alsook die uit de normsteekproef). De effectgrootte van voor- naar natoets voor deze groep met drie workshops was van dezelfde orde van grootte als die gevonden in de studie op kleinere schaal bij vier workshops, namelijk  $d = 0.55$  (in *Hoofdstuk 3* was ook  $d = 0.55$  gevonden). Leerlingen van leerkrachten uit de andere experimentele condities (met één of twee workshops) hadden overigens vrijwel dezelfde scoretoename als de leerlingen van leerkrachten uit de controle conditie (zonder workshops). Deze resultaten duiden erop dat het bijstaan van leerkrachten in het gebruik van formatieve toetstechnieken daadwerkelijk bij kan dragen aan het bevorderen van de prestaties van leerlingen. Deze leerwinst hangt wel af van het aantal nascholingsessies dat de leerkrachten bijwonen; alleen in de conditie met drie sessies van één uur werd een significant grotere leerwinst gevonden. De bevinding van geen groter leereffect in de experimentele groepen met één of twee workshops was niet onverwacht. In veel van de onderzoeken waarin (grote) positieve leereffecten van het gebruik van formatieve toetstechnieken werden gevonden, was dit pas na een intensief en langdurig begeleidingstraject. In dit onderzoek was het uiteindelijke toegevoegde leereffect gelijkwaardig aan die in de voorgaande studie met meer begeleiding, daarbij was de leervooruitgang ruim 40% groter dan de normale vooruitgang over deze periode. Het gedifferentieerde effect kan naast door het aantal workshops ook zijn beïnvloed door andere factoren waar niet voor gecontroleerd was in deze studie, zoals de motivatie van de leerkracht voor het gebruik van de toetstechnieken of de reeds bestaande toetspraktijk.

---

<sup>2</sup> Deze effectgrootte gaat om het verschil tussen de condities en is dus als volgt berekend:  $d = (\text{gemiddelde derde experimentele conditie} - \text{gemiddelde controleconditie}) / \text{gepoolde standaarddeviatie}$ .

## Hoofdstuk 5. Het vaststellen van toetsprofielen van leerkrachten

In dit onderzoek – beschreven in *Hoofdstuk 5* – is een secundaire, diepgravende analyse van de vragenlijstdata over de toetspraktijk van leerkrachten uitgevoerd. Hiervan was het doel om het toetsen door de leerkracht te karakteriseren naar gelang het type keuzes hij of zij maakt bij het verzamelen van informatie over de leerlingen. Deze secundaire analyse richtte zich in het bijzonder op het vaststellen van betekenisvolle toetsprofielen van leerkrachten, die wellicht gebruikt kunnen worden in verdere nascholing op dit gebied. Om deze profielen vast te stellen zijn de antwoorden van leerkrachten op de vragenlijst die in *Hoofdstuk 2* is beschreven gebruikt. Allereerst is met een aantal exploratieve factoranalyses de onderliggende (latente) factorstructuur van de vragenlijst onderzocht. Een oplossing met vijf factoren bleek het beste te passen. Op basis van de bijbehorende items kregen de factoren de volgende namen: *Doelgerichtheid van het toetsen* (items over de doelen van het toetsen), *Authenticiteit van het toetsen* (items over het gebruik van authentieke toetstechnieken), *Relevantie van toetsen* (items over de relevantie van toetsen), *Diversiteit van probleemformaat* (items over de typen items bij het toetsen) en *Belang van het toetsen van vaardigheden en begrip* (items over het belang van het toetsen van deze zaken). Om hun toetspraktijk in kaart te brengen en de leerkrachten aan verschillende toetsprofielen toe te wijzen, is een latente klasse analyse uitgevoerd met alle itemscores als input. Op deze manier kon gekeken worden of er verschillen waren in de toetspraktijk van leerkrachten van verschillende profielen op de vijf factoren. Het model met vier latente klassen (profielen) bleek het beste te passen.

De grootste groep leerkrachten hoorde bij het profiel van de *mainstream toetsers* (35.5% van de leerkrachten). Leerkrachten met dit profiel gebruikten geregeld verschillende soorten toetsen, observationele en instrumentele, voor zowel summatieve als formatieve doelen. Op alle factoren scoorden deze leerkrachten rondom het gemiddelde. De tweede groep, qua grootte, waren de *enthousiaste toetsers* (28.5% van de leerkrachten). Leerkrachten met dit profiel waren zich zeer bewust van de verschillende mogelijkheden van toetsen en gebruikten ze hier dan ook veelvuldig voor. Op alle factoren scoorden deze leerkrachten boven het gemiddelde, met een uitschieter bij *Doelgerichtheid van het toetsen*. Een bijna even grote groep waren de *minder enthousiaste toetsers*

(25.8% van de leerkrachten). Deze leerkrachten vonden toetsen in het algemeen niet zo nuttig en gebruikten ze dus ook minder. Op alle factoren scoorden ze onder het gemiddelde. Ten slotte waren er nog de *alternatieve toetsers* (10.3% van de leerkrachten), die een wat ambigue opstelling ten opzichte van toetsen lieten zien. Deze leerkrachten hadden bovengemiddeld eigen inbreng in het toetsen, maar vonden toetsen verder noch belangrijk noch nuttig. Op basis van deze toetsprofielen kan het volgende worden geconcludeerd. Ten eerste kan met de toetsprofielen de toetspraktijk van de leerkrachten beschreven worden, waardoor ze gebruikt kunnen worden om op maat gemaakte nascholing aan te bieden. Daarbij maakt deze karakterisering een verbinding tussen de toetspraktijk en toetsvaardigheden van de leerkrachten, die goed gebruikt kan worden in de ontwikkeling van verdere ideeën en theorieën over toetsen.

## 6. Conclusie

In het laatste hoofdstuk worden de resultaten van deze verschillende onderzoeken samengevat en is gekeken naar de betekenis hiervan voor de onderwijspraktijk en verder onderzoek. Al met al blijkt uit de resultaten van de onderzoeken in dit proefschrift dat het ondersteunen van leerkrachten bij het gebruik van formatieve toetstechnieken bij rekenen-wiskunde inderdaad tot leerwinst bij de leerlingen kan leiden. Hiermee is andermaal de effectiviteit van het gebruik door leerkrachten van dit soort toetsen, gericht op het vooruit helpen van het leerproces, aangetoond. Dit is een bemoedigend resultaat voor de onderwijspraktijk van rekenen-wiskunde in Nederland, omdat deze leerwinst al gevonden werd na een kort nascholingstraject, van in totaal drie uur. Met een relatief kleine tijdsinvestering voor de leerkracht werd al een behoorlijk leereffect bij de leerlingen gevonden. In het voortdurende streven naar het verbeteren van het reken-wiskundeonderwijs kan het gebruik van deze formatieve toetstechnieken door leerkrachten een belangrijke rol spelen.

Door de stapsgewijze opzet van de onderzoeken, van kleinschalige studies om de haalbaarheid en uitvoerbaarheid te bekijken, naar een grootschaliger onderzoek waarin de effectiviteit werd onderzocht, kunnen de materialen en de opzet van de nascholing vrijwel direct overgenomen worden als blauwdruk voor de algemene onderwijspraktijk. In een samenwerkingsverband met een onderwijsbegeleidingsdienst en een PABO wordt nu gewerkt aan het aanbieden

van de formatieve toetstechnieken voor rekenen-wiskunde aan zittende en toekomstige leerkrachten van het basisonderwijs. Belangrijk in deze nascholing is dat rekening wordt gehouden met een aantal kenmerken, die in dit promotieonderzoek heel belangrijk werden gevonden door de deelnemende leerkrachten. In de eerste plaats, het feit dat de technieken expliciet als aanpasbaar werden aangeboden, leerkrachten konden ze zo aanpassen dat ze pasten bij de eigen onderwijspraktijk. Hierdoor maakten de leerkrachten zich de technieken eigen en gingen ze deel uitmaken van hun eigen repertoire van toetstechnieken. Ten tweede zijn ook de reflectieve discussies in de workshops als heel waardevol ervaren. In een nascholing zouden zulke terugkomdagen zeker ook geïntegreerd moeten worden.

Vanzelfsprekend zijn er ook nog genoeg vragen open voor verder onderzoek. Allereerst is er de vraag van het lange termijn effect van de ondersteuning in het gebruik van de formatieve toetstechnieken bij rekenen-wiskunde en de rol die het toetsprofiel van de leerkracht daarin speelt. Leerlingen die tijdens de nascholing in de klassen van de deelnemende leerkrachten zaten lieten een vergrote leerwinst zien, maar of deze voorsprong behouden blijft of dat er het volgende jaar een terugval is, zijn nog open vragen. Het zou goed kunnen dat het toetsprofiel van de leerkracht daar een invloedrijke rol in speelt. Ook of, en hoe, de deelnemende leerkrachten de toetstechnieken in volgende jaren gebruiken of als ze andere leerjaren onderwijzen is nog niet onderzocht. De internationale betekenis van de resultaten is ook nog een open vraagstuk. Weliswaar zijn er eerste ervaringen in China met het gebruik van deze zelfde soort formatieve toetstechnieken, waaruit bleek dat een simpele vertaling niet voldoende was. Voor de toetstechnieken daar gebruikt konden worden diende eerst een uitvoerige analyse gemaakt worden van het Chinese reken-wiskundecurriculum, de onderwijspraktijk en de lesmethoden. Zelfs na deze analyses en uitvoerige aanpassingen in de formatieve toetstechnieken was het nog veelgevraagd voor de Chinese leerkrachten, aangezien de technieken een compleet andere toets- en leercultuur veronderstellen, waarin toetsen dienen om van te leren en om mee af te rekenen. Deze resultaten laten zien dat wellicht ook voor andere leerjaren in het Nederlandse onderwijs eerst zo een terdege analyse van de leerlijnen gemaakt zou moeten worden. Dit soort onderzoeken in andere (onderwijs)culturen is dan ook bijzonder waardevol voor het verdere begrip van de werking en waarde van de formatieve toetstechnieken.

## Dankwoord

Allereerst, en bovenal, Marja. Ik kan je eigenlijk niet genoeg bedanken, zonder jou was dit proefschrift er niet geweest. Dank voor alles. Dank voor je onuitputtelijke kennis van zaken, je enthousiasme, je inspirerende ideeën, je uitgesproken meningen, je aandacht en dagelijkse nabijheid in begeleiding, professioneel en persoonlijk, je precisie in formuleringen, je onvermoeibare werk aan artikelen, je ongelooflijke inzet en nog veel meer. Dank ook voor je vrijwel grenzeloze vertrouwen in mij en mijn kunnen, dat heeft me bijzonder gesterkt gedurende dit promotietraject. Ik prijs mezelf gelukkig dat ik door zo een bijzonder en wijs mens begeleid heb mogen worden en dankzij jou zoveel heb kunnen leren. Ik hoop dit de komende jaren voort te zetten.

Hartelijk dank ook aan collega's buiten Utrecht die bij het ICA project betrokken waren: Conny Bodin van Onderwijsadvies, Arno van Houwelingen van de Haagse Hogeschool, Roel Pots en collega's bij OPOZ voor de hulp bij het opzetten van het onderzoek, het beschikbaar stellen van vergaderruimte, en het werven van leerkrachten voor de ontwikkelfase van het onderzoek. Ook de collega's van Cito/Universteit Twente, Anton Béguin, Theo Eggen, Floor Scheltens en Jorine Vermeulen gedurende het project goede uitwisselingen gehad, dank voor de vruchtbare samenwerking.

Alle leerkrachten en schooldirecties die hebben meegewerkt aan de deelonderzoeken. Extra veel dank aan de mede-ontwikkelaars van Jaar 1 in Zoetermeer: Jolanda (Piramide), Rosanne (Triangel), Sonia (Zwanenbos) en Angela (Mr. Verwers). Dank voor jullie tijd, gastvrijheid en wil om mee te werken aan een nog ietwat onduidelijk omschreven project. Dankzij jullie zijn de technieken zo geworden als ze zijn. Ook de leerkrachten van de tweede evaluatieronde in Jaar 2: Marja (de Driemaster), Madeleine (de Baanbreker), Hans ('t Plankier), Tim (Martin Luther King), Marga en Sara (Michiel de Ruijter), veel dank voor jullie feedback en openheid. En natuurlijk ook veel dank aan de leerkrachten en scholen die in de laatste experimentele fase mee hebben gedaan: Ans, Ineke en Jan (Rietveldschool), Femkelien (Kompas), Nelleke en Claudia (Wakersduin), Marianne en René (Fonkeling), Willy (Stroomdal), Els (Op 'e Hichte), Martine (Rehoboth), Mireille, Jose en Gina (Komeet), André, Karin en Marieke (Herman Gorter), Ans (St Nicolaas), Dianne (Prinses Marijke), Monique (Schinveld), Teresia (met de Bijbel), Wilma

(Lichtstraat), Berna (Titus Brandsma), Josine (de Klinker), Gert-Jan (Johannes Calvijn), Els (de Notenkraker), Jochem (Montessorischool Kralingen), Linda (Beatrixschool) en ook de vele leerkrachten die de vragenlijst hebben ingevuld.

Dank ook aan collega's Michiel Doorman en Mieke Abels voor de inspirerende ideeën bij de eerste ontwerpen van de technieken en het samenwerken rondom de NWD. Arthur Bakker, dank voor de vele inspirerende gesprekken, filosofische beschouwingen, nadenken over methoden van onderzoek en het bijzonder prettige samenwerken aan de Research Methods-cursus. Andere collega's van het Freudenthal Instituut voor fijne gesprekken in de wandelgangen, op of rond conferenties: Dolly van Eerde, Dédé de Haan, Paul Drijvers, Jan van Maanen, Jantien Smit, Sietske Tacoma, Bas Holleman, Ank van der Heijden, Mariozee Wintermans, Christian Bokhove, Wim van Velthoven. Nathalie Kuijpers, hartelijk dank voor de hulp met layoutzaken!

Kamergenoten voor kortere of langere duur: Adri Dierdorp dank voor het carpoolen, je attenties en vele kopjes koffies. Nathalie Martel pour les quelques conversations en Français à l'institut et l'instauration de nos déjeuners (diners?) réguliers. Ariyadi Wijaya, terimah kasih banyak, it was a pleasure being your roommate for three years. Shinya Ito, thank you for your kindness every single time you visited, I hope you will be able to continue your work on Freudenthal's legacy in Utrecht. Xiaoyan Zhao, thanks for your inspiring enthusiasm, your tenacious work ethic, and your incessant questioning and wanting to have things made completely clear. Ilona Friso-van den Bos, dank voor je positieve instelling en adviezen deze laatste maanden en Ali Sangari, thank you for showing yet another cultural perspective on life, education and assessment.

Dank ook aan collega's van ICO. In het bijzonder Jorrick Beckers, dank voor de vele droogkomische reflecties, altijd een fijne afleiding; en Marrit Jansma, dank voor de prettige mailconversaties die het promovendusschap en amateurvoetbal in leuk Fries perspectief plaatsten. Also many thanks to international mathematics education colleagues for their inspirational talks at PME and CERME.

Naast de voornoemde 'professionele' contacten had ik dit onderzoek ook niet kunnen vervolmaken zonder de steun en afleiding van de volgende vrienden en familieleden.

Darinka, vriendin en collega op afstand. Dank voor je inspirerende onderzoek aan Harvard, bezoeken over en weer, hier en daar. Telkens weer een bijzonder welkome afleiding en motivatie als we elkaar spreken. En dat dankzij SSL ☺.

Anna, merci pour ta gentillesse et conversations, irrégulières mais toujours bienvenues. Je suis très content, et pas peu fier, que l'on ait réussi à faire perdurer notre amitié malgré la distance (Amsterdam, Toulouse, Londres, et bientôt Vancouver...). Nos discussions sur tout, sur la recherche, être doctorant, la valeur des données (f)MRI ; c'est toujours un grand plaisir.

Stéphane Vautier, merci pour tes encouragements, ton amitié et l'opportunité de réfléchir (et publier) avec toi sur des questions de mesure et des lois dans la psychologie. Ce qui m'a permis, malgré nos prises de tête avec le paradigme régnant, de vivre en tant qu'étudiant tout ce qui est associé au processus de la publication. Norman Verhelst, dank voor je gastvrijheid in Tiel en je expertise op psychometrisch gebied, al heb ik daar in dit promotietraject niet direct gebruik van gemaakt, je bent een inspiratie geweest voor me.

De tijgers van RapiDos, zonder de vriendschap, aanheid, lekker gaan en de sportieve uitlaatklep die jullie me bieden was dit me nooit gelukt. Al blijft voetballend succes tegenwoordig iets vaker uit, ik ben heel blij dat dit zootje ongeregeld toch al ruim 10 jaar bij elkaar heeft weten te blijven.

Mijn beide ouders, jullie zijn ieder op eigen wijze altijd bepalend voor keuzes en paden die ik ingeslagen ben, daarin ook altijd bijzonder geïnteresseerd en ondersteunend geweest, veel dank daarvoor. Wouter en Tim, ik ben blij dat jullie naast mijn grote broers ook mijn paranimfen zijn. Suzanne, Mexico is echt heel ver weg.

Pour finir en beauté ; Clémentine en Théo, les amours de ma vie. Clem, mon cœur, merci. Merci m'avoir suivi quand en quelques jours, j'avais postulé, eu un entretien, été embauché et qu'il fallait partir de Toulouse, laissant ta famille, tes amis, ton pays et le travail. Merci pour tout ça et plus encore. Merci de m'aimer comme tu le fais. Merci pour ta compréhension quand encore une fois je devais travailler le soir car j'avais attendu le dernier moment pour finir. Merci de t'intégrer si bien à Amsterdam. Merci de nous avoir donné notre Théo. Merci pour ta force dans les moments difficiles que l'on a traversés avant son venu. Merci d'avoir accepté de devenir ma femme et de porter un nom qui ne se prononce pas en français. Je t'aime.

## Curriculum vitae

Michiel Veldhuis was born on February 17, 1986 in Amsterdam (the Netherlands). He obtained his secondary school degree in 2004 at the St. Ignatiusgymnasium in Amsterdam. In 2005, he enrolled in a bachelor program in Psychology at the University of Amsterdam, graduating with an honours bachelor's degree in Clinical Developmental Psychology in 2008. That same year, he moved to Toulouse (France) and started pursuing a bachelor's degree in Applied Mathematics and Informatics at the University of Toulouse-II-Le Mirail. In 2009, he simultaneously enrolled in the master program of Psychology (*maîtrise*) at the same university. In July 2011, he graduated with both a bachelor's degree in Applied Mathematics and Informatics (*Licence*) and a research master's degree in Psychology of Cognitive Processes (*Master 2 de Recherche*). In September 2011, Michiel started his PhD research in mathematics education at the Freudenthal Institute in connection to the "Education and Learning" Research Program of the Faculty of Social and Behavioural Sciences of Utrecht University. The PhD research was carried out under the supervision of Prof. dr. Marja van den Heuvel-Panhuizen, and was performed within the ICA project (Improving Classroom Assessment, funded by the Netherlands Organization for Scientific Research, NWO MaGW/PROO: Project 411-10-750), which was a joint project with Twente University/Cito. Over the course of the ICA project Michiel published several articles about the ICA project and the connected ICA-China project, and he presented research findings at national and international conferences. Recently, Michiel was invited by the European Society for Research in Mathematics Education to be a Member of the CERME 10 International Program Committee. He also fulfilled all requirements of the Interuniversity Center for Educational Sciences (ICO) Research School in the Netherlands as ICO PhD member.

## Overview of publications related to this thesis

- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). Primary school teachers' assessment profiles in mathematics education. *PLoS ONE*, 9(1), e86817.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). *Exploring the feasibility and effectiveness of assessment techniques to improve student learning in primary mathematics education*. In C. Nicol, S. Oesterle, P. Liljedahl & D. Allan (Eds.) Proceedings of the Joint Meeting of the International Group of Psychology of Mathematics Education (PME) 38 and Psychology of Mathematics Education – North America (PME-NA) 36, Vol 5., pp. 329-336.
- Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen, J.A., Eggen, T.J.H.M. (2013). Teachers' use of classroom assessment in primary school mathematics education in the Netherlands. *CADMO*, 21(2), 35-53.

Articles currently (April 2015) under review:

- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2015). *Exploring teachers' use of classroom assessment techniques in primary mathematics education*. Manuscript submitted for publication.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2015). *Supporting primary school teachers to improve their assessment practice in mathematics: effects on students' learning*. Manuscript submitted for publication.
- Zhao, X., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (2015). *Chinese teachers' use of classroom assessment techniques in primary mathematics education*. Manuscript submitted for publication.

## Overview of presentations related to this thesis

- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2015). *Improving classroom assessment in primary mathematics education*. Paper presented at the 9th Congress of the European society for Research in Mathematics Education, Prague, Czech Republic, February 4-8.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). *Making mathematics teachers better assessors: improving student learning in primary education*. Paper presented at the 15th Association for Educational Assessment-Europe Conference, Tallinn, Estonia, November 6-8.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). *Exploring the feasibility and effectiveness of assessment techniques to improve student learning in primary mathematics education*. Paper presented at the 38th Psychology of Mathematics Education Conference, Vancouver, Canada, July 15-20.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). *Exploring the feasibility and effectiveness of classroom assessment techniques in primary mathematics education*. Poster presented at the 6<sup>th</sup> Expert Meeting on Mathematical Thinking and Learning, Leiden, Netherlands, April 4<sup>th</sup>.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). *Het verbeteren van de toetspraktijk bij rekenen-wiskunde in groep 5* [Improvement of classroom assessment in Grade 3 mathematics]. Paper presented at NRO-PROO Conference 'Mathematics and language in primary education', Utrecht, Netherlands, February 7<sup>th</sup>.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2014). *Het gebruik van formatieve toetstechnieken bij rekenen-wiskunde in groep 5* [The use of formative assessment techniques in Grade 3 mathematics]. Paper presented at the Panama Conference, Noordwijkerhout, Netherlands, January 16-17.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2013). *Exploring the feasibility and effectiveness of assessment techniques to improve student learning in primary mathematics education*. Paper presented at the ICO National Fall School, Maastricht, Netherlands, November 7-8.
- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2013). *Improving classroom assessment: bij rekenen-wiskunde in groep 5 van het basisonderwijs* [Improving classroom assessment: In third grade mathematics education]. Presentation at the PROO-Symposium 'Mathematics in primary school' at the Onderwijs Research Dagen 2013, Brussels, Belgium, May 29-31.

- Veldhuis, M., & Van den Heuvel-Panhuizen, M. (2013). *Teachers' mathematics assessment profiles in primary education*. Poster presented at the 5<sup>th</sup> Expert Meeting on Mathematical Thinking and Learning, Walferdange, Luxembourg, March 1<sup>st</sup>.
- Veldhuis, M. (2012). *Teacher profiles in mathematics education*. Round-table presentation at ICO International Fall School, Girona, Spain, November 5-9.
- Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen, J., & Eggen, T. (2012). *Teachers' use of classroom assessment in primary school mathematics education in the Netherlands*. Paper presented at EARLI SIG 1 Conference, Brussels, August 28-31.
- Veldhuis, M. (2012). *Ontwikkeling van assessmenttechnieken bij rekenen-wiskunde in groep 5* [Development of assessment techniques for mathematics in 3rd grade]. Invited talk at Cito-POVO group, Arnhem, Netherlands, July 2<sup>nd</sup>.

## Flisme Scientific Library

(formerly published as CD- $\beta$  Scientific Library)

89. Jupri, A. (2015). *The use of applets to improve Indonesian student performance in algebra.*
88. Wijaya, A. (2015). *Context-based mathematics tasks in Indonesia: Toward better practice and achievement.*
87. Klerk, S. (2015). *Galen reconsidered. Studying drug properties and the foundations of medicine in the Dutch Republic ca. 1550-1700.*
86. Krüger, J. (2014). *Actoren en factoren achter het wiskundecurriculum sinds 1600.*
85. Lijnse, P. L. (2014). *Omzien in verwondering. Een persoonlijke terugblik op 40 jaar werken in de natuurkundedidactiek.* Utrecht University, Utrecht.
84. Weelie, D. van (2014). *Recontextualiseren van het concept biodiversiteit.* Utrecht University, Utrecht.
83. Bakker, M. (2014). *Using mini-games for learning multiplication and division: a longitudinal effect study.*
82. Ngô Vũ Thu Hằng (2014). *Design of a social constructivism-based curriculum for primary science education in Confucian heritage culture.*
81. Sun, Lei (2014). *From rhetoric to practice: enhancing environmental literacy of pupils in China.*
80. Mazereeuw, M. (2013). *The functionality of biological knowledge in the workplace. Integrating school and workplace learning about reproduction.*
79. Dierdorp, A. (2013). *Learning correlation and regression within authentic contexts.*
78. Dolfing, R. (2013). *Teachers' Professional Development in Context-based Chemistry Education. Strategies to Support Teachers in Developing Domain-specific Expertise.*
77. Mil, M. H. W. van (2013). *Learning and teaching the molecular basis of life.*
76. Antwi, V. (2013). *Interactive teaching of mechanics in a Ghanaian university context.*
75. Smit, J. (2013). *Scaffolding language in multilingual mathematics classrooms.*
74. Stolk, M. J. (2013). *Empowering chemistry teachers for context-based education. Towards a framework for design and evaluation of a teacher professional development programme in curriculum innovations.*
73. Agung, S. (2013). *Facilitating professional development of Madrasah chemistry teachers. Analysis of its establishment in the decentralized educational system of Indonesia.*
72. Wierdsma, M. (2012). *Recontextualising cellular respiration.*
71. Peltenburg, M. (2012). *Mathematical potential of special education students.*

70. Moolenbroek, A. van (2012). *Be aware of behaviour. Learning and teaching behavioural biology in secondary education.*
69. Prins, G. T., Vos, M. A. J. & Pilot, A. (2011). *Leerlingpercepties van onderzoek & ontwerpen in het technasium.*
68. Bokhove, Chr. (2011). *Use of ICT for acquiring, practicing and assessing algebraic expertise.*
67. Boerwinkel, D. J. & Waarlo, A. J. (2011). *Genomics education for decision-making. Proceedings of the second invitational workshop on genomics education, 2-3 December 2010.*
66. Kolovou, A. (2011). *Mathematical problem solving in primary school.*
65. Meijer, M. R. (2011). *Macro-meso-micro thinking with structure-property relations for chemistry. An explorative design-based study.*
64. Kortland, J. & Klaassen, C. J. W. M. (2010). *Designing theory-based teaching-learning sequences for science. Proceedings of the symposium in honour of Piet Lijnse at the time of his retirement as professor of Physics Didactics at Utrecht University.*
63. Prins, G. T. (2010). *Teaching and learning of modelling in chemistry education. Authentic practices as contexts for learning.*
62. Boerwinkel, D. J. & Waarlo, A. J. (2010). *Rethinking science curricula in the genomics era. Proceedings of an invitational workshop.*
61. Ormel, B. J. B. (2010). *Het natuurwetenschappelijk modelleren van dynamische systemen. Naar een didactiek voor het voortgezet onderwijs.*
60. Hammann, M., Waarlo, A. J., & Boersma, K. Th. (Eds.) (2010). *The nature of research in biological education: Old and new perspectives on theoretical and methodological issues – A selection of papers presented at the VIIth Conference of European Researchers in Didactics of Biology.*
59. Van Nes, F. (2009). *Young children's spatial structuring ability and emerging number sense.*
58. Engelbarts, M. (2009). *Op weg naar een didactiek voor natuurkunde-experimenten op afstand. Ontwerp en evaluatie van een via internet uitvoerbaar experiment voor leerlingen uit het voortgezet onderwijs.*
57. Buijs, K. (2008). *Leren vermenigvuldigen met meercijferige getallen.*
56. Westra, R. H. V. (2008). *Learning and teaching ecosystem behaviour in secondary education: Systems thinking and modelling in authentic practices.*
55. Hovinga, D. (2007). *Ont-dekken en toe-dekken: Leren over de veelvormige relatie van mensen met natuur in NME-leertrajecten duurzame ontwikkeling.*
54. Westra, A. S. (2006). *A new approach to teaching and learning mechanics.*
53. Van Berkel, B. (2005). *The structure of school chemistry: A quest for conditions for escape.*
52. Westbroek, H. B. (2005). *Characteristics of meaningful chemistry education: The case of water quality.*

51. Doorman, L. M. (2005). *Modelling motion: from trace graphs to instantaneous change*.
50. Bakker, A. (2004). *Design research in statistics education: on symbolizing and computer tools*.
49. Verhoeff, R. P. (2003). *Towards systems thinking in cell biology education*.
48. Drijvers, P. (2003). *Learning algebra in a computer algebra environment. Design research on the understanding of the concept of parameter*.
47. Van den Boer, C. (2003). *Een zoektocht naar verklaringen voor achterblijvende prestaties van allochtone leerlingen in het wiskundeonderwijs*.
46. Boerwinkel, D.J. (2003). *Het vormfunctieperspectief als leerdoel van natuuronderwijs. Leren kijken door de ontwerpersbril*.
45. Keijzer, R. (2003). *Teaching formal mathematics in primary education. Fraction learning as mathematising process*.
44. Smits, Th. J. M. (2003). *Werken aan kwaliteitsverbetering van leerlingonderzoek: Een studie naar de ontwikkeling en het resultaat van een scholing voor docenten*.
43. Knippels, M. C. P. J. (2002). *Coping with the abstract and complex nature of genetics in biology education – The yo-yo learning and teaching strategy*.
42. Dressler, M. (2002). *Education in Israel on collaborative management of shared water resources*.
41. Van Amerom, B.A. (2002). *Reinvention of early algebra: Developmental research on the transition from arithmetic to algebra*.
40. Van Groenestijn, M. (2002). *A gateway to numeracy. A study of numeracy in adult basic education*.
39. Menne, J. J. M. (2001). *Met sprongen vooruit: een productief oefenprogramma voor zwakke rekenaars in het getallengebied tot 100 – een onderwijsexperiment*.
38. De Jong, O., Savelsbergh, E.R. & Alblas, A. (2001). *Teaching for scientific literacy: context, competency, and curriculum*.
37. Kortland, J. (2001). *A problem-posing approach to teaching decision making about the waste issue*.
36. Lijmbach, S., Broens, M., & Hovinga, D. (2000). *Duurzaamheid als leergebied; conceptuele analyse en educatieve uitwerking*.
35. Margadant-van Arcken, M. & Van den Berg, C. (2000). *Natuur in pluralistisch perspectief – Theoretisch kader en voorbeeldlesmateriaal voor het omgaan met een veelheid aan natuurbeelden*.
34. Janssen, F. J. J. M. (1999). *Ontwerpend leren in het biologieonderwijs. Uitgewerkt en beproefd voor immunologie in het voortgezet onderwijs*.
33. De Moor, E. W. A. (1999). *Van vormleer naar realistische meetkunde – Een historisch-didactisch onderzoek van het meetkundeonderwijs aan kinderen van vier tot veertien jaar in Nederland gedurende de negentiende en twintigste eeuw*.

32. Van den Heuvel-Panhuizen, M. & Vermeer, H. J. (1999). *Verschillen tussen meisjes en jongens bij het vak rekenen-wiskunde op de basisschool – Eindrapport MOOJ-onderzoek.*
31. Beeftink, C. (2000). *Met het oog op integratie – Een studie over integratie van leerstof uit de natuurwetenschappelijke vakken in de tweede fase van het voortgezet onderwijs.*
30. Vollebregt, M. J. (1998). *A problem posing approach to teaching an initial particle model.*
29. Klein, A. S. (1998). *Flexibilization of mental arithmeticsstrategies on a different knowledge base – The empty number line in a realistic versus gradual program design.*
28. Genseberger, R. (1997). *Interessegeoriënteerd natuur- en scheikundeonderwijs – Een studie naar onderwijsontwikkeling op de Open Schoolgemeenschap Bijlmer.*
27. Kaper, W. H. (1997). *Thermodynamica leren onderwijzen.*
26. Gravemeijer, K. (1997). *The role of context and models in the development of mathematical strategies and procedures.*
25. Acampo, J. J. C. (1997). *Teaching electrochemical cells – A study on teachers' conceptions and teaching problems in secondary education.*
24. Reygel, P. C. F. (1997). *Het thema 'reproductie' in het schoolvak biologie.*
23. Roebertsen, H. (1996). *Integratie en toepassing van biologische kennis – Ontwikkeling en onderzoek van een curriculum rond het thema 'Lichaamsprocessen en Vergift'.*
22. Lijnse, P. L. & Wubbels, T. (1996). *Over natuurkundedidactiek, curriculumontwikkeling en lerarenopleiding.*
21. Buddingh', J. (1997). *Regulatie en homeostase als onderwijsthema: een biologie-didactisch onderzoek.*
20. Van Hoeve-Brouwer G. M. (1996). *Teaching structures in chemistry – An educational structure for chemical bonding.*
19. Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education.*
18. Klaassen, C. W. J. M. (1995). *A problem-posing approach to teaching the topic of radioactivity.*
17. De Jong, O., Van Roon, P. H. & De Vos, W. (1995). *Perspectives on research in chemical education.*
16. Van Keulen, H. (1995). *Making sense – Simulation-of-research in organic chemistry education.*
15. Doorman, L. M., Drijvers, P. & Kindt, M. (1994). *De grafische rekenmachine in het wiskundeonderwijs.*
14. Gravemeijer, K. (1994). *Realistic mathematics education.*
13. Lijnse, P. L. (Ed.) (1993). *European research in science education.*
12. Zuidema, J. & Van der Gaag, L. (1993). *De volgende opgave van de computer.*

11. Gravemeijer, K, Van den Heuvel Panhuizen, M., Van Donselaar, G., Ruesink, N., Streefland, L., Vermeulen, W., Te Woerd, E., & Van der Ploeg, D. (1993). *Methoden in het reken-wiskundeonderwijs, een rijke context voor vergelijkend onderzoek.*
10. Van der Valk, A. E. (1992). *Ontwikkeling in Energieonderwijs.*
9. Streefland, L. (Ed.) (1991). *Realistic mathematics education in primary schools.*
8. Van Galen, F., Dolk, M., Feijs, E., & Jonker, V. (1991). *Interactieve video in de nascholing reken-wiskunde.*
7. Elzenga, H. E. (1991). *Kwaliteit van kwantiteit.*
6. Lijnse, P. L., Licht, P., De Vos, W. & Waarlo, A. J. (Eds.) (1990). *Relating macroscopic phenomena to microscopic particles: a central problem in secondary science education.*
5. Van Driel, J. H. (1990). *Betrokken bij evenwicht.*
4. Vogelesang, M. J. (1990). *Een onverdeelbare eenheid.*
3. Wierstra, R. F. A. (1990). *Natuurkunde-onderwijs tussen leefwereld en vakstructuur.*
2. Eijkelhof, H. M. C. (1990). *Radiation and risk in physics education.*
1. Lijnse, P. L. & De Vos, W. (Eds.) (1990). *Didactiek in perspectief.*

## ICO Dissertation Series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the following universities: Eindhoven University of Technology, Erasmus University Rotterdam, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Antwerp, University of Ghent, University of Groningen, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University (and formerly Radboud University Nijmegen and Tilburg University).

Recent publications in the ICO Dissertation Series (updated April 2015):

303. Strien, J.L.H. van (19-12-2014) *Who to Trust and What to Believe? Effects of Prior Attitudes and Epistemic Beliefs on Processing and Justification of Conflicting Information From Multiple Sources*. Heerlen: Open University of the Netherlands.
302. Huizinga, T. (12-12-2014) *Developing curriculum design expertise through teacher design teams*. Enschede: University of Twente.
301. Leenaars, F.A.J. (10-12-2014) *Drawing gears and chains of reasoning*. Enschede: University of Twente.
300. Gabelica, C. (4-12-2014) *Moving Teams Forward. Effects of feedback and team reflexivity on team performance*. Maastricht: Maastricht University.
299. Wijnia, L. (14-11-2014) *Motivation and Achievement in Problem-Based Learning: The Role of Interest, Tutors, and Self-Directed Study*. Rotterdam: Erasmus University Rotterdam.
298. Gaikhorst, L. (29-10-2014) *Supporting beginning teachers in urban environments*. Amsterdam: University of Amsterdam.
297. Khaled, A.E. (7-10-2014) *Innovations in Hands-on Simulations for Competence Development. Authenticity and ownership of learning and their effects on student learning in secondary and higher vocational education*. Wageningen: Wageningen University.
296. Engelen, J. (11-09-2014) *Comprehending Texts and Pictures: Interactions Between Linguistic and Visual Processes in Children and Adults*. Rotterdam: Erasmus University Rotterdam.
295. Rijt, J.W.H. van der, (11-9-2014) *Instilling a thirst for learning. Understanding the role of proactive feedback and help seeking in stimulating workplace learning*. Maastricht: Maastricht University.
294. Rutten, N.P.G. (5-9-2014) *Teaching with simulations*. Enschede: University of Twente.
293. Hu, Y. (26-6-2014) *The role of research in university teaching: A comparison of Chinese and Dutch teachers*. Leiden: Leiden university.
292. Baars, M.A. (6-6-2014) *Instructional Strategies for Improving Self-Monitoring of Learning to Solve Problems*. Rotterdam: Erasmus University Rotterdam

291. Coninx, N.S. (28-05-2014) *Measuring effectiveness of synchronous coaching using bug-in-ear device of pre-service teachers*. Eindhoven: Eindhoven University of Technology.
290. Loon, M. van (8-5-2014) *Fostering Monitoring and Regulation of Learning*. Maastricht: Maastricht University.
289. Bakker, M. (16-04-2014) *Using mini-games for learning multiplication and division: A longitudinal effect study*. Utrecht: Utrecht University.
288. Mascareno, M.N. (11-4-2014) *Learning Opportunities in Kindergarten Classrooms. Teacher-child interactions and child developmental outcomes*. Groningen: University of Groningen.
287. Frambach, J.M. (26-3-2014) *The Cultural Complexity of problem-based learning across the world*. Maastricht: Maastricht University.
286. Karimi, S. (14-3-2014) *Analysing and Promoting Entrepreneurship in Iranian Higher Education: Entrepreneurial Attitudes, Intentions and Opportunity Identification*. Wageningen: Wageningen University.
285. Kuijk, M.F. van (13-03-2014). *Raising the bar for reading comprehension. The effects of a teacher professional development program targeting goals, data use, and instruction*. Groningen: University of Groningen.
284. Hagemans, M.G. (07-03-2014) *On regulation in inquiry learning*. Enschede: University of Twente.
283. Smet, M.J.R. de (31-1-2014). *Composing the unwritten text: Effects of electronic outlining on students' argumentative writing performance*. Heerlen: Open University of the Netherlands.
282. Zwet, J. van der (30-1-2014). *Identity, interaction, and power. Explaining the affordances of doctor-student interaction during clerkships*. Maastricht: Maastricht University.
281. Cviko, A. (19-12-2013) *Teacher Roles and Pupil Outcomes. In technology-rich early literacy learning*. Enschede: University of Twente.
280. Kamp, R.J.A. (28-11-2013) *Peer feedback to enhance learning in problem-based tutorial groups*. Maastricht: Maastricht University.
279. Lucero, M.L. (21-11-2013) *Considering teacher cognitions in teacher professional development: Studies involving Ecuadorian primary school teachers*. Ghent: Ghent University.
278. Dolfing, R. (23-10-2013) *Teachers' Professional Development in Context-based Chemistry Education. Strategies to Support Teachers in Developing Domain-specific Expertise*. Utrecht: Utrecht University.
277. Popov, V. (8-10-2013) *Scripting Intercultural Computer-Supported Collaborative Learning in Higher Education*. Wageningen: Wageningen University.
276. Bronkhorst, L.H. (4-10-2013) *Research-based teacher education: Interactions between research and teaching*. Utrecht: Utrecht University.
275. Bezdán, E. (4-10-2013) *Graphical Overviews in Hypertext Learning Environments: When One Size Does Not Fit All*. Heerlen: Open University of the Netherlands.