Educational Studies in Mathematics (2019) 102:435–456
https://doi.org/10.1007/s10649-019-09905-7

Assessing mathematical thinking as part of curriculum reform in the Netherlands

Paul Drijvers [1] · Hanneke Kodde-Buitenhuis [1] · Michiel Doorman [1]

Published online: 4 July 2019

Abstract
Assessment is a crucial factor in the implementation of curriculum reform. Little is known, however, on how curriculum changes can be reflected adequately in assessment, particularly if the reform concerns process skills. This issue was investigated for the case of assessing mathematical thinking in a mathematics curriculum reform for 15–18-year-old students in the Netherlands. From 2011 until 2017, these reform curricula were field tested in pilot schools, while other schools used the regular curricula. The research question is how this reform was reflected in national examination papers and in student performance on corresponding assignments. To address this question, we developed a theory-based model for mathematical thinking, analyzed pilot and regular examination papers, and carried out a quantitative and qualitative analysis of students' work on assignments that invite mathematical thinking. The results were that the pilot examination papers did address mathematical thinking to a greater degree than the regular papers, but that there was a decrease over time. Pilot school students outperformed their peers in regular schools on assignments that invite mathematical thinking by 4–5% on average and showed more diversity in problem-solving strategies. To explain the decreasing presence of mathematical thinking in examination papers, we conjecture that conservative forces within the assessment construction process may push back change.

Keywords Assessment · Curriculum reform · Mathematical thinking

✉ Paul Drijvers
p.drijvers@uu.nl

Hanneke Kodde-Buitenhuis
hannekebuitenhuis@gmail.com

Michiel Doorman
m.doorman@uu.nl

[1] Freudenthal Institute, Utrecht University, PO Box 85.170, 3508 AD Utrecht, The Netherlands

1 Introduction

It is well documented that the implementation of curriculum reform and curricular innovation is far from straightforward (Stanic & Kilpatrick, 1992). As a recent case, the implementation of the Common Core State Standards for mathematics raised a strong debate in the USA (Larson & Kanold, 2016). Scaling up from small-scale pilot sites to nationwide educational practice is a complex enterprise, in which different processes ask for coherence, integration, interaction, and alignment: the production of curriculum documentation, the design of teaching resources, the dissemination of core ideas, the delivery of professional development activities for teachers, and monitoring the implementation through supporting research (Fullan, 2007; Kuiper, 2009; Wahlström & Sundberg, 2015). As Goodlad (1985) pointed out, curriculum development processes involve the interplay of three distinct perspectives: the substantive perspective, the technical-professional perspective, and the social-political perspective.

A key factor in curriculum reform is the establishment of new assessment practices. Developing assessment practices that meet new curricular goals often turns out to be the closing piece in curriculum reform implementation processes. In the meantime, to a great extent, such assessment practices do determine the success of the new curricula's implementation. Assessment practices, and high-stakes centralized summative tests such as national final examinations in particular, to a large extent drive text book design and teaching practice; as such, they are crucial in implementing educational change. The case of mathematics curriculum reform addressed in this paper is no exception (Fried & Amit, 2016). As the Mathematical Sciences Education Board [MSEB] and National Research Council [NRC] (1993) expressed it, "if current assessment practices prevail, reform in school mathematics is not likely to succeed" (p. 31).

However, fine-tuning curriculum reform and assessment practices is not only crucial but also problematic. Much remains unknown about factors that affect a successful alignment. This holds even more if the reform involves competences that are difficult to assess, compared to more straightforward goals. For mathematics curricula in particular, recent research shows that a mismatch between curriculum and assessment with respect to competences such as mathematical thinking or modeling is likely to occur in the Netherlands and elsewhere (Drüke-Noe & Kühn, 2017; Vos, 2013). The theme of this article, therefore, is to investigate the alignment of curriculum reform and assessment, and the factors that may inhibit it, for the case of mathematics.

To address this theme, we study the case of a recent curriculum reform in the Netherlands for 15–18-year-old pre-higher education students. One central aim of this curriculum reform was to foster students' mathematical thinking, which is considered a central higher order learning goal of mathematics education worldwide. In the light of the importance attributed to these skills, and their central role in the Dutch curriculum reform, the study addresses the following research question: how is the curriculum reform with respect to mathematical thinking reflected in national examination papers in the Netherlands and in student performance on corresponding assignments?

2 Theoretical framework

The framework that guided this retrospective analysis included overarching notions about assessment in the context of curriculum reform and a domain-specific lens on mathematical thinking skills, so central in the current Dutch curriculum reform.

2.1 Assessment of reform curricula

Four notions about assessment as part of curriculum reform implementation are relevant for this study. First, the classical criterion for appropriate assessment is validity. Informally speaking, validity means that the test measures what it is meant to measure. Validation is the process of gathering evidence that provides a basis for the sound interpretation of test scores (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). In the case of curriculum reform, this type of evidence is hard to acquire, as the "what is meant to be measured" is subject to changes. This particularly holds if the curriculum reform includes underlying skills and competences, such as mathematical thinking or modeling competences, that are not straightforward to assess (Vos, 2013), and if the assessment format does not change along with the curriculum, as was the case in the Dutch reform (Kuiper, Folmer, & Ottevanger, 2013). The validity perspective in this study is helpful to study the link between intended curriculum and assessment and to consider the validity of the assessment format in the light of mathematical thinking as learning goal.

The second notion about assessment is that it should reflect the teaching and learning as they took place in the reform education. From this perspective, changes in assessment should reflect developments in teaching and learning. However, educational practice does not change easily, and assessment may be used as a lever to implement reform. This may imply a "Catch-22" situation[1] in curriculum reform implementation: changes in assessment are limited because the educational practice that they should reflect is changing slowly, whereas educational practice will not change as long as assessment is not really different. An important factor here is the backwash effect of assessment: it is well documented in research that summative and centralized assessment, such as national final examinations, very much guides teachers and text book authors (e.g., see Watkins, Dahlin, & Ekholm, 2005). This perspective may be useful in understanding the phenomenon under study, i.e., the implementation of mathematical thinking in national assessments.

A third notion of interest concerns taxonomies to classify learning goals, which may be used to structure assessment matrices and student models. Many of these taxonomies, such as Bloom's (e.g., see Krathwohl, 2002), are hierarchical in the sense that they distinguish lower and higher order learning goals. The notion of mathematical thinking competence would be at the high end of such a taxonomy. Within this high end, the threefold model by Drijvers (2015), presented in more detail below, will serve as a non-hierarchical taxonomy and can be used as a framework to analyze national final tests.

Fourth and final, Kuiper's model of curriculum reform implementation stimuli may be useful to position assessment in the curriculum reform implementation as a whole. Kuiper (2009) described three types of stimuli that affect the implementation process of curriculum reform, namely top-down, bottom-up, and from-aside stimuli. Speaking in general (see Fig. 1 left), bottom-up elements included the needs expressed by teachers and the initiatives taken to deal with these needs. From-aside elements included professional development activities, educational resources, and practice-oriented research studies. In the specific Dutch assessment reform case (Fig. 1 right), the top-down elements included official curriculum descriptions and ministry regulations, provided by the assessment authority College voor Toetsen en Examens [CvTE]. From-aside stimuli included the work by assessment experts from the Cito assessment

[1] Named after the novel by Heller (1961).

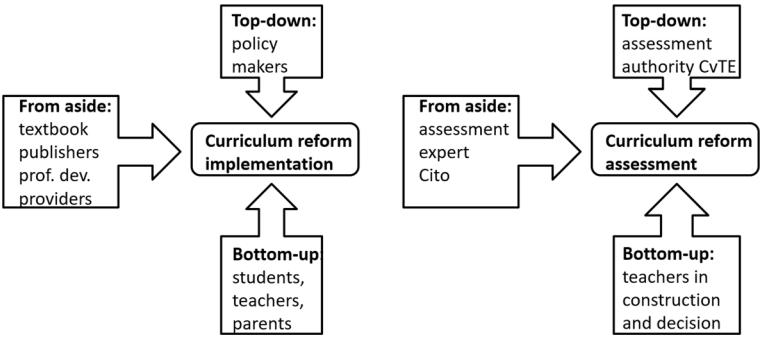

Fig. 1 Curriculum reform implementation stimuli (left, inspired by Kuiper, 2009) adapted to national assessment in the Netherlands (right)

institute. As we will see, the content of the national examination papers is to a large extent determined by teachers, who provide input from the bottom-up. As such, assessment is influenced by different stimuli, and the national final examinations can be considered as boundary objects (Akkerman & Bakker, 2011) in the Dutch reform process. The model in Fig. 1 (right) may help to explain assessment practices as they developed over time in the Netherlands.

2.2 Mathematical thinking

The importance of mathematical thinking as a key higher order learning goal in mathematics education is widely accepted. As Pólya (1963, p. 605) already wrote: "First and foremost, it [mathematics education, PD] should teach those young people to THINK." In line with this, Skemp (1976) convincingly argued that one should teach for relational rather than instrumental understanding. The National Research Council (1989, p. 31) highlighted the power of mathematical thinking and described some key elements: "[…] mathematics offers distinctive modes of thought which are both versatile and powerful, including modeling, abstraction, optimization, logical analysis, inference from data, and use of symbols." More recently, the Common Core State Standards included learning goals that clearly refer to mathematical thinking, such as "make sense of problems and persevere in solving them," "reason abstractly and quantitatively," "model with mathematics," and "use appropriate tools strategically."[2] Katz (2014) explained that mathematical thinking is an important ingredient of inquiry-based mathematics education. In addition to this shared opinion on the importance of mathematical thinking, it is also widely acknowledged that it is not straightforward for teachers to move from "answer telling and procedure teaching" to raising thought-provoking questions and teaching mathematical thinking (Mason, 2000). The danger of not making this shift, however, is that mathematics in school differs drastically from mathematics in a professional or academic setting (Devlin, 2012).

In spite of the consensus on the importance of mathematical thinking as a learning goal and on challenges this may raise for teachers, views on what mathematical thinking really is are less unanimous. A first main approach to mathematical thinking stresses problem-solving (e.g., see Mason, 2000; Pólya, 1962; Schoenfeld, 1992, 2013, 2014). Problem-solving refers to Pólya's description of "finding a way out of a difficulty, a way around an obstacle, attaining an

[2] http://www.corestandards.org/Math/Practice/

aim which was not immediately attainable" (Pólya, 1962, p. v). Problems, in this view, are tasks that are non-routine to the student and invite to think of a possible solution strategy (Doorman et al., 2007). What is non-routine depends on the student's experience, preliminary knowledge, talent, and skills. Developing a repertoire of heuristics is an important element in problem-solving skills (van Streun, 2001). Looking back on this body of literature, we agree that problem-solving is a key component of mathematical thinking. In our view, however, it is not exclusively the domain of mathematics, and, more important, mathematical thinking encompasses more than problem-solving.

A second angle to consider mathematical thinking concerns modeling. Modeling involves connecting mathematics and the world around us, applying mathematics, and inventing mathematics to solve problems. As such, it is an indispensable element in sense making mathematics education. Mathematical means are used, for example, to understand phenomena, to predict developments, or to optimize processes (e.g., see Blum, Galbraith, Henn, & Niss, 2007; Kaiser, Blomhøj, & Sriraman, 2006). In countries such as the Netherlands (de Lange, 1987), and more recently, Austria (Siller et al., 2015), a modeling approach has guided curriculum design to an important extent. Without wanting to underestimate the importance of modeling and the type of mathematical thinking and sense making it involves, we do believe that while modeling is a very ambitious learning goal, it too does not completely cover the broad sense of mathematical thinking.

A third angle, finally, to consider mathematical thinking is the abstraction perspective. Abstraction is "an activity by which we become aware of similarities […] among our experiences" (Skemp, 1986, p. 21), or, to phrase it differently, "the isolation of specific attributes of a concept so that they can be considered separately from the other attributes" (Tall, 1988, p. 2). As a result of abstraction, one enters the world of mathematical objects and their relationships (Mason, 1989). White and Mitchelmore (2010) developed an interesting model for teaching abstraction in phases. In line with this, we see generalization and abstraction as key processes in mathematical thinking that may have been underestimated in recent work on mathematical thinking. Because of this, we would like to include it in our approach.

To summarize the above, we conclude that problem-solving, modeling, and abstracting are indeed important elements of mathematical thinking. In line with extensive discussions within the Dutch reform curriculum committee (cTWO, 2007), we feel that an approach to mathematical thinking that does justice to all three elements is needed. As a result, we structure the debate by setting up a model for mathematical thinking that integrates problem-solving, modeling, and abstracting as core aspects (Drijvers, 2015), and in which each of the three elements follows the above definitions. This wide lens integrates several views on mathematical thinking, at the price of not focusing on one of its aspects in more detail. More information on the way in which these three notions are operationalized can be found in Section 3 and in Appendices 1 and 2.

This threefold mathematical thinking model consisting of the triad problem-solving–modeling–abstraction not only reflects the view of the Dutch mathematics curriculum reform committee but it also connects to the theory of Realistic Mathematics Education (RME). RME is an instruction theory that has greatly influenced Dutch mathematics education and that considers mathematics as a human activity that should be experienced as meaningful by the students (van den Heuvel-Panhuizen & Drijvers, 2014). A problem-solving approach can foster this. Furthermore, the notion of mathematization plays a central role in RME. Within mathematization, horizontal and vertical mathematization are distinguished. Horizontal mathematization, mathematizing reality, includes using mathematical tools to organize and solve problems positioned in real-life situations (Treffers, 1987). As such, it clearly connects to the

modeling aspect mentioned above. Vertical mathematization, mathematizing mathematics, refers to the process of reorganization within the mathematical system resulting in shortcuts by using connections between concepts and strategies. Through vertical mathematization, one enters the abstract world of symbols; as such, it reflects the abstraction aspect of the mathematical thinking model.

3 Context of the study

In the Netherlands, pre-higher secondary education consists of a 5- and a 6-year program. In these programs, students can choose to do Mathematics A, Mathematics B, or Mathematics C, with Mathematics D as an additional option. Mathematics B is the mathematics curriculum that best prepares for higher education in science, technology, engineering, and mathematics, and it is central in the debate on the mathematics curriculum reform. Globally speaking, the approach and "spirit" of the Dutch Mathematics B curriculum can be compared to A-level mathematics courses in the UK. About 25% of the students in the 5-year program enroll in Mathematics B and about 50% of the students in the 6-year program.[3]

Figure 2 depicts the timeline for the mathematics curriculum reform process for 15–18-year-old pre-higher education students in the Netherlands in which this study was situated. The reform involved both general senior higher education (5-year program, grades 7–11) and the pre-university program (6-year program, grades 7–12).[4] The process started in 2005, after a similar reform had been initiated for chemistry, biology, and physics 2 years before. The curriculum reform committee Commissie Toekomst Wiskundonderwijs [cTWO], established by the Ministry of Education, consisted of 14 members, including teachers, mathematicians, teacher educators, and researchers. Seven curricula were to be designed: three for the different streams in the 5-year program and four for the different streams in the 6-year program. Based on the intermediate vision statement report (cTWO, 2007), pilot teaching materials were designed by teams of authors and field tested in 16 pilot schools. Schools volunteered to participate. Most pilot schools tested two or three of the seven curricula. Five of the 16 schools opted for the so-called Mathematics B curriculum for the 6-year program and six for the 5-year program. In 2013, the cTWO final report was accepted by the ministry (cTWO, 2013). It included curriculum prescriptions with attainment goals (the "what") but did not prescribe teaching approaches or detailed time schedules (the "how"). In 2014, a professional development program was set up. The new mathematics curricula were implemented nationwide in 2015 in grade 10 (CvTE, 2016). The first nationwide national examinations took place in 2017 in grade 11 (5-year curriculum) and in 2018 in grade 12 (6-year curriculum).

The national final examinations in the Netherlands are designed with great care. Based on the syllabi, a construction group, consisting of a small number of experienced mathematics teachers, designs assignments under the guidance of an assessment expert from Cito, a Dutch institute for testing and assessment. In several cycles over a period of 2–3 years, these assignments and concepts of examination papers go back and forth between the construction group, the Cito assessment expert, and a second committee, the assessment committee. The latter committee, which is appointed by the national assessment authority (CvTE), consists of

[3] https://www.duo.nl/open_onderwijsdata/databestanden/vo/leerlingen/

[4] For an overview of the Dutch educational system, see www.epnuffic.nl/en/study-and-work-in-holland/dutch-education-system.

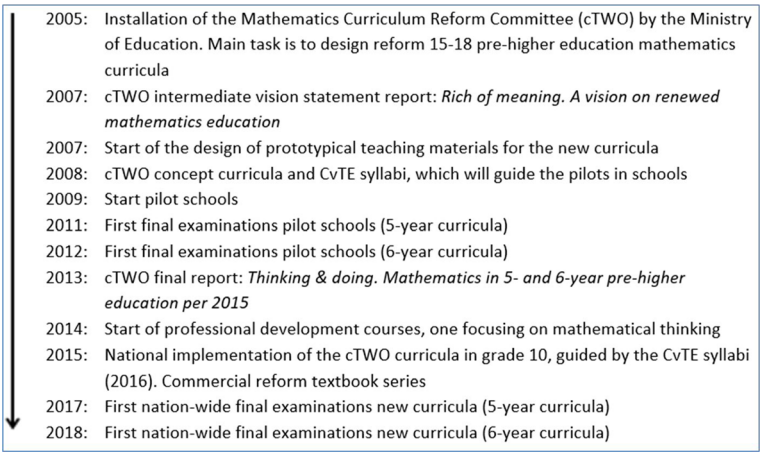

Fig. 2 Dutch curriculum reform timeline

two mathematics teachers and one chair (usually from higher education) and decides on the examination papers. Assignments are field tested and examination papers are proofread by a limited number of external experts. This construction process is coordinated by the assessment expert from Cito. After delivery, the examination results are evaluated, not only through psychometric and statistical analyses but also through sessions with teachers and monitoring teacher reactions on discussion boards, such as the online forum moderated by the Dutch Association of Mathematics Teachers. Students' results on the final national examination only account for 50% of their final grade for mathematics. The other 50% comes from school assessment. Even if schools are quite free to set their own assessment formats for this school-internal assessment, the overall impression is that most school examinations prepare for the national examination and have a similar style and format. The backwash effect seems to have a large impact on school examinations (Kuiper, van Silfhout, & Trimbos, 2017).

For students in the pilot schools where the new curricula were tested, small-scale national pilot examinations were set up from 2011 until 2017, alongside the nationwide regular examinations. These pilot examinations included (1) completely new assignments, fit to the new curricula, (2) assignments that are identical to regular examination assignments, and (3) assignments that are adapted from regular examination assignments, to better reflect the curriculum reform. Appendix 2 contains an example of the latter category.

As for the political context, the curriculum reform addressed in this paper took place in a climate of a Dutch version of a "Math wars" debate between mathematicians and mathematics educators, similar to the one described by Schoenfeld (2004). Both within and outside the reform committee, mathematicians were suspicious with respect to the Realistic Mathematics Education approach. As a consequence, the RME view did not guide the reform process; this view was not very much represented in the Reform Committee, and to an even smaller extent in the Assessment Committee. To not further fuel this—sometimes unproductive—debate, it seemed appropriate to highlight mathematical thinking independently from realistic mathematics education.

The first author of this paper was involved in the curriculum reform process, from 2005 as a member of the reform committee cTWO and from 2011 as the chair of the national assessment committee for Mathematics B, the mathematics curriculum that prepares for higher education and university studies in science, technology, engineering, and mathematics. The second author carried out most of the analyses.

4 Methods

In this study, we analyzed the regular and pilot national final examinations during the curriculum reform process from the first pilot examinations in 2011 until the first nationwide final examinations in 2017. The analyses in this study focused on the Mathematics B curricula and included the two steps of an analysis of national examination papers for regular and pilot curricula and of a quantitative and qualitative analysis of examination results from regular and pilot curriculum students.

4.1 Analysis of national examination papers for regular and pilot curricula

The goal of the analysis of national examination assignments of regular and pilot curricula was to assess how the reform focus on mathematical thinking was reflected in the national pilot and regular examinations from the years 2011 to 2017. Data consisted of the 25 examination papers over these years, each delivered in a written 3-h examination session.

To analyze these examination papers, an instrument was designed in which the model by Drijvers (2015) was slightly adapted: within the modeling element in mathematical thinking, a distinction was made between the process of modeling and the model as an object. The modeling process refers to translating between the world of the problem situation and mathematics (cf. horizontal mathematization), whereas the model as an object includes adapting a model, analyzing its properties, or comparing different models. Similarly, within the abstraction element, a distinction was made between the process of abstracting and the use of abstract mathematical objects (Kodde-Buitenhuis, 2015). In the process of abstracting, similarities and differences in mathematical situations lead to the formation of meaningful mathematical objects, whereas the abstract object category refers to thinking about mathematical objects, their properties, and the relations between them. Figure 3 depicts this model, which led to the codebook presented in Table 1.

To assign codes to the assignments, each assignment was coded dichotomously (yes/no) with respect to each of the five categories shown in Table 1. Appendices 1 and 2 show some

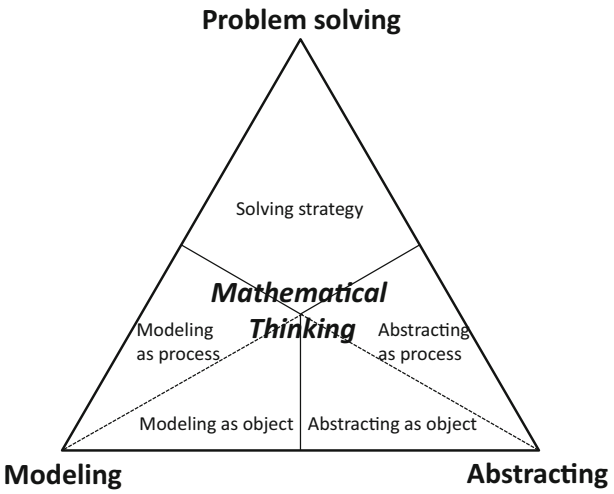

Fig. 3 The model used for the examination paper analysis (Kodde-Buitenhuis, 2015)

Table 1 Codebook for examination paper analysis (summarized code descriptions)

Code	Code description
Problem-solving	Find a strategy to solve a non-routine problem, which may include multiple ways and may consist of multiple thinking steps.
Modeling	
Modeling as process	Translate between a problem situation and mathematical terms or vice versa.
Model as object	Customize a model, analyze model features, or compare different models.
Abstracting	
Abstracting as process	Extract similarities and differences from concrete situations to form meaningful mathematical objects or apply such knowledge in a new concrete situation.
Abstract object	Think about mathematical objects, their properties, and relations between them.

exemplary assignments to give insight in the coding process. If none of the codes applied, the assignment was considered as not inviting mathematical thinking; if at least one of the codes applied, it was considered an "MT assignment" and the number of credit points was added to the category. In 7 out of 85 cases, an assignment was assigned two codes. Appendix 3 shows an example of such coding.

To assess the interrater reliability, three different coders coded a subset of 18 examination items using the codebook. This led to an interrater reliability of Fleiss' kappa = 0.70, which can be considered as a good agreement (Landis & Koch, 1977). Based on the coding, the overall proportions of assignments that invite mathematical thinking were assessed, as well as proportions for each of the model's categories. In this way, we could compare regular and pilot examination papers and were able to identify the trend over time.

4.2 Quantitative and qualitative analysis of examination results from regular and pilot curriculum students

The goal of the quantitative analysis of examination results from regular and pilot curriculum students was to assess how students from five pilot schools performed on assignments that appeal to mathematical thinking, compared to their peers in the regular curricula. To assess the students' performance with respect to mathematical thinking in the pilot examinations, we identified 15 assignments (seven for the 5-year stream and eight for the 6-year stream) that invite mathematical thinking and that were identical in the regular and pilot examinations. One was deleted from the analysis, as the assessment authority decided not to include it in the grading due to some critical comments by teachers, who considered the assignment as not matching the curriculum description.

Data consist of the pilot students' results, as they had been assigned in the regular grading process, and the results from students in the regular curriculum for these 14 identical assignments that involved mathematical thinking. Table 2 provides the numbers of students involved. To analyze the data, we compared the average proportion of the maximum score that was obtained for the assignment, the so-called p values—not to be confused with significances in statistical hypothesis testing. For each of the overlapping assignments, the hypotheses that the pilot and regular students performed equally well were tested with two-sample t tests, and effect sizes were calculated. The t test was considered appropriate given the independence of the two groups, the sample sizes, and the assumption of normality.

Table 2 Numbers of students involved in the analysis of Mathematics B examination results

Year	5-year regular	5-year pilot	6-year regular	6-year pilot
2011	10,387	138	–	–
2012	11,311	135	15,683	243
2013	10,789	121	14,935	178
2014	11,327	124	14,226	166
2015	10,593	109	14,925	95
2016	13,232	150	15,169	107
2017	13,178	–	17,059	123
Total	80,817	777	91,997	912

To back up the quantitative analysis with more detailed information, the qualitative analysis of pilot curriculum students' examination papers aimed to provide insight into the work done by the student and into the underlying signs of (a lack of) mathematical thinking. The analysis consisted of a manual qualitative analysis on a sample of students' written work on the Mathematics B 2014 examination (6-year program), including 128 students from three regular schools and 110 students from four pilot schools. In terms of overall student achievements, these schools can be regarded as average schools.

The data analysis focused on the assignments that invited mathematical thinking in terms of the model used for the examination paper analysis described above. For this limited set of assignments, an open coding procedure was carried out, leading to three main categories of differences of student work: the use of representations, of different problem-solving strategies, and of algebraic versus numerical methods. Appendix 3 shows an example.

5 Results

5.1 The national examination papers for regular and pilot curricula

The results of the analysis of national examination papers in the years 2011–2017, for the 5- and 6-year stream and for the regular curricula and the pilot reform curricula, are shown in Table 3. The percentages reflect the relative frequency of assignments that invite mathematical thinking in terms of problem-solving, modeling, and abstraction. In the final row, the average percentages are calculated. A special case here is the 2017 examination of the 5-year stream:

Table 3 Percentage of credit points of assignments inviting mathematical thinking in regular and pilot national examinations for the 5- and 6-year pre-higher education streams

Year	5-year regular	5-year pilot	6-year regular	6-year pilot
2011	16%	42%	–	–
2012	5%	26%	41%	60%
2013	5%	24%	33%	53%
2014	11%	32%	22%	50%
2015	21%	38%	35%	38%
2016	10%	22%	35%	38%
2017	22%		29%	27%
Average	11%	31%	33%	44%

Table 4 Frequencies of each of the mathematical thinking model's components in regular and pilot examinations (2011–2017) for the 5- and 6-year pre-higher education streams

Model component	5-year regular	5-year pilot	6-year regular	6-year pilot	Total
Problem-solving	12	28	30	37	107
Modeling as process	3	2	0	6	11
Model as object	0	0	0	0	0
Abstracting as process	0	1	0	0	1
Abstract object	2	2	4	9	17

this was the first nationwide test according to the new curriculum. This percentage was not included in the average in the final row.

Table 3 reveals several findings. First, it shows that mathematical thinking assignments are less frequent in the 5-year pre-higher education stream than in the 6-year pre-university stream. Second, it shows that the frequency of thought-provoking assignments was higher in the pilot examinations than in the regular ones (31% versus 11%, and 44% versus 33%, respectively). Third, we notice a decreasing trend in the pilot 6-year stream, with the final year's pilot examination paper being similar to the regular paper with respect to mathematical thinking, whereas the difference in the first examinations in 2012 was considerable.

Table 4 shows the frequencies of each of the categories of the mathematical thinking model for all examination papers. It reveals that problem-solving is highly dominant, whereas assignments involving modeling and abstraction are scarce. This is in agreement with the analysis by van Streun (2014). Models as objects, for example for comparison or customization, did not appear in the analysis, and the process of abstraction is almost absent as well.

To illustrate the way in which examination assignments were changed in the curriculum reform, Appendix 2 shows an assignment from the 2012 regular examination (6-year program) and its adapted version in the pilot examination.

5.2 Examination results from regular and pilot curriculum students

The comparison of student performances on assignments that were part of both the regular and pilot examinations and invite mathematical thinking is shown in Tables 5 and 6. After the columns with the year and the assignment number in the pilot and regular examinations, the p values as well as their differences are provided. We recall that the p value is the average percentage of the maximum credit that the students could obtain. Next, the t values are reported, as well as the degrees of freedom and the Cohen d effect sizes.

In Table 5, we see that the 5-year stream pilot students outperformed their peers in the regular curriculum by 5.4% points on average. For three out of the seven assignments, the differences are statistically significant and the effect sizes can be considered small. For the 6-year stream results shown in Table 6, there was a 4.1%-point average improvement in favor of the pilot students. There are three significant improvements, with small to medium effect sizes, but there is also one small negative effect size in 2017. The latter assignment focused on integral calculus, which was a more prominent topic in the regular than in the new curriculum. These results suggest a modest improvement with respect to the students' mathematical thinking, which can be attributed to the new curriculum.

Table 5 Overview of *p* values for identical assignments (5-year program)

Year	Assign. # pilot	Assign. # regular	*p* value pilot	*p* value regular	*p* value diff	*t* value	*df*	Cohen's *d*
2011	3	3	40.0	41.9	− 1.8	− 0.57	10,523	− 0.05
2011	17	18	70.0	59.9	10.1	2.51*	10,523	0.22
2011	18	19	61.4	47.1	14.3	3.99***	10,523	0.34
2012	14	7	87.2	79.4	7.8	2.73**	11,444	0.24
2014	11	13	37.7	35.0	2.7	0.72	11,449	0.07
2015	3	3	79.1	76.0	3.0	0.83	10,700	0.08
2016	8	9	43.3	41.3	2.0	0.72	13,380	0.06
Average			59.8	54.4	5.4			

*Significance $p < 0.05$ (two-tailed)

**Significance $p < 0.01$ (two-tailed)

***Significance $p < 0.001$ (two-tailed)

As a further look at the student results, the qualitative analysis of a sample of students' written work on the 6-year stream Mathematics B 2014 examination suggests three types of differences within the assignments that call for mathematical thinking:

– Students from pilot schools showed more diversity in problem-solving strategies than their peers in the regular program on identical assignments. For example, on one pilot assignment, 7 out of 110 pilot students (6%) invented an alternative solving strategy, whereas this was 3 out of 128 (2%) for the regular students, even if both percentages are quite low. This diversity also appeared in other assignments in the pilot examinations that invite mathematical thinking.
– Sketching graphs and switching between representations were shown more frequently in pilot schools. For example, on one pilot assignment, 19 out of 110 pilot students (17%) invented an alternative solving strategy, whereas this was 3 out of 128 (2%) for the regular students. Appendix 3 shows another example.

Table 6 Overview of *p* values for identical assignments (6-year program)

Year	Assign. # pilot	Assign. # regular	*p* value pilot	*p* value regular	*p* value diff	*t* value	*df*	Cohen's *d*
2013	2	2	60.3	56.7	3.7	1.18	15,111	0.09
2013	16	14	50.5	50.4	0.1	0.04	15,111	0.00
2014	15	6	65.6	70.4	− 4.8	− 1.94	14,390	− 0.15
2015	4	3	57.2	56.2	1.0	0.33	15,018	0.03
2015	13	14	61.1	34.6	26.4	6.23***	15,018	0.64
2016	2	3	52.0	39.3	12.7	3.79***	15,274	0.37
2017	14	13	45.7	56.3	− 10.6	− 3.48***	17,180	− 0.31
Average			56.0	52.0	4.1			

*Significance $p < 0.05$ (two-tailed)

**Significance $p < 0.01$ (two-tailed)

***Significance $p < 0.001$ (two-tailed)

– Finally, students in pilot schools used exact algebraic methods more often, whereas students in the regular program made more use of the graphing calculator. For example, on one pilot assignment, 43 out of 110 students solved the equation $(3-3r)^2+(1-r)^2=(1+r)^2$ algebraically, where approximations provided by the graphing calculator would have been sufficient.

6 Conclusion and discussion

In this paper, we investigated the alignment of curriculum reform and assessment for the case of mathematical thinking, one of the key points in the recent curriculum reform in the Netherlands. The research question was the following: How is the curriculum reform with respect to mathematical thinking reflected in national examinations papers in the Netherlands and in student performance on corresponding assignments? To address this question, we analyzed national examination papers for regular and pilot curricula, we analyzed quantitatively the examination results from regular and pilot curriculum students on common assignments, and we backed up the findings with a qualitative analysis of pilot curriculum students' examination work.

With respect to the examination papers, we conclude that the pilot examination papers did pay more attention to mathematical thinking than the regular ones. Problem-solving was the most prominent aspect by far. However, mathematical thinking became less visible in the 5-year program's examination papers than in the 6-year program's examinations. The most striking result is that we notice a decreasing number of assignments appealing to mathematical thinking over time, i.e., during the period from 2011 to 2017. This is a concern, which is discussed below.

With respect to student performance, we conclude that the students in pilot schools performed better on mathematical thinking tasks that were identical in both examination papers than the students enrolled in the regular curricula. Taking into account the size of the increase, on average 5.4% of the total score for the 5-year program and 4.1% for the 6-year program, we consider this as just a modest improvement. Also, there was much variation among the different assignments with respect to the performance differences and most effect sizes were small. The qualitative analysis of student work suggested that pilot students showed more diversity in problem-solving strategies than their peers in the regular program. Sketching graphs and switching between representations was shown more frequently in pilot schools and exact algebraic methods were used more often.

In short, we conclude that the curriculum reform has indeed led to more assignments in national examination papers that invite mathematical thinking and that the students in pilot schools show a modest improvement on such tasks compared to the students enrolled in the regular curriculum. However, there is a tendency to reduce attention for mathematical thinking in the examination papers, and we consider the improvement in student achievement a modest one.

Before discussing these conclusions, let us first address the study's limitations. Related to the examination results, we should acknowledge that both the number of assignments and the number of students enrolled in the pilot curricula were small. Also, the pilot schools had not been selected randomly, but had volunteered. Even if there is evidence that these schools did not perform significantly better overall than regular schools, there could be some teacher effects involved, such as teachers in pilot schools being more qualified or more positive about the curriculum reform. Therefore, we cannot be completely sure about the causes of the achievement effects that we

found. And, even more seriously, we do not know what caused the (mostly positive) effects, as they may result from the interplay between specific content knowledge, probably better represented in the new curricula, and mathematical thinking in general. In particular, the new 6-year program curricula are more coherent in their focus on algebraic and analytical skills, as Euclidean geometry was replaced by analytical geometry. As the effects we found hold both for the 5-year and the 6-year program, we do not consider this as a cause. Also, we do not have any evidence of different assessment practices in regular and pilot schools outside the two variants of the national examination papers.

Let us now discuss the study and reflect on its theoretical and empirical outcomes in the light of the literature described in the theoretical framework. Even if the research question has an empirical and descriptive character, the study's main theoretical value lies in the model of mathematical thinking as a triangle of problem-solving, modeling, and abstracting. This triangle, visualized in Fig. 3 and elaborated in the Codebook in Table 1, integrates the different views on mathematical thinking as they are described in literature (e.g., Blum et al., 2007; Mason, 2000; Pólya, 1962, 1963; Schoenfeld, 1992; White & Mitchelmore, 2010). Whereas problem-solving, modeling, and abstracting are usually considered in isolation, we now integrate these views on mathematical thinking in one single model, which proved to be applicable in the context of the study. As such, the model is a step ahead in doing justice to the multi-faceted character of mathematical thinking as it emerges from literature. The model may be helpful to assess the balance between the three elements of mathematical thinking, to ensure that each component is addressed, and to monitor developments over time. As a further refinement of the model, we acknowledge that in retrospect, there was no need to refine the coding of the modeling category and the abstraction category into a process and an object sub-code. The model might benefit from simplifying it into its basic threefold structure.

A second point to discuss at a theoretical level is that Goodlad's notions (1985) and Kuiper's model (2009) on curriculum reform and assessment offer useful perspectives to study curriculum reform processes. Kuiper's model (see Fig. 1) was helpful in identifying factors that may explain the difficulties in implementing curriculum reform and in maintaining the new elements in the reform curricula. The continuous interplay of the "top-down," "from-aside," and "bottom-up" players depicted in the Kuiper (2009) model (Fig. 1, right hand side) might make the educational system resistant to change.

At the empirical level, the mathematical thinking model showed to be helpful in investigating the balance of the three mathematical thinking elements in the Dutch national examinations. In this case, it was striking that problem-solving was apparently considered the most important or feasible aspect for inclusion. Abstraction and modeling were not prominent in the examination papers under consideration, neither in the regular ones nor in the pilot examinations, even though they had been recognized as important aspects of mathematical thinking (Drijvers, 2015). These findings are in line with the results reported by Vos (2013), who claims that modeling in Dutch examination papers usually appears only in mechanistic and reproductive ways. At an international level, these findings reflect the work by Drüke-Noe and Kühn (2017), who show that there is a general tendency in central written examinations to mainly include assignments with low cognitive demands. This brings up the question whether written national examinations are the best way to assess these types of mathematical thinking skills. In the Dutch examination system, with its dual character of school assessment and national assessment, we wonder if the school assessment, with its more flexible format, could play a more important role in assessing mathematical thinking.

Another empirical finding, generated by the use of the model, was the decrease over time of examination assignments that invite mathematical thinking. This is a point of concern, as the presence of mathematical thinking in pilot examination papers was at the heart of the reform intentions. What is causing this downward trend? As mentioned above, the national examinations may be less suitable to assess mathematical thinking. Also, seen from a social-political perspective (Goodlad, 1985), there may be intrinsically conservative mechanisms within the assessment construction process that are resistant to change. One factor might be the Dutch "Math wars" debate on Realistic Mathematics Education, which caused the examination construction to be carefully monitored by critical forces from outside. Indeed, the 2012 6-year pilot examination was criticized by teachers, who found it too difficult for their students. We conjecture that this led to the examination board being more careful in pushing mathematical thinking too hard in national examinations. Another factor might be the relationship between the actors in the assessment process and the reform committee; we conjecture that the fact that the reform committee member left the assessment authority in 2013 was not in favor of a close link between the two.

In terms of the Kuiper (2009) model, the continuous interplay of the "top-down," "from-aside," and "bottom-up" players might also explain the system's resistance to change and its trend to reduce change over time. The iterative revisions of potential examination assignments by different stakeholders may reinforce the tendency of being (too?) careful and, as a consequence, of avoiding or simplifying assignments that originally invited mathematical thinking. Such a general resistance to change in assessment practices may hold particularly for mathematical thinking, with its global and somewhat vague character, compared to specific curriculum changes, such as adding a specific topic, e.g., complex numbers, to the calculus strand. A reform that aims at competence change is harder, as it will challenge assessment validity and reliability. As a "top-down" stimulus to ensure this reform to happen, one option would be for the assessment authorities to require the national examination papers to contain a certain percentage of assignments that invite mathematical thinking, to ensure its visibility, and to mobilize the lever function and the backwash effect of national examinations. For teachers, we suggest that they exploit the freedom they have in their school-based assessment to assess mathematical thinking skills in a way they see as appropriate.

A final point for discussion concerns the fact that less attention is paid to mathematical thinking in the 5-year strand examination papers than in the 6-year papers, even if the reform committee cTWO claimed that mathematical thinking is relevant for all students. Is mathematical thinking considered too hard for the 5-year stream, or less relevant? We would agree that the level of mathematical thinking required should be adapted to the target group, but not that it would be impossible or less important for students in the 5-year program. If we consider mathematical thinking to be relative to the student's preliminary knowledge, skills, and talent, there is in principle no reason for this difference. Also, in the light of the attention paid to mathematical literacy (Organisation for Economic Co-operation and Development [OECD], 2017), we believe it is worthwhile to investigate further how mathematical thinking skills could be addressed in assessment that did not address the highest level students.

Of course, these conclusions and this discussion lead to new questions to explore. A first question is how mathematical thinking could be assessed in assessment formats that differ from the regular written and centralized examinations. The Dutch model of dual assessment would provide an excellent environment for such a study. Second, an important question is how mathematical thinking goals can be addressed in assessment for those students who are not high-achievers in mathematics, but who might all the same benefit from mathematical thinking skills in their future professional and private lives.

Acknowledgments The authors thank Theo van den Bogaart, Marieke Bor-de Vries, Wilmad Kuiper, and Maarten Pieters for their valuable comments on earlier versions of this paper.

Appendix 1. Applying the mathematical thinking model to assignments

This appendix illustrates the way in which examination assignments were coded according to the mathematical thinking model consisting of problem-solving, modeling, and abstracting and to the code book shown in Table 1 in particular. We subsequently provide an example of each of the three elements.

Problem-solving

Assignment 2 of the 6-year program pilot examination paper in 2013 concerns the so-called Antoine's equation on the relation between damp pressure and temperature. The following equation is provided:

$$\log P = k - \frac{m}{T - n} \quad \text{with } T > m$$

Here, P is the damp pressure in bar and T the temperature in kelvin, and k, m, and n are constants that depend on the type of liquid. Also, a graph is provided (see Fig. 4).

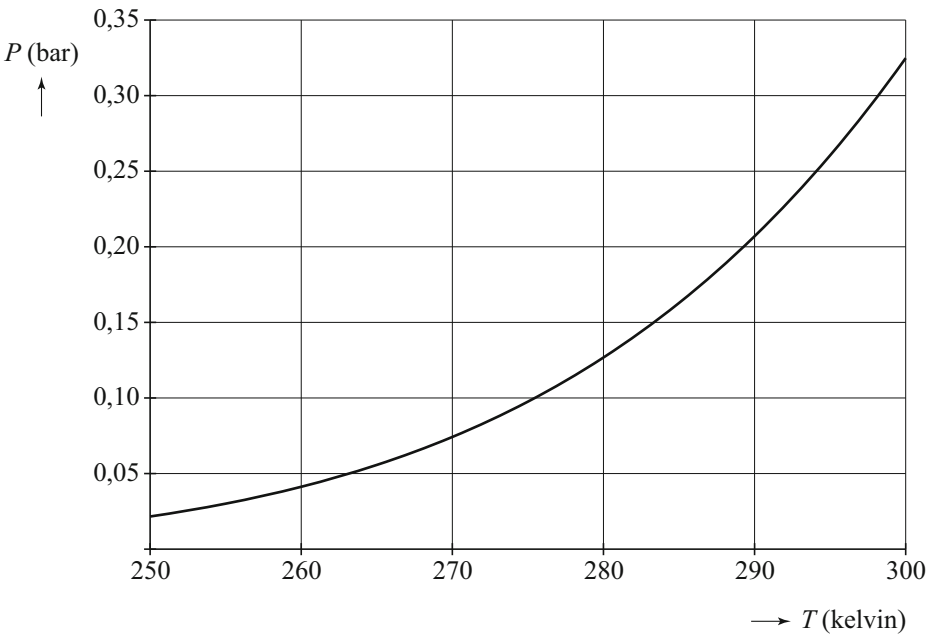

Fig. 4 The pressure–temperature graph (2013 6-year program pilot examination)

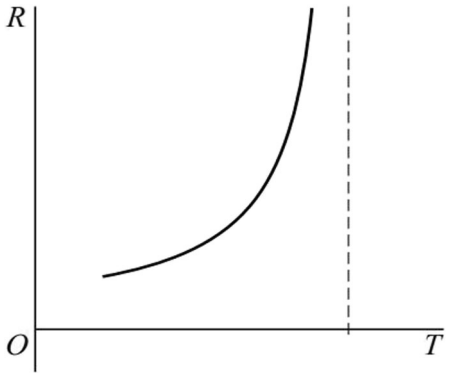

Fig. 5 The survival time–temperature graph (2011 5-year program pilot examination)

The task, now, is to show through reasoning with the formula, but without differentiation, that the function is indeed increasing. As this is a non-routine task to these students, it is coded as a problem-solving assignment.

Modeling

Assignment 3 of the 5-year program pilot examination paper in 2011 concerns the survival time of somebody who falls into cold water. The following formula is provided:

$$R = 15 + \frac{7.2}{0.0785 - 0.0034T} \text{ with } R > 0 \text{ and } T > 5.0$$

Here, R stands for the survival time in minutes and T for the water temperature in degree Celsius. Also, a global graph is provided (see Fig. 5).

The task here is to calculate the value of T that corresponds to the graph's vertical asymptote and to explain the meaning of the vertical asymptote for the situation of the person in the water. The credits for the latter aspect, the interpretation of the mathematical phenomenon in terms of the problem situation, are coded as modeling, and modeling as process, in particular.

Abstracting

Assignment 17 of the 5-year program pilot examination paper in 2011 shows a logarithm table (see Fig. 6). An example is provided, showing how to calculate log 1½ as the difference of log 3 and log 2. Next, the task is to calculate log 24 without the use of a calculator. This task is considered an abstraction task and, more precisely, is coded "abstract object," as the logarithm is considered an abstract object here that needs to be manipulated based on mathematical relations and not on the meaning of logarithm as it was taught in terms of context situations, i.e., the logarithm as the time needed to reach a specific growth, given the growth factor per time.

n	$\log n$
1	0
2	0,3010
3	0,4771
4	0,6021
5	0,6990
6	0,7782
7	0,8451
8	0,9031
9	0,9542
10	1
100	2
1000	3

Fig. 6 The logarithm table (2011 5-year program pilot examination)

Appendix 2. Adaptation, coding, and results of a pilot examination assignment

This appendix illustrates the way in which examination assignments changed in the curriculum reform, how the coding took place for the pilot version, and what the results were.

Figure 7 shows part of an assignment from the 2012 examination paper (6-year program), both the regular and the pilot examination. It concerns the shape of a wine glass. One of the assignments for the students, both in the regular and in the pilot examination paper, is to set up a formula for the curve *CD*. The introduction to the assignment, however, is quite different for the two examination papers. The regular paper provides the following introduction:

> To graph *CD*, a downwards opened parabola is used with *C* as its vertex. A formula for this parabola can be found by first moving the curve *CD* so that *C* becomes (0, 0). Figure 8 shows the curve *CD* and its image *OE* and also the translation is shown. Curve

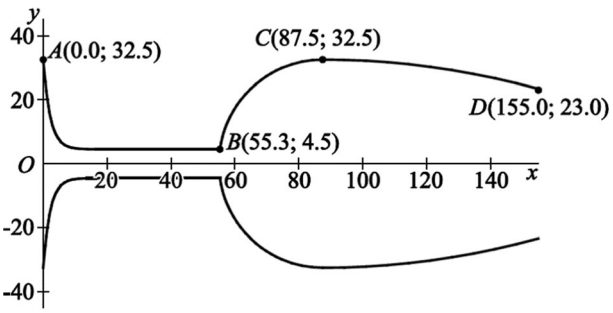

Fig. 7 The graph with the wine glass assignment (2012 6-year program examination)

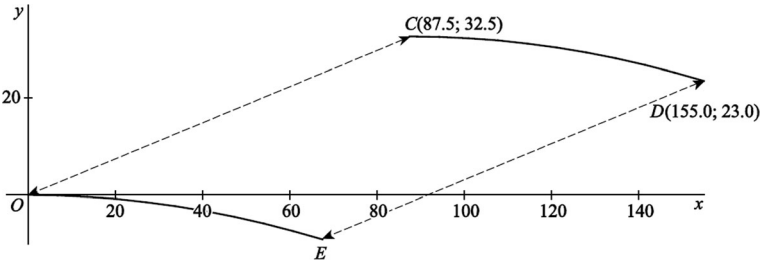

Fig. 8 The graph joining the suggested approach in the regular examination version

OE is part of a downwards opened parabola with vertex *O*, so has a formula of the form $y = a \cdot x^2$, with $a < 0$. Now the translation can be used that maps *OE* to *CD*. This translation is also shown in Fig. 8.

This introduction guides the student to a particular solution strategy (i.e., translating the curve). In the pilot examination paper, however, Fig. 8 is not included, and the introduction is limited to the following.

To graph *CD*, a downwards opened parabola is used with *C* as its vertex.

Clearly, in the pilot version, much more mathematical thinking is required to find a possible solution strategy (either translating the graph or trying to find a parabola directly) and to carry out the assignments that emerge from the chosen strategy, as the students have to figure out the strategy themselves, rather than following the approach suggested in the regular version of the assignment. The marking scheme of both examination papers, provided by the assessment authorities, assigned different full credit for the pilot version and the regular version, namely five instead of four:

- *C*(87.5, 32.5) is the vertex of the parabola, so a formula for *CD* has the form $y = a(x - 87.5)^2 + 32.5$ (2 points)

- *D*(155.0, 23.0) is a point on the curve *CD*, so $23.0 = a(155.0–87.5)2 + 32.5$ (1 point)
- Describe how this equation can be solved (1 point)
- This gives for *a* the value -0.002 (or more precise), so a formula for the curve CD is $y = 0.002(x - 87.5)^2 + 32.5$ (1 point)

To code the pilot version of the assignment according to the code book shown in Table 1, 3 credit points were coded as problem-solving, as the first two bullets in the marking scheme were considered non-routine tasks to the students.

The overall results for this assignment show that pilot students ($N = 243$) on average scored 39.8% of the maximum score and students in regular schools 51.8% ($N = 15{,}683$). The guidance apparently led to higher scores for the regular students.

Appendix 3. An example of the analysis of student work

To illustrate the differences between the written work by pilot students and regular students, Fig. 9 (left hand side) shows an assignment from the pre-university stream

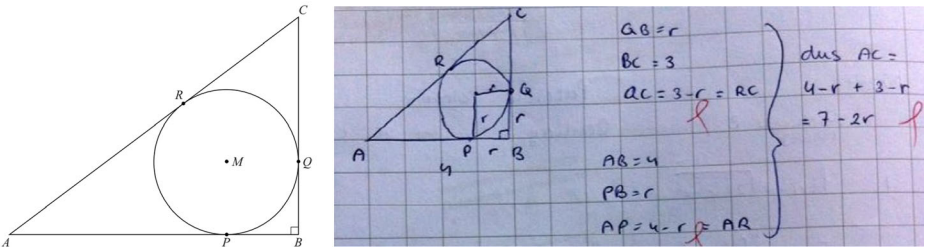

Fig. 9 Part of a pilot assignment (left) and corresponding student work (right)

pilot examination paper in 2014. A rectangular triangle ABC is shown, with $AB = 4$ and $BC = 3$. The assignment is to prove that the radius of the circle is equal to 1. The student work shown on the right hand side contains a sketch with r as a newly introduced variable, which leads to equation models. Introducing a new variable is considered a manifestation of the process of modeling, and drawing a sketch manifests problem-solving skills.

References

Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research, 81*, 132–169.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Blum, W., Galbraith, P. L., Henn, H.-W., & Niss, M. (2007). *Modelling and Applications in Mathematics Education. The 14th ICMI Study*. New ICMI Study Series Vol. 10. New York, NY: Springer.

College voor Toetsen en Examens (2016). Wiskunde B VWO. Syllabus Centraal Examen 2018 (bij het nieuwe examenprogramma). Utrecht: CvTE. https://www.examenblad.nl/examenstof/syllabus-2018-wiskunde-b-vwo/2018/f=/syllabus_wiskunde_b_vwo_2018.pdf.

Commissie Toekomst Wiskundeonderwijs (2007). Rijk aan betekenis. Visie op vernieuwd wiskundeonderwijs. [Rich of meaning. A vision on renewed mathematics education.] Utrecht, the Netherlands: cTWO. http://www.fi.uu.nl/ctwo/.

Commissie Toekomst Wiskundeonderwijs (2013). *Denken & doen, wiskunde op havo en vwo per 2015*. [Thinking & doing. Mathematics in 5- and 6-year pre-higher education per 2015.] Utrecht: cTWO. http://www.fi.uu.nl/ctwo/publicaties/docs/CTWO-Eindrapport.pdf.

de Lange, J. (1987). *Mathematics, insight and meaning*. Doctoral thesis. Utrecht, the Netherlands: OW & OC.

Devlin, K. (2012). *Introduction to mathematical thinking*. Petaluma, CA: Devlin.

Doorman, L. M., Drijvers, P. H. M., Dekker, G. H., van den Heuvel-Panhuizen, M. H. A. M., de Lange, J., & Wijers, M. M. (2007). Problem solving as a challenge for mathematics education in the Netherlands. *ZDM - International Journal on Mathematics Education, 39*(5–6), 405–418.

Drijvers, P. (2015). *Denken over wiskunde, onderwijs en ICT*. [Thinking about mathematics, education and ICT.] Inaugural lecture. Utrecht, the Netherlands: Universiteit Utrecht. http://www.fisme.science.uu.nl/publicaties/literatuur/Oratie_Paul_Drijvers_facsimile_20150521.pdf.

Drüke-Noe, C. & Kühn, S.M. (2017). Cognitive demand of mathematics tasks set in European statewide exit exams—are some competences more demanding than others? In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education*

(CERME10, February 1–5, 2017) (pp. 3484–3491). Dublin, Ireland: DCU Institute of Education and ERME.
Fried, M. N., & Amit, M. (2016). Reform as an issue for mathematics education research: Thinking about change, communication, and cooperation. In L. English & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (3rd ed., pp. 257–274). New York, NY: Routledge.
Fullan, M. (2007). *The new meaning of educational change* (Fourth ed.). New York, NY: Teachers College Press.
Goodlad, J.I. (1985). Curriculum as a field of study. In T. Husén & N.T. Postlethwaite (Eds.), *The International Encyclopedia of Education* (pp. 1141–1144). Oxford, U.K.: Pergamon Press.
Heller, J. (1961). *Catch-22*. New York, NY: Simon & Schuster.
Kaiser, G., Blomhøj, M., & Sriraman, B. (2006). Towards a didactical theory for mathematical modelling. *Zentralblatt für Didactik der Mathematik, 38*(2), 82–85.
Katz, J. D. (2014). *Developing mathematical thinking. A guide to rethinking the mathematics classroom*. Lanham, MD: Rowman & Littlefield.
Kodde-Buitenhuis, J. W. (2015). *Wiskundig denken in de pilot examens van de nieuwe wiskundecurricula havo/vwo.* [Mathematical thinking in pilot examinations of new curricula.] Internal rapport. Arnhem, the Netherlands: Cito.
Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Practice, 41*(4), 212–218.
Kuiper, W. (2009). *Curriculumevaluatie en verantwoorde vernieuwing van bètaonderwijs*. [Curriculum evaluation and responsible reform of science education.] Inaugural lecture. Utrecht, the Netherlands: Universiteit Utrecht.
Kuiper, W., Folmer, E., & Ottevanger, W. (2013). Aligning science curriculum renewal efforts and assessment practices. In D. Corrigan, R. Gunstone, & A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy* (pp. 101–118). New York, NY: Springer.
Kuiper, W., van Silfhout, G., & Trimbos, B. (2017). Curriculum en toetsing. [Curriculum and assessment.] In E. Folmer, A. Koopmans-van Noorel, & W. Kuiper (Eds.), *Curriculumspiegel 2017* [Curriculum mirror 2017], (pp. 83–110). Enschede, the Netherlands: SLO.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
Larson, M.R., & Kanold, T.D. (2016). The Common core mathematics debate. In M.R. Larson, & T.D. Kanold (Eds.), *Balancing the equation* (pp. 43–56). Reston, VA: NCTM.
Mason, J. (1989). Mathematical abstraction ad the result of a delicate shift of attention. *For the Learning of Mathematics, 9*(2), 2–8.
Mason, J. (2000). Asking mathematical questions mathematically. *International Journal of Mathematical Education in Science and Technology, 31*(1), 97–111.
Mathematical Sciences Education Board, & National Research Council (1993). *Measuring what counts: A conceptual guide for mathematics assessment*. Washington D.C.: National Academies Press.
National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.
OECD (2017). PISA 2015 assessment and analytical framework. Science, reading, mathematic, financial literacy and collaborative problem solving. Paris, France: OECD Publishing.
Pólya, G. (1962). *Mathematical discovery. On understanding, learning, and teaching problem solving (Vol. 1)*. New York – London – Sydney: Wiley & Sons.
Pólya, G. (1963). On learning, teaching, and learning teaching. *The American Mathematical Monthly, 70*(6), 605–619.
Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 334–370). New York, NY: Macmillan.
Schoenfeld, A. (2004). The math wars. *Educational Policy, 18*, 253–286.
Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast, 10*(1–2), 9–34.
Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? *Educational Researcher, 43*(8), 404–412.
Siller, H.-S., Bruder, R., Hascher, T., Linnemann, T., Steinfeld, J., & Sattlberger, E. (2015). Competency level modelling for school leaving examination. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education* (CERME9, 4–8 February 2015) (pp. 2716–2723). Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME.
Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 20–26.
Skemp, R. R. (1986). *The psychology of learning mathematics (2nd ed.)*. Harmondsworth, U.K.: Penguin.

Stanic, G. M. A., & Kilpatrick, J. (1992). Mathematics curriculum reform in the United States: A historical perspective. *International Journal of Educational Research, 17*(5), 407–417.

Tall, D. (1988). *The nature of advanced mathematical thinking. Discussion paper for the working group on advanced mathematical thinking*. PME-XII, Vezprém, Hungary. https://homepages.warwick.ac.uk/staff/David.Tall/pdfs/dot1988i-nature-of-amt-pme.pdf.

Treffers, A. (1987). Three dimensions. A model of goal and theory description in mathematics instruction—the Wiskobas project. Dordrecht, the Netherlands: D. Reidel Publishing Company.

van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 521–525). Dordrecht, Heidelberg, New York, London: Springer.

van Streun, A. (2001). *Het denken bevorderen*. [Promoting thinking.] Inaugural lecture. Groningen, the Netherlands: RUG.

van Streun, A. (2014). Onderwijzen en toetsen van wiskundige denkactiviteiten. [Teaching and assessing mathematical thinking activities.] Enschede, the Netherlands: SLO. http://www.slo.nl/downloads/2014/onderwijzen-en-toetsen-van-wiskundige-denkactiviteiten.pdf/.

Vos, P. (2013). Assessment of modelling in mathematics examination papers: Ready-made models and reproductive mathematising. In G.A. Stillman et al. (Eds.), Teaching mathematical modelling: Connecting to research and practice, international perspectives on the teaching and learning of mathematical modelling (pp. 479–488). Dordrecht, the Netherlands: Springer.

Wahlström, N., & Sundberg, D. (2015). *Theory-based evaluation of the curriculum Lgr 11*. Sweden: Institute for Evaluation of Labour Market and Education Policy. http://www.ifau.se/Upload/pdf/se/2015/wp2015-11-Theory-based-evaluation-of-the-curriculum-Lgr_11.pdf.

Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Science, 33*(4), 283–309.

White, P., & Mitchelmore, M. (2010). Teaching for abstraction: A model. *Mathematical Thinking and Learning, 12*(3), 205–226.