Educational Studies in Mathematics (2023) 114:315–336
https://doi.org/10.1007/s10649-023-10233-0

When teachers construct tests for assessing students' competencies: a taxonomy

Semir Becevic[1]

Accepted: 13 April 2023 / Published online: 29 June 2023

Abstract

Little is known about how teachers construct tests. For that reason, this study addresses the use of teacher-constructed tests for assessing educational goals, expressed in terms of student mathematical competencies. The focus is on meanings that upper secondary school mathematics teachers assign to their own test construction practices for assessing educational goals, expressed in terms of mathematical competencies in the curriculum. The methodological approach of grounded theory, underlined by symbolic interactionism, is applied to semi-structured interviews with teachers. The core category, the emerging taxonomy, is derived by revealing distinctions in degree of paying attention to competencies: no attention, superficial attention, and qualitative attention, as well as two different phases of the assessment: constructional and marking. Finally, a couple of possible implications for developing and improving test construction are offered. This includes collaborative work, inside and outside of schools, with both prospective and in-service teachers, for improvement of competence implementation in regular teaching and learning in alignment with mathematical content.

Keywords Teacher-constructed tests · Assessment · Competencies · Upper secondary school · Grounded theory · Symbolic interactionism · Noticing

1 Introduction

For quite some time, various notions of mathematical competence and competencies, as well as the corresponding notions of skills, abilities, capabilities, and proficiencies, have been given increasing importance in the formulation of learning goals and standards in mathematics (Biehler, 2019; Niss & Højgaard, 2019; Niss et al., 2017; Pettersen & Braeken, 2019), and in the mathematics education literature worldwide (Department of Education, 2013; Kilpatrick et al., 2001; Niss, 1993; Niss & Højgaard, 2002; Swedish National Agency for Education, 2000). And they cover results of mental processes connected to doing mathematics by several competency frameworks (Kilpatrick, 2020). These

✉ Semir Becevic
semir.becevic@gu.se

[1] Department of Pedagogical, Curricular and Professional Studies, Faculty of Education, University of Gothenburg, Box 100, S-405 30 Gothenburg, Sweden

notions have different origins and sometimes overlap (Weinert, 1999). What they have in common is that they communicate goals for student learning and are used for educational evaluation and development. Teacher-constructed tests might well serve as a method for this evaluation and development, as well as for coordination of the intended (ideal and formal), the implemented (perceived and operational), and the attained (experiential and learned) curriculum (Osta, 2020).

In this article, *teacher-constructed tests* are defined as written tests that are constructed and/or chosen by teachers for assessment (Goos, 2020; Watt, 2005), and their proper alignment with curricula and classroom instruction is a prerequisite for making students sufficiently aware of their goal attainment (Phelan & Phelan, 2010). As one of the central means of assessment, teacher-constructed tests play an important role in the interoperation of educational and assessment practices; this is the case all over the world (Brookhart, 1994; Niss et al., 2016; Oescher & Kirby, 1990).

Therefore, it is surprising that so little research has been conducted about teachers' test construction. How teachers address curriculum requirements in their assessment tests, what they pay attention to, how they choose topics, and how they construct their assessment tests are still not known or well-enough understood. Teacher-constructed tests are one area where insights may be obtained into meanings that teachers assign in their interpretations of educational goals, topics, or competency requirements. These meanings provide the basis for what teachers notice as relevant for their choice of test tasks.

The aim of this study is to relate meanings that teachers assign to their test construction practices to teachers' noticing of what is relevant when choosing tasks, and assessing educational goals, expressed in terms of mathematical competencies.

The research question is as follows: What do teachers pay attention to when they recall and describe their test practices in relation to the requirements in the national curriculum for assessing students' *procedural*, *conceptual*, *reasoning*, *problem-solving*, *modelling*, and *communication competencies* on three qualitative levels?

The focus is not on teachers' test construction practices per se but on teachers' sense-making processes on previous test construction practices. These sense-making processes are based on what they attend to, interpret, and make decisions about, throughout the test examples that they provide. This will be investigated through interviews with teachers about their test construction practices. Although the question will be answered only for Swedish teachers, the aim is to provide knowledge, which could be generalized to international contexts.

2 Theoretical framing

Symbolic interactionism is a theoretical approach that can guide the reconstruction of teachers' sense-making, both theoretically and methodologically. This approach is used to capture what the teachers consider is relevant to test construction, in alignment with the competency requirements from the curriculum.

Symbolic interactionism is based on three main premises (Blumer, 1969, p. 2):

1. Human beings act toward things on the basis of meanings that the things have for them
2. The meaning of such things is derived from, or arises out of, the social interaction that one has with one's fellows
3. These meanings are handled in, and modified through, an interpretive process used by the person dealing with the things he/she encounters.

Leaning on these premises, I start from the assumption that teachers act towards test construction based on meanings that this test construction, and the associated competency requirements from the curriculum, has for them. Based on these meanings, that is, what the teacher attends to, teachers' test construction is negotiated by communication with the researcher in a social interaction during the interview. This social interaction within the interview between the teacher and the researcher is the main source of creation as well as recreation of (previous) meanings. Meanings are also handled and modified through an interpretive process when recalling test construction within the interviews. When recalling and describing these test constructional demands throughout the interviews, teachers attend to, interpret, and make decisions concerning their test construction. In summary, symbolic interactionism allows for identification and reconstruction of creation and recreation of meanings in the interviews, that is, what teachers consider to be relevant, interpret and make decisions about on their test construction in relation to the competency requirements. Teachers' attending or paying attention to is therefore conceptualized by the notion of noticing (Criswell & Krall, 2017; Goodwin, 1994; Jacobs et al., 2020) that is, attending to, interpreting, and deciding (Jacobs et al., 2020) when recalling and describing their assessment test construction.

As the focus of this article is on recalling and describing teachers' sense-making in relation to assessment of mathematical competencies by teacher-constructed tests, a short literature review of competencies, assessment, and teacher-constructed tests follows below.

3 Competencies, assessment, and tests — a literature review

3.1 Competence

Mastering a competence, according to Niss and Højgaard (2011), involves both one's preparedness and *ability* to conduct particular mathematical actions. Lithner et al. (2010) present a mathematical competencies research framework (MCRF), which aims to facilitate empirical analyses of students' competency development. Competencies, within MCRF, are described as problem-solving, reasoning, procedural, representation, connection, and communication competencies. Even though the idea behind the MCRF project has been directed towards clarifying the role that national tests have for competence goal implementation and there are consequently natural similarities between the framework and the curriculum in Sweden, no explicit connection has been nationally addressed by the Swedish National Agency of Education.

In this paper, I use Niss's definitions of competence and competency: (2003, pp. 6–7):

> To possess a *competence* (to be competent) in some domain of personal, professional or social life is to master (to a fair degree, appropriate to the conditions and circumstances) essential aspects of life in that domain. *Mathematical competence* then means the ability to understand, judge, do, and use mathematics in a variety of intra- and extra-mathematical contexts and situations in which mathematics plays or could play a role. […] A *mathematical competency* is a clearly recognisable and distinct, major constituent of mathematical competence.

Examples of such competencies are mathematical reasoning, mathematical modelling, or mathematical problem-solving. In the English version of the article mentioned above, Niss and Højgaard (2011, p. 49) accept the notion of *competency* as "a well-informed readiness to act appropriately in situations involving a certain type of mathematical challenge."

3.2 Assessment

Following Niss (1993, p. 3), I consider assessment in mathematics to be:

> […] the judging of the mathematical capability, performance and achievement –– all three notions to be taken in their broadest sense – of students whether as individuals or in groups […]

Assessing competency thus involves being able to reveal, evaluate, and characterize students' mathematical improvement and competencies (Niss & Højgaard, 2011, p. 87). *Classroom assessment* is conceived by Goos (2020, p. 572) in terms of activities undertaken by teachers to obtain, understand, and utilize evidence of student learning to guide subsequent action. Test construction should hence create opportunities for teachers to adopt these above-mentioned processes in relation to the competencies as they are presented by mathematics curricula.

The notion of assessment task in this article means a particular assessment item or assignment that might be found in tests, in textbooks, or as a part of classroom teaching practices. Bergqvist and Lithner (2012) point out modest opportunities for students to experience and develop some aspects of creative mathematical reasoning from teachers' task-solving presentations. These aspects of creative mathematical reasoning, tightly connected to mathematical competencies, therefore also become very important in a practical assessment context. This is confirmed by Horoks and Pilet (2017) who characterize teachers' assessment task practices from three perspectives: the distance between assessment tasks and tasks from previous teaching practices, the depth of assessment information obtained, and the interpretation and utilization of the information (curricula, external assessment). Swan and Burkhardt (2012, as cited in Suurtamm et al., 2016) study assessment tasks in the context of design principles for high-quality assessment and the complexity of mathematical competence, presented nowadays in the terms of educational and learning outcomes worldwide. Such tasks are further considered as broader than typical tasks, requiring higher cognitive and language levels, as well as taking more time to solve, and are often complex and unfamiliar to students. The authors also recognize some difficulties in relating student performances to levels or stages of mathematical competencies when designing tests. Thus, assessing competencies is a challenge for teachers, as Niss et al. (2016, p. 630) point out:

> […] there is a huge task lying in front of us in making competency notions understood, embraced and owned by teachers […]

This task may also involve teacher-constructed tests.

3.3 Tests

Tests are a relied-upon and regular part of teachers' assessment practices (Nieminen & Atjonen, 2022; Senk et al., 1997; Stiggins & Bridgeford, 1985) and might be conceived of as the implemented (experiential and learned) curriculum (Osta, 2020). Nortvedt and Buchholtz (2018) address the existence of teachers' testing practices that emphasize continuous assessment of students' procedural, as opposed to problem-solving, skills. Traditional tests focus mainly on teaching or what has been taught, as opposed to learning by connecting different areas of content (Burkhardt, 2007).

Wiggins (1992) describes three difficulties that teachers in mathematics and other subjects face in the construction of tests. First, it is problematic to incorporate non-routine tasks, as these may prove too challenging for students, who may not have the ability to engage in original thinking to solve these problems. Second, constructing tasks that address students' knowledge at different levels is difficult and time-consuming. Third, the criteria for assessing higher-order thinking are complex to comprehend and use. Based on differences in students' learning and disparities among teachers in setting subject content and assessing its fit to students' competencies and higher cognitive processes, Watt (2005) regards tests as invalid, but nevertheless finds that teachers perceive them to be useful. Recalling studies by Simon and Forgette-Giroux (2000) and Firestone et al. (2000), Watt (2005) states that "traditional" testing has difficulty assessing mathematical competencies and higher-order cognitive processes, as opposed to simply testing the successful execution of mathematical procedures. Niss (1993) offers one possible answer as to why this is the case: "The more complex abilities a task encompasses, the more difficult it is to interpret its outcome in a reliable way" (p. 19).

Frary et al. (1993) detect problems experienced by teachers in interpreting scores in criterion-based assessment, where some predetermined criteria illustrate (levels of) expected performance. They find that teachers' training in developing tests is insufficient, and that teachers' test construction efforts are mainly grounded in their own experience and the support of their colleagues. Nichols and Sugrue (1999) have also detected weak abilities among teachers to interpret scores and to detect higher-order thinking in students.

New teachers in the USA often do not feel confident in testing, and those who are more experienced use common textbooks and teacher's guides as a help (Burke, 2009). Senk et al. (1997) also found teachers' tests to be greatly influenced by tests from the textbooks. US high-school textbooks that offer both evaluation of students' mathematical knowledge and insight into students' management of mathematical processes (reasoning, communication, etc.) indicate inconsistencies regarding opportunities for students' engagement in these processes (Hunsader et al., 2014). Similar results have been identified with regard to problem-solving in textbooks from twelve countries (Jäder et al., 2020). Mathematical learning in Sweden, closely in line with the USA (see above), is predominantly represented by textbooks, tests, and teaching (Lithner, 2004), where all three components are tightly interwoven. Students' textbook task-solving has been characterized by a dominance of rote learning, trivial reasoning, and imitative strategies (Sidenvall et al., 2015). The notions of *imitative* and *creative* mathematical reasoning, developed by Lithner et al. (2010), are used by these authors to compare task requirements on national and teacher-constructed tests. The distinction between imitative and creative mathematical reasoning is made based on whether students have previously engaged with the same types of tasks. Their results indicate that success on teacher-constructed tests requires imitative reasoning, whereas tasks on national tests require creative mathematical reasoning. If students mainly practise imitative algorithmic tasks in teaching activities (Boesen et al., 2014), they demonstrate difficulties when given creative tasks (Norqvist et al., 2019). Boesen (2006) explains this difference between the two modes of mathematical reasoning found in the content comparison of teacher-made and national tests in terms of insufficient familiarity with these modes of mathematical reasoning, assessment expectations, and results.

Although the results may seem disappointing, current research does not seriously take into account the perspective of teachers, who are most familiar with their own students, as well as teaching and learning classroom situations. In acknowledging a situated perspective (Boaler, 2000; Lave & Wenger, 1991) on learning in the classroom, the teachers' sense-making about their assessment practices and the assessment requirements of the curricula become relevant. For this reason, a short description of curriculum development also follows.

4 The Swedish context

Since the 1960s, Swedish curriculum development in mathematics has undergone a change from focusing mostly or solely on the teaching of subject content to instead focusing on instilling *abilities*, the national equivalent of the notion of competencies, and paying more attention to assessment. This change is evident in a comparison of four sets of curricula from 1965 (Swedish National Board of Education [Skolöverstyrelsen], 1965), 1994, 2000, and 2011. The 2011 curriculum that forms the basis for this study, described below, includes both subject content and assessment criteria for each of the courses at the national level. The same is valid for the current curriculum after the most recent revision in 2021 (Swedish National Agency for Education, 2021). Teachers in Sweden are supposed to follow the instructions presented by the curriculum with some degree of freedom, for instance, regarding choosing modes of teaching and assessment strategies in achieving the educational goals.

In the 1994 curriculum (Swedish National Agency for Education, 1994) for upper secondary schools, where students are aged between 16 and 19 years, two qualitative levels — G, pass, and VG, pass with distinction — were used. The 2000 revision (Swedish National Agency for Education, 2000) added an additional level, MVG, pass with special distinction. In the 2000 curriculum, the formulation of the grade criteria already reflected a focus on the ability to handle concepts and procedures and to solve problems, as can be seen in the following (my own translation from the Swedish):

> G: Students use appropriate mathematical concepts, methods, models, and procedures to formulate and solve problems in one step.
>
> VG: Students use appropriate mathematical concepts, methods, models, and procedures to formulate and solve different types of problems.
>
> MVG: Students formulate and develop problems, choose general methods and models for problem-solving, as well as demonstrate clear thinking in correct mathematical language.
> (Swedish National Agency for Education, 2000)

The 2011 curriculum introduced seven abilities, not explicitly defined except for the interpretive formulations given by the commentary material, at three qualitative levels, with some overlap with the abilities previously referenced. These are conceptual, procedural, reasoning, problem-solving, modelling, and communication abilities, as well as the ability to use and relate mathematics to other subjects, both within and outside school. The last of these competencies is not included in national and teachers' testing. The 2011 curriculum also introduced grading criteria differentiating between three qualitative levels E, C, and A for each of the abilities. If all requirements for the lower grade level, E or C, are fulfilled but not all requirements for the higher grade level, C or A, the intermediate

Table 1 Qualitative levels of modelling ability

E	D	C	B	A
In their work students re-express and transform realistic problem situations into mathematical formulations by applying *given* mathematical models		In their work students re-express and transform realistic problem situations into mathematical formulations by *choosing* and applying mathematical models		In their work students re-express and transform realistic problem situations into mathematical formulations by *choosing*, applying, and *adapting* mathematical models

Note. The qualitative grade levels E, D, C, B, and A for the modelling competency are presented in the first row with the written text from the curriculum within the subsequent columns E, C, and A

D and B are the intermediate grade levels, applied if all the requirements for the lower level, E or C, are fulfilled but not for the higher levels, C or A, respectively

grade levels D and B are applicable. There are two main constituent parts of the 2011 curriculum: the first concerns the subject content included in courses, the content knowledge, while the second concerns grading criteria, the knowledge requirements. The examples in Table 1 offer descriptions of different qualitative levels for conceptual ability, in text, and modelling ability. Similar descriptions can be found for the other abilities.

> E: Students can *with some certainty* show the key concepts in action, and *in basic terms* describe their meaning using *some* other representation. In addition, students switch *with some certainty* between these representations. Students can with *some certainty* use concepts […].
>
> C: Students can *with some certainty* show the key concepts in action, and *in detail* describe their meaning using *some* other representations. In addition, students switch *with some certainty* between these representations. Students can with *some certainty* use concepts […].
>
> A: Students can *with certainty* show the key concepts in action, and *in detail* describe their meaning using *several* other representations. In addition, students are able to switch *with certainty* between these different representations. Students can with *certainty* use concepts […].
> (Swedish National Agency for Education, 2012)

5 Methodology and method

The methodological approach of grounded theory (Glaser & Strauss, 1967) was employed when working with the interviews, both in developing the questions and in transcribing, re-interpreting, and coding the teachers' sense-making in relation to their test construction and the competency requirements from the curriculum. In accordance with symbolic interactionism (Clark et al., 2021), the role of grounded theory was to capture the meaning-making processes through which the teachers construct their own tests for assessing mathematical competence. The interview establishes a social interaction between a teacher and the researcher, in which the teacher's sense-making is communicated to the researcher (Blumer, 1969). By recalling and describing their test construction within the interviews, the teachers notice what they attend to, interpret, and make decisions about (Jacobs et al., 2020).

5.1 Sample

Teachers were invited to participate voluntarily in the study. The information they received, via email or phone, included a short description of the study, its aim and structure, and what was asked of participants, including a clarification of the ethical considerations (Swedish Research Council, 2023; Uppsala University, 2023).

Six mathematics teachers, whose teaching experience ranges from 2 to over 20 years, teaching at different upper secondary schools in Sweden, both public and private, as well as in different upper secondary school programmes, were involved in the study. Two of the teachers' jobs were specifically concentrated on the development of mathematics at schools, and one of the teachers had a research programming background. Even though the teacher sampling is not representative, it helps to identify some teachers' ways of relating test practices to student competencies.

5.2 Data collection

The empirical data in this study was collected by means of semi-structured individual interviews with the teachers, conducted by the author via telephone calls at predetermined times with the interviewees. During the interviews, each of the teachers related the meaning that she/he assigned to her/his own test construction to the requirements from the national curriculum for assessing students' competencies on three qualitative levels. Semi-structured interviews provided opportunity and flexibility for teachers to relate meanings assigned to their own test construction to the competency requirements. At the same time, opportunities were created for the interviewer to pose relevant questions, request explanations, and obtain relevant knowledge that may be used in deciding and steering the topics discussed (Robson, 2002). The interviews, in total, lasted 4 h and 25 min. All interviews were audiotaped and transcribed using Transana software. Memos were also written to record the author's observations during the coding process. Additionally, several of each teacher's most recent tests were gathered in advance to serve as a basis for the interviews, in total 19 tests with 253 tasks. Eighty-four tasks were composed of sub-tasks. In most of the tests, the tasks included associated scores (143) and/or rubrics (145) (see example tasks 10 and 11 below). Some of the teachers handed in students' results (40 tasks), assessment instructions (128 tasks), and their own written assessment examples (24 tasks). Three (50 tasks) of these 19 tests were taken directly from the textbooks.

These tests constituted the basis for the design of the first interview questions and the basic question framework. The basic question framework, which concerns the purposes of testing, test construction, selection of tasks and grading, as well as the alignment of teachers' tests with the *competency* requirements of the curriculum, would be further adopted and developed in the course of theoretical sampling. Gradually, the successive interviews, transcripts, and memos led to the acquisition of new knowledge, and the identification of new knowledge gaps, which had to be filled by developing additional questions. This increasing knowledge and category collection, *theoretical sampling* (Glaser & Strauss, 1967, p. 45), (see

Sect. 5.3, how memos are used, and Table 3), was used to map out plans for generating new empirical and theoretical knowledge, categories, with associated *properties* (characteristics) of test construction and their *dimensions* (possible variations), from one interview to the other. This involved continual analytical comparisons by the author between the old and the new emerging categories and their successive conceptualization.

Theoretical sampling and the author's *sensitivity*, that is "the ability to pick up on subtle nuances and cues in the data that infer or point to meaning" (Corbin & Strauss, 2008b, p. 19), to emerging data served as the main resources for analytical progress.

5.3 Analyses of the data

Using the grounded theory approach (Glaser & Strauss, 1967), each transcript, divided into small excerpts (see below), accompanying memos and the knowledge acquired served as a source for the adjustment of previous interview questions and the formulation of new ones, in an iterative process completed after each round of interviews: interview – transcript – analyses – conceptualization (categories, memos) – new questions – interview—... The purpose was the generation of a *core category*, an explanatory whole, concerning the research aim in focus. Evaluation criteria developed by Corbin and Strauss (2008a) guided the analyses. The criteria encompass *fitting* experiences from the research field, *usefulness* for application and development, *conceptualisation* that findings are built on, *contextualisation of concepts*, *logic*, such as logical flow of ideas, *depth*, *variation*, *creativity*, *sensitivity*, and *evidence from memos*.

Let us look at an excerpt from one of the interview transcripts, together with the interviewed teacher's constructed test-scoring rubric (Table 2), to highlight a part of the initial analysis. The first column presents the mathematical competencies (translated into English) while the letters E, C, and A stand for the three qualitative levels of the competencies from the curriculum. The numbers indicate the maximal scoring for the competencies and their qualitative levels. The rubric displays 2 scores for the conceptual competency on E level and 1 score for the modelling competency on E and C qualitative levels, respectively. Scoring on E level is a prerequisite for scoring on C level. The annotation 3/1/0 implies 3 scores on E qualitative competency level, 1 score on C, and 0 scores on A.

The following extract from a memo presenting the researcher's analytical considerations throughout the analysis of the transcript, a *category* labelled as *concept*, revolves around the actions of thinking, reinterpretation of the teachers' interpretation, forming opinions, and developing ideas about content in the transcript. This category might remain the same or be changed during the process of generating additional categories, depending on newly acquired knowledge insights.

> 10. The volume of water in a swimming pool is found by using the mathematical formula $y = 500 - 1.2x$, where y stands for the amount of water in m^3 that is left in the swimming pool after x minutes. (3/1/0)
>
> a) What does 500 mean in the formula?
> b) What does 1.2 mean in the formula?
> c) When is the swimming pool empty?

Table 2 Scoring rubric

	E	C	A
Conceptual	2		
Procedural			
Problem-solving			
Modelling	1	1	
Reasoning			
Communication			

Note. The first column includes the competencies while three qualitative levels of the competencies E, C, and A are placed in the first row. The numbers present the maximal scoring for conceptual and modelling competencies and their qualitative levels from the curriculum

> **Teacher**: […] Those first questions […] there's a straight line and there are the kind of constants that mathematics textbooks prioritize *k* and *m* [often called m and y-intercept in English] […] and that's conceptual competency […] as it says […] by the Swedish National Agency of Education […]: 'The student can with some certainty describe the signification of key concepts […] using some representations and with some certainty describe connections between them'. I'm confident about that. But then, when the swimming pool is empty and you're supposed to make an interpretation, I feel, it's really difficult to see the difference between the E and C levels. If we look at what's written [in the curriculum] for modelling competency, you're supposed to *apply* given models for quality level E of modelling competency, while you're supposed to *choose* and *apply* models at the C quality level. […] You have to choose somehow […], some mathematical activity has to be done. When students write only the solution to a task, without showing very much, they've still *applied*, but nothing more. […] I want to see a little more than just division or something like that in solutions. […] When I'm discussing it now, I realize that points might also be assigned for reasoning and communication competencies as well. If after your initial decision on how to score an item you look at it again, you realize there's a little bit of everything. (Teacher 6, Time: 0:17:34–0:19:52)
> **Memo**: *Scoring item*
> The teacher's focus is initially concerned with mathematical content. This mathematical content concerns not only the concept of a straight line ($y = kx + m$) but also the *k* and *m* values as its constituents. From the CONTENT KNOWLEDGE (that is, mathematics included in the curriculum) the interviewee/teacher shifts gradually to the KNOWLEDGE REQUIREMENTS (that is, mathematical competencies and the *assessment* of these competencies within the curriculum). In *assessment* terms, the straight line and the values *k* and *m* are identified as "*Conceptual COMPETENCY*". Prospective correct solutions to the first two assignments are therefore scored with two points and identified as *conceptual COMPETENCY* at the lowest *QUALITY LEVEL* (E) (Table 2).
> The interviewee has feelings of great confidence in assigning points to mathematical content. […]

In the transcript and the extract from the memo above, the following notions are initially treated as categories with associated properties and dimensions: conceptual competency, modelling competency, scoring (either expected or conceivable and actual student solutions), phases (i.e., constructional and marking phases), and qualitative levels of competencies (or degrees of attention to them). These are then further developed by subsequent interviews.

The procedure of the analysis and creation of theoretical sampling (Table 3) had already begun during transcription of the first recorded interview. The transcribed texts were first analysed, line by line, and memos were added, then divided into small but meaningful and relatively independent sections, based on content (see the excerpt above), and finally placed in separate folders by the Transana software. The folders were named with key words. Grounded theory calls for the *open coding* (Table 3) of such texts. Through this coding and key words, some initial categories, with related preliminary properties and dimensions, were identified.

This newly acquired knowledge, that is, an initial theoretical sampling conducted as a way of collecting data based on theoretical insights, results in categories inside the folders and was used as a source for the next interview and so on. These existing folders were filled in and new folders were created to cover novel knowledge. The transcribed sections in the folders underwent a process of *comparative analysis* through *constant comparisons* between already acquired and familiar knowledge about the teachers' test practices as assigned by interpreted meanings of these practices within the social interaction during the interviews. This process provided opportunities for discovering new interpretations and additional categories as well as their initial conceptualization. The generation of new categories or their conceptualization was derived from identified properties and dimensions of previous categories, through *axial coding*. The axial coding generated new categories by merging categories from open coding (see Table 3). This gradually broadened, as well as deepened, the theoretical samplings, for instance, different degrees of attention given to the competencies and the mathematics by teachers, through distinct phases of test construction.

Finally, through *selective coding*, a process of unification and refinement of the explanatory whole, that is the core category, was completed. The role of the selective coding was to continue the process of developing an overall theoretical framework to describe test construction based on the meanings that the teachers assigned to their test practices in the interviews. In concrete terms, with the division into mathematics (i.e., the content knowledge), and the competencies (i.e., knowledge requirements), a couple of still unclassified categories with various focuses remained from the axial coding. For instance, some of these categories, such as the test construction, the access to student solutions, and the variable teacher interpretations related to some given assessment context, laid the theoretical foundation for the consideration of two main categories: *attention to competencies* and *phase* in test construction by teachers. Furthermore, *quality of paying attention to competencies* and *division of phases* are identified as two properties within these main categories with two accompanying dimensions: three different degrees of the attention to the competencies; no attention, superficial attention, and qualitative attention, as well as two phases; constructional and marking. These six categories mirror the teachers' noticing of what they attend to, interpret, and make decisions about. During this final selective coding, at the same time, all of the interviews and transcripts were, one after the other, regularly and thoroughly reanalysed and revised against the emerging results and possible reconceptualization. This process was considered to be complete when neither additional analyses of theoretical sampling nor new interview insights about test

Table 3 The process of analysis

Coding	Open	Axial	Selective	
Examples	Earlier curriculum, testing mathematics, mastering the whole course, alignment …	Mathematics and content knowledge at different levels	Focus on content knowledge	No attention Superficial attention Qualitative attention
	Student solutions, boring test, assessment of what, uncertainty, grade limits, steering by national tests, challenge, test arrangement, alignment with national testing, summative and formative tests …	Assessment, grades, curriculum, national tests, peer-assessment, test, test construction, testing aims, result	Focus on competencies	No attention Superficial attention Qualitative attention
	Opportunities for students, comprehension, assessment of competencies, formative aims, higher competencies …	Competencies, knowledge requirements, test assessment		

construction suggested any need to create new conceptualization or categories. Expressed in terms of grounded theory, the *saturation* of the data collection has been achieved (Glaser & Strauss, 1967). Table 4 presents the generated core category "*The emerging taxonomy*".

The generation of a core category through three types of coding and this continuous process of comparative analysis from a bottom-up research perspective is also adopted and suggested as a part of validation within the grounded theory approach: "let the research findings speak for themselves" (Corbin & Strauss, 2008a, p. 305). Ten general and evaluative criteria for qualitative findings drawn from the approach of grounded theory have also been elaborated by the authors. These evaluative criteria have been continuously applied here in the analysis for checking and improving the quality of the results (see the first paragraph in this section).

Table 4 The emerging taxonomy

Degrees of attention		Phases of testing	
		Constructional (C)	Marking (M)
(QA)	Attention to qualitative level of competencies in addition to mathematics of varying complexity	Considers whether selected competencies can be shown at different levels	Considers whether there is evidence that a student has the competency at a certain qualitative level
(SA)	Attention to presence or absence of competencies in addition to mathematics of varying complexity	Considers whether selected competencies can be shown by the student	Considers whether or not the competencies are demonstrated
(NA)	No attention to competencies, only to mathematics of varying complexity	Considers what mathematics concepts to include and how difficult the tasks should be	Considers how complex the mathematics is: how many steps are needed in a solution?

Note. The three levels of attention to the competencies, no attention (NA), superficial attention (SA), and qualitative attention (QA) within the constructional (C) and marking phase (M) of testing conduct the taxonomy of teachers' test construction C-NA, M-NA, C-SA, M-SA, C-QA, M-QA

6 Results

The core category "the emerging taxonomy" of teachers' test construction practices, with categories arranged taxonomically from less to more comprehensive, is empirically derived from reinterpretations of the meanings created by the interviewees using the methodology of grounded theory. The taxonomy emphasizes teachers' generated meanings and is presented in Table 4.

As a part of making decisions, two different phases—*constructional* (C) and *marking* (M)—are detected, with three different degrees of attention to competencies considered in each. The division between phases is based on teachers' access to student solutions, which are present in the marking but not the constructional phase. As a part of attending, three degrees of attention to competencies are also identified in this study: *no attention* (NA), *superficial attention* (SA), and *qualitative attention* (QA). In the first degree, there is no attention to competencies, as the name implies. In the second degree, attention to competencies is of a superficial nature: only the presence or absence of competencies is considered. In the third degree, the qualitative level of competencies is considered.

This taxonomy does not necessarily mirror the individual teacher's predisposition. A single teacher might be in any of the categories within test construction and scoring tasks at different times. In the excerpt, the teachers describe sequences of actions, in terms of meanings assigned to their test construction, with a starting point in mathematics, that is, the content knowledge, and with a subsequent transition to competencies, that is, the knowledge requirements (see the excerpt and memo in Sect. 5.3). Similar action sequences, from the content knowledge to the knowledge requirements, are also apparent in scoring processes. The taxonomy provides a theoretical framing for the empirical field of this study in alignment with the evaluative criteria described above. Teachers' noticing, characterized by attending to, interpreting, and decision making, constitutes the *core variable*.

The following section gives a detailed exemplification of the phases of testing, the degrees of attention, and the emergent taxonomy with its content. Because the result is a product of the methodological and analytical approach of grounded theory, as described and exemplified earlier, it will be presented without any further analyses.

6.1 Phases

The constructional phase, occurring before students take the test, comprises interpreting the curricular requirements and making decisions about *test preparation*, *task construction*, and *task selection,* as well as *task and test rubric construction*. Rubrics are constructed to specify the expectations for possible solutions with regard to competencies, and include assigning competencies and scores to the tasks, weighting the test tasks, and grading the whole test. Alone or in cooperation with colleagues involved in the same courses, depending on varying school assessment policies, teachers construct, select, and assign competencies and scores to tasks in relation to what they attend to and what they anticipate that the student solutions might look like. Tests are thus initially based on what teachers notice, and on their expectations, and these expectations and the rubrics are judged and revised or reinterpreted by teachers and their colleagues in the marking phase, after the acquisition of student solutions. The marking phase is the phase in which teachers evaluate and assess actual student solutions. In this phase, by means of reinterpretations, teachers obtain new assessment-related insights, which are then evaluated and may be adopted:

> T: After every test we have assessment meetings and there are many things we look at afterwards; unfortunately, too much. And so, you see, we should maybe give additional points for this, or we might have changed our opinions about [values]. We have an assessment meeting tomorrow for the course Mathematics 2C and I know I have some opinions that I want to check with my colleagues about. We'll change the assignment of points a little.
> (Sample: Phases, Clip: Assessment conference, Time: 11:59:09–12:28:03)

6.2 The degrees of attention to competencies throughout the phases

The three degrees of attention, NA, SA, and QA, are exemplified through the two phases C and M. These three degrees of attention together with the two phases have created the six categories of the taxonomy. For each of these categories, a specific teacher has been empirically chosen. For this reason, the teachers are not given distinct names.

The first degree of attention to competencies, no attention (NA), relates to noticing to and being concerned with subject content only, paying no attention to competencies or connections between them or between subject content and competencies. Let us consider an interview excerpt on the constructional phase (C-NA):

> I: If we look at how good the correspondence is between your tests and the competencies described in the curriculum …
> T: I haven't thought specifically about it.
> I think I have to read a little about the goals of the curriculum in order to see.
> (Sample: C-NA, Clip: Alignment, Time: 16:59:00–17:27:02)

With students' solutions in mind, in the marking phase (M) with this degree of attention (NA), the teacher notices only how many mathematical steps are present in the solution to the task. Competencies are not considered. The notation 1/2/0 below still means 1 point on the E qualitative level, 2 points on the C level, and 0 points on the A level. The notion of the VG grade, pass with distinction, mentioned in the *Swedish context* section, refers to the previous curriculum (1994 and 2000):

> 11. John, Malin, and Anna weigh 120 kg together. John weighs 12kg more than Anna. Malin weighs twice as much as Anna. Let Anna's weight be x kg. Write the expressions for John's and Malin's weights. Calculate Anna's weight by setting up and solving an equation. 1/2/0
> I: Could you explain how you were thinking there with 1/2/0?
> T: Yes, I was thinking, John, Malin, and Anna weigh 120 kg together. There I thought, if they [the students] have done something on this task, if they've understood the text, they have to have set up some kind of [mathematical] expression then, made some attempt; I've given 1 E point. If they've set Anna's weight to a number of kilos [x] and then they can express the other people's weight based on Anna's weight, and write a correct mathematical expression, I've given 1 VG point too. And, if they manage to solve for Anna's weight [x] by setting up that equation correctly, then I've given 2 VG points; that is, full points.
> (Sample: M-NA, Clip: Mathematical steps, Time: 14:18.0–15:21.3)

In the second degree of attention to competencies, superficial attention (SA), attention is paid to subject content and the presence or absence of competencies without reference to qualitative levels. Here is an example pertaining to the constructional phase (C):

> T: Regarding test construction, it was just about finding the subject content and then making sure there were tasks for the different grade criteria, grade levels: G [pass], VG [pass with distinction], and what we need for 'sun tasks' [meaning tasks that are specially designed to test higher-level mathematical competencies]. And then, of course, you have to try to find some tasks that have to do with concepts and some tasks that have to do with procedures and modelling.
> (Sample: C-SA, Clip: Concept and procedure, Time: 05:30:02–06:19.08)

Competencies here are considered only on a trivial level, in terms of whether they are present or not. Conceptual, procedural, and modelling competencies are linked to test tasks, but with no connection to the quality.

In the marking phase (M), superficial attention (SA) is limited to presence or absence of competencies in a rubric on the first page of the test. The qualitative levels of competencies are not explicitly noticed inside the rubric but are mainly decided based on interpretations of the mathematical steps needed to reach an appropriate task solution. The idiosyncratic combination of teacher expectations and mathematical steps is a weak indication of actual qualitative competency. In the next excerpt, a rubric refers to both the mathematics tested and competencies required without their qualitative levels:

> T: First I look at this rubric: if you [the student] can manage these things well, all the crosses to the left, so to speak [...]. Sometimes, not every time actually, but sometimes I also use an upper part [of the rubric with the competencies], and I note that it's like the level, a little bit, of these different competencies that I'm assessing. [...] This is on quite a high level, maybe up to C [level] or even better.
> (Sample: M-SA, Clip: Abilities, Time: 18:45.6–20:54.4)

Although the competencies are listed in the teacher's rubric as well as parts of the subject content, this assessment information contains no differentiation between qualitative levels. A slight differentiation, however, is provided with regard to subject content: "can do well" and "can do partially". Level C, which the teacher talks about here, thus assesses both mathematical steps and competency in this case.

At the qualitative degree of attention (QA), teachers notice both subject content and the qualitative level of competency. Again, let us look at an excerpt from the constructional phase (C) where requirements and actions to complete the task revealed the presence of the conceptual and the procedural competency without a clear link to the competency levels via scoring:

> I: You write 2E, for example, or 1C, and I want to know how these points and levels come into the picture.
> T: You could say this: that the test I constructed, or yes, I chose tasks, and then I did solutions later. Then I analyse my solutions. And then I think a little bit about whether there are alternative solutions, of course. […] So, I try to just break it down then. What is it that you're doing and what does it require? And like I say, then it's almost always the case that you begin, as I see it, with concept and procedure points.
> I: … on these different … different levels …
> T: … yes, exactly.
> I: So that you assess, in a sense, if I understand correctly, you actually assess your expectations in some sense … about solutions …
> T: … *mmm* [indicating agreement] … *mmm.*
> (Sample: C-QA, Clip: Different levels, Time: 0:21:42.3–0:23:12.4)

Finally, let us look at an excerpt representing noticing in the marking phase (M) and qualitative attention (QA), for a task called "How high is the church tower?" (Fig. 1).

> I: I see you've written here 1/2/1. Can you briefly
> describe what you were thinking about levels, competencies?
> T: First of all, you have to realize that here I need to use 'tan' to describe this situation. The description of the situation this has to do with comes under modelling. But I think they get problem-solving points because they haven't seen a task like this before. And then, they need to think in a new way. But both of those are on the C level. Then, to solve the 'tan' thing itself is not so complicated, so that's one procedure point at the G level. Or sorry, E level.
> …If you add a height [the height of the person] that you don't know, you've shown a higher abstraction level and a certain reasoning competency … … you've shown good … high quality here.
> (Sample: M-QA, Clip: Tower, Time (part of): 14:36:8–18.03.1)

Here, the teacher makes a link between the question and modelling competency, as the problem comes from a real situation, and suggests that problem-solving competency is related to students' previous experience, or lack thereof, with tasks of a similar type. The teacher's noticing of the competencies at qualitative levels is evident from the use of the words "higher" and "certain".

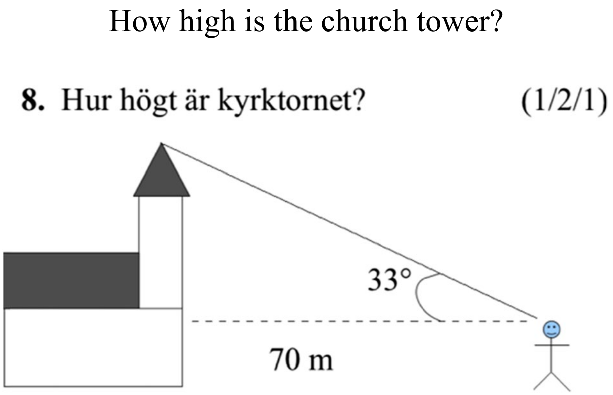
How high is the church tower?
8. Hur högt är kyrktornet? (1/2/1)
33°
70 m

Fig. 1 How high is the church tower? Note. The annotation 1/2/1 implies 1 score on E, 2 scores on C, and 1 score on A qualitative competency level (own translation of the task title)

7 Discussion

The emerging taxonomy (Table 4) of meanings assigned by teachers to their own test practices for evaluating mathematical competencies captures a theoretical framing of test construction practices from the competency perspective. It reveals a great amount of possible variation in how teachers relate their test practices to requirements to assess students' competencies. This variation is not only evident between teachers themselves, but also between their individual test and task construction in relation to assessment at different qualitative levels, both empirically, in the evidence from the interviews, and theoretically, depending on different levels of paying attention to or noticing competencies and depending on two different phases.

The results of the study should not simply be seen as categories and their conceptualization that represent individual teachers' positions in relation to competency considerations in tests. Rather, the results present and reveal ways in which meanings assigned to test practices might vary. An individual teacher may exhibit several of these approaches at different stages of their work with tests. The variation might exist not only between teachers' different tests, but also between different tasks within the same test, which is something empirically and theoretically derived, as well as being unexpected. As an example, both the transcript excerpt, from the section analyses of the data (Sect. 5.3; Teacher 6), and the transcript excerpt illustrating the category C-SA, from the results section (Sect. 6.1; Sample: C-SA), are taken from the same teacher. The quotation, "sometimes, not every time actually, but sometimes I also use an upper part [of the rubric with the competencies]" might serve as another example from another teacher in the illustration of the category M-SA. This study and the emerging taxonomy illustrate challenges in assessing competency on qualitative levels, which might broaden the scope and importance of the study to an international research public.

These practices and challenges in competency assessment are in line with the literature that highlights the complexity and difficulties involved in competency assessment (Niss, 1993; Niss et al., 2016; Suurtamm et al., 2016; Watt, 2005; Wiggins, 1992). This complexity might have effects on both teaching and testing (Burke, 2009; Horoks & Pilet, 2017; Senk et al., 1997) and teachers' tendency to mainly focus on procedural competency in teaching practices (Boesen et al., 2014). Corresponding conclusions have also been identified with regard to testing mainly mathematical procedures (Boesen et al., 2010; Burkhardt, 2007; Nortvedt & Buchholtz, 2018; Watt, 2005). For these reasons, the taxonomy may contribute to clarifying, promoting, and improving teachers' test construction practices for competency assessment.

A part of the literature review points not only to a strong influence of competency assessment on classroom teaching and testing, but also to an analogous influence from a textbook perspective (Burke, 2009; Hunsader et al., 2014; Jäder et al., 2020; Lithner, 2004; Senk et al., 1997; Sidenvall et al., 2015). Furthermore, an important distance between teaching and assessment tasks has been noticed (Horoks & Pilet, 2017). If teachers practise more procedural exercises with students during teaching and are compelled to produce new approaches to testing other competencies, problems also might emerge (Norqvist et al., 2019). There is perhaps a reason to look carefully at coordination between competency requirements, textbooks, teaching, and testing for comprehension and explanation of assessment variation within the taxonomy. For instance, insufficient coordination between teaching and testing practices might make test construction more time-consuming. Teachers have to deviate sometimes from applying tasks from teaching sequences in test construction if they want to test, for instance, problem-solving, creative reasoning, and modelling competencies. This raises the question of the ecological validity of test constructions.

Beneficially, the results indicate some of the possible modes for test construction. The degrees of attention inform how competencies might be included throughout the interplay between the constructional and marking phases. Newly acquired insights and knowledge from the marking phase should be analysed, identified, followed up, and applied with formative purposes in mind. This could have great relevance for the development, adjustment, and coordination of new instructional, teaching, and testing practices.

The methodological approach of grounded theory, accompanied by the author's experience of and sensitivity to both the educational level and the phenomenon in focus (Corbin & Strauss, 2008c), took the direction of the study away from the interviews and transcripts as a whole towards a theoretical explanation of the test construction, that is, away from the particular teachers in this study.

Teachers in Sweden, including the teachers in this study, are expected to follow the national assessment instructions in the curriculum. Although this study is limited to the Swedish context, the results provide general directions with respect to the degree of attention and the two decision phases that could be adapted and also applied to how teachers recall and describe their assessment test construction in other countries.

8 Implications

This study has reconstructed teachers' reflective sense-making of their previously conducted test constructions. However, it does not provide insight into how teachers actually construct tests and what they notice during these processes. As the interplay between the constructional and marking phases are relevant in any test construction, the taxonomy may serve as a sensitizing framework for future research on teachers' test construction. In addition to the theoretical potential of the taxonomy to guide future research, the interplay between the constructional and marking phases also has practical implications. For example, the analysis of students' actual solutions during the marking phase can be applied to test improvement by refining the scoring and grading of prospective tests. Student solutions, well utilized by teachers, will be key in this process and could lead to well-elaborated task designs.

Conscious consideration of this interplay might actually constitute a mechanism for improving the quality of teaching before test construction and testing, too. This in turn can improve teachers' awareness of and familiarity with assessment of mathematical competencies as well as boosting their coordination of teaching and competency requirements, independently of the textbook if needed.

The interviews were not conducted during the actual test construction but afterwards, which might leave possible research space for further investigation, as other perspectives may arise in data if it is collected during construction, with additional consequences for the interpretations of the interviewees. The interview was shaped as a social interaction, which is directed by specific expectations. This may be different when various teachers construct tests or recall test constructions during social interactions. This is an area for further study that could complement the findings in this paper.

The taxonomy clearly includes attending and deciding as the parts of noticing, while interpretations are related to the theoretical framing and are therefore baked into the results. Further studies may be able to identify specific interpretations in teacher noticing in test constructions.

While the small number of the interviewees may raise doubts about theoretical saturation and subsequent findings in further studies, the variation and the high number of tasks together with grounded theory and the evaluative criteria in this study strengthen the evidence of the findings.

However, assessment practices also need to continue to reflect not only competencies but also important subject content. The taxonomy, with variation based on attention to competencies, as presented in this paper, may assist in training in test construction for both pre-service and in-service teachers.

Although introducing the idea of competency in school mathematics seems very relevant from many perspectives, both within and outside of school mathematics, its implementation in schools requires careful planning, training, and continual follow-ups throughout mathematical school classroom practices regarding both teaching and assessment.

Funding Open access funding provided by University of Gothenburg.

Data Availability The data that support the findings of this study are available from the corresponding author upon request.

References

Bergqvist, T., & Lithner, J. (2012). Mathematical reasoning in teachers' presentations. *The Journal of Mathematical Behavior, 31*(2), 252–269.

Biehler, R. (2019). Allgemeinbildung, Mathematical Literacy, and Competence Orientation. In H. N. Jahnke & L. Hefendehl-Hebeker (Eds.), *Traditions in German-Speaking Mathematics Education Research* (pp. 141–170). Springer Cham. https://doi.org/10.1007/978-3-030-11069-7

Blumer, H. (1969). *Symbolic interactionism: perspective and method*. Prentice-Hall, Inc.

Boaler, J. (2000). Exploring Situated Insights into Research and Learning. *Journal for Research in Mathematics Education, 31*(1), 113–119.

Boesen, J. (2006). *Assessing mathematical creativity : comparing national and teacher-made tests, explaining differences and examining impact* (Publication Number 34) [Doctoral thesis, comprehensive summary, Matematik och matematisk statistik, Umeå universitet]. DiVA. Umeå. Retrieved October 19, 2012, from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-833

Boesen, J., Lithner, J., & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning students use. *Educational Studies in Mathematics, 75*(1), 89–105.

Boesen, J., Helenius, O., Bergqvist, E., Bergqvist, T., Lithner, J., Palm, T., & Palmberg, B. (2014). Developing mathematical competence: From the intended to the enacted curriculum. *The Journal of Mathematical Behavior, 33*, 72–87.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*(4), 279–301.

Burke, K. (2009). *How to assess authentic learning*. Corwin. Retrieved February 3, 2023, from https://books.google.se/books?id=_ES5DlimlB8C&lpg=PP1&hl=sv&pg=PR4#v=onepage&q&f=true

Burkhardt, H. (2007). Mathematical proficiency: What is important? How can it be measured? In A. H. Schoenfeld (Ed.), *Assessing mathematical proficiency* (Vol. 53, pp. 77–97). Cambridge University Press.

Clark, T., Bryman, A., Foster, L., & Sloan, L. (2021). *Bryman's social research methods* (Sixth ed.). Oxford University Press.

Corbin, J., & Strauss, A. (2008a). Criteria for Evaluation. In J. Corbin & A. Strauss (Eds.), *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory* (pp. 297–312). SAGE Publications.

Corbin, J., & Strauss, A. (2008b). Practical Considerations. In J. Corbin & A. Strauss (Eds.), *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory* (pp. 19–44). SAGE Publications.

Corbin, J., & Strauss, A. (2008c). Strategies for Qualitative Data Analysis. In J. Corbin & A. Strauss (Eds.), *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory* (pp. 65–86). SAGE Publications.

Criswell, B., & Krall, R. M. (2017). Teacher Noticing in Various Grade Bands and Contexts: Commentary. In E. O. Schack, M. H. Fisher, & J. A. Wilhelm (Eds.), *Teacher Noticing: Bridging and Broadening Perspectives, Contexts, and Frameworks* (pp. 21–30). Springer International Publishing.

Department of Education. (2013). *National curriculum in England: secondary curriculum*. Retrieved November 7, 2013, from https://www.gov.uk/government/publications/national-curriculum-in-england-secondary-curriculum

Firestone, W. A., Winter, J., & Fitz, J. (2000). Different Assessments, Common Practice? Mathematics testing and teaching in the USA and England and Wales. *Assessment in Education: Principles, Policy & Practice, 7*(1), 13–37.

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12*(3), 23–30.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*(3), 606–633.

Goos, M. (2020). Mathematics Classroom Assessment. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 572–576). Springer International Publishing.

Horoks, J., & Pilet, J. (2017). Assessment in mathematics as a lever to promote students' learning and teachers' professional development. 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. 3572–3579). DCU Institute of Education and ERME.

Hunsader, P. D., Thompson, D. R., Zorin, B., Mohn, A. L., Zakrzewski, J., Karadeniz, I., Fisher, E. C., & MacDonald, G. (2014). Assessments accompanying published textbooks: The extent to which mathematical processes are evident. *ZDM-Mathematics Education, 46*(5), 797–813.

Jacobs, V. R., Philipp, R. A., & Sherin, M. G. (2020). Noticing of Mathematics Teachers. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 639–641). Springer International Publishing.

Jäder, J., Lithner, J., & Sidenvall, J. (2020). Mathematical problem solving in textbooks from twelve countries. *International Journal of Mathematical Education in Science and Technology, 51*(7), 1120–1136.

Kilpatrick, J. (2020). Competency Frameworks in Mathematics Education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 110–113). Springer International Publishing.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Retrieved November 12, 2012, from http://www.nap.edu/catalog/9822.html

Lave, J., & Wenger, E. (1991). *Situated learning*. Cambridge University Press.

Lithner, J. (2004). Mathematical reasoning in calculus textbook exercises. *The Journal of Mathematical Behavior, 23*(4), 405–427.

Lithner, J., Bergqvist, E., Bergqvist, T., Boesen, J., Palm, T., & Palmberg, B. (2010, January 26–27). *Mathematical competences: A research framework* [Conference session]. Mathematics and mathematics education: Cultural and social dimensions. Madif 7: The seventh mathematics educational research seminar, Stockholm, Sweden. Retrieved August 30, 2018, from http://matematikdidaktik.org/index.php/madifs-skriftserie/

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18*(2), 18–29.

Nieminen, J. H., & Atjonen, P. (2022). The assessment culture of mathematics in Finland: a student perspective. *Research in Mathematics Education*, 1–20. https://doi.org/10.1080/14794802.2022.2045626

Niss, M. (1993). Assessment in Mathematics Education and its Effects. In M. Niss (Ed.), *Investigations into Assessment in Mathematics Education. An ICMI Study*. Kluwer Academic Publishers.

Niss, M. (2003). *Mathematical competences and the learning of mathematics: The Danish KOM project* [Conference session] 3td Mediterranean Conference on Mathematics Education, Athens, Hellas

Niss, M., & Højgaard, J. T. (2002). *Kompetencer og matematiklæring. Ideer og inspiration til udvikling af matematikundervisning i Danmark* (Vol. Uddannelsesstyrelsens temahæfteserie Nr 18).

Niss, M., & Højgaard, T. (Eds.). (2011). *Competencies and Mathematical Learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark*. Roskilde: Roskilde Universitet. (IMFUFA-tekst: i, om og med matematik og fysik; Nr. 485).

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics, 102*(1), 9–28.

Niss, M., Bruder, R., Planas, N., Turner, R., & Villa-Ochoa, J. A. (2016). Survey team on: Conceptualisation of the role of competencies, knowing and knowledge in mathematics education research. *ZDM-Mathematics Education, 48*(5), 611–632.

Niss, M., Bruder, R., Planas, N., Turner, R., & Villa-Ochoa, J. A. (2017). Conceptualisation of the Role of Competencies, Knowing and Knowledge in Mathematics Education Research. In G. Kaiser (Ed.), *Proceedings of the 13th International Congress on Mathematical Education ICME–13* (pp. 235–248). Springer International Publishing AG.

Norqvist, M., Jonsson, B., Lithner, J., Qwillbard, T., & Holm, L. (2019). Investigating algorithmic and creative reasoning strategies by eye tracking. *The Journal of Mathematical Behavior, 55*, 1–14.

Nortvedt, G. A., & Buchholtz, N. (2018). Assessment in mathematics education: Responding to issues regarding methodology, policy, and equity. *ZDM-Mathematics Education, 50*(4), 555–570.

Oescher, J., & Kirby, C. P. (1990). *Assessing teacher made-test in secondary math and science classrooms* Annual meeting of the National Council of Measurement in Education, Boston Massachusetts.

Osta, I. (2020). Mathematics Curriculum Evaluation. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 576–582). Springer International Publishing.

Pettersen, A., & Braeken, J. (2019). Mathematical Competency Demands of Assessment Items: A Search for Empirical Evidence. *International Journal of Science and Mathematics Education, 17*(2), 405–425.

Phelan, J., & Phelan, J. (2010). Classroom Assessment Tasks and Tests. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 209–219). Elsevier.

Robson, C. (2002). *Real World Research*. Blackwell Publishing.

Senk, S. L., Beckmann, C. E., & Thompson, D. R. (1997). Assessment and Grading in High School Mathematics Classrooms. *Journal for Research in Mathematics Education, 28*(2), 187–215.

Sidenvall, J., Lithner, J., & Jäder, J. (2015). Students' reasoning in mathematics textbook task-solving. *International Journal of Mathematical Education in Science and Technology, 46*(4), 533–552.

Simon, M., & Forgette-Giroux, R. (2000). Impact of a Content Selection Framework on Portfolio Assessment at the Classroom Level. *Assessment in Education: Principles, Policy & Practice, 7*(1), 83–100.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, *22*(4), 271–286.

Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., & Vos, P. (2016). Assessment in Mathematics Education. *Springer Cham*. https://doi.org/10.1007/978-3-319-32394-7_1

Swan, M., & Burkhardt, H. (2012). A designer speaks: Designing assessment of performance in mathematics. *Educational Designer: Journal of the International Society for Design and Development in Education*, *2*(5), 1–41.

Swedish National Agency for Education. (1994). *Matematik [Mathematics] (in Swedish)*. Nationellt Centrum för Matematikutbildning [National Center for Mathematics Education]. Retrieved November 20, 2012, from http://ncm.gu.se/media/kursplaner/gym/kursplangymA-E94.pdf

Swedish National Agency for Education. (2000). *Matematik [Mathematics] (in Swedish)*. Nationellt Centrum för Matematikutbildning [National Center for Mathematics Education]. Retrieved November 20, 2012, from http://ncm.gu.se/media/kursplaner/gym/Gym2000.pdf

Swedish National Agency for Education. (2012). *Matematik [Mathematics] (in Swedish)*. Retrieved January 10, 2013, from https://www.skolverket.se/download/18.4fc05a3f164131a74181063/1535372298267/Mathematics-swedish-school.pdf

Swedish National Agency for Education. (2021). *Ämne - Matematik [Subject - Mathematics] (In Swedish)*. Retrieved February 2, 2022, from https://www.skolverket.se/undervisning/gymnasieskolan/laroplan-program-och-amnen-i-gymnasieskolan/gymnasieprogrammen/amne?url=-996270488%2Fsyllabuscw%2Fjsp%2Fsubject.htm%3FsubjectCode%3DMAT%26version%3D11%26tos%3Dgy&sv.url=12.5dfee44715d35a5cdfa92a3

Swedish National Board of Education [Skolöverstyrelsen]. (1965). *Löroplan för gymnasiet [Curriculum for Upper secondary school](in Swedish)*. Nationellt Centrum för Matematikutbildningen [Nationel Center for Mathematics Education] Retrieved November 20, 2012, from http://ncm.gu.se/media/kursplaner/gym/LLgym1965.pdf

Swedish Research Council. (2023). *Conducting Ethical Research*. Retrieved February 3, 2023, from https://www.vr.se/english/applying-for-funding/requirements-terms-and-conditions/conducting-ethical-research.html

Uppsala University. (2023). *CODEX rules and guidelines for research*. Retrieved February 3, 2023, from https://www.codex.uu.se/?languageId=1

Watt, H. (2005). Attitudes to the Use of Alternative Assessment Methods in Mathematics: A Study with Secondary Mathematics Teachers in Sydney. *Australia. Educational Studies in Mathematics, 58*(1), 21–44.

Weinert, F. E. (1999). *Concepts of Competence. OECD Project Definitions and Selection of Competencies: Theoretical and Conceptual Foundations (DeSeCo)*. Neuchâtel.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, *49*(8), 26–33.

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