FRAMEWORK
FOR CLASSROOM ASSESSMENT
IN MATHEMATICS

CONTENTS

Jan de Lange

Freudenthal Institute
&
National Center for Improving Student Learning
and Achievement in Mathematics and Science

September 1999

FRAMEWORK FOR CLASSROOM ASSESSMENT IN MATHEMATICS

This document is not *the* framework for classroom assessment in mathematics. One might even argue that this is not *a* framework. There have been several efforts to design and describe frameworks in assessment or, more specifically, in mathematics assessment. We mention several "framework" publications:

- Third International Mathematics and Science Study's (TIMSS) monograph, *Curriculum Frameworks for Mathematics and Science* (Robitaille et al., 1993).
- *Measuring Student Knowledge and Skills: A New Framework for Assessment* (Organization for Economic Cooperation and Development [OECD], 1999).
- "A Framework for Reflecting on Assessment and Evaluation" (Aikenhead, 1997).
- "A Framework for Developing Cognitively Diagnostic Assessments" (Nichols, 1994).
- "A Framework for Authentic Assessment in Mathematics" (Lajoie, 1991).
- "Toward an Assessment Framework for School Mathematics" (Goldin, 1992).

Goldin's title holds for all frameworks in the sense that we are continuously on the way *toward* a framework. In particular, it holds for the present one. This framework is the result of some 20 years of developmental research on classroom assessment practices. These experiences made clear how important and neglected classroom assessment is—in the U.S. even more than in most other countries because of the emphasis in the U.S. on standardized tests. A most timely overview of the research literature in classroom assessment by Black and Wiliam (1998) made our task in some ways more complex but also easier.

We have deliberately chosen to connect our framework with the OECD (1999) framework, designed for the Program of International Student Assessment (PISA)—not only because it reflects our philosophy reasonably well, but also because we need to connect internal and external assessment frameworks as much as we can. The framework presented here is under continuous development. As a product of the National Center for Improving Student Learning and Achievement in Mathematics and Science (NCISLA), it tries to incorporate examples and practices that relate to the theme chosen by the Center: Learning for Understanding. This theme certainly holds for the researchers at the Center: As we make progress, we will learn, our understanding of classroom assessment will improve over time, and so will this framework.

The structure of the framework is clear: We first discuss our philosophy, resulting in principles. Then we discuss what we consider important in mathematics education: the mathematical literacy and the organization of the mathematical content. The mathematical competencies that are needed can be categorized into three "levels" and the mathematical concepts into strands or "big ideas." We then discuss the whole array of formats and tools that are available for classroom assessment. Feedback and scoring are discussed before finally discussing more practical realizations of such a framework into the classroom. The *Great Assessment Picture Book* for mathematics (Mathematics GAP Book; Dekker & Querelle, in press) supports this framework, illustrating many of its ideas and concepts.

Aims

The aim of classroom assessment is to produce information that contributes to the teaching and learning process and assists in educational decision making, where decision makers include students, teachers, parents, and administrators.

The aim of mathematics education is to help students become mathematically literate. This means that the individual can deal with the mathematics involved in real world problems (i.e. nature, society, culture—including mathematics) as needed for that individual's current and future private life (as an intelligent citizen) and occupational life (future study or work) and that the individual understands and appreciates mathematics as a scientific discipline.

The aim of a framework for classroom assessment in mathematics is to bring the aim of classroom assessment together with the aim of mathematics education in a seamless and coherent way, with optimal results for the teaching and learning process, and with concrete suggestions about how to carry out classroom assessment in the classroom situation.

Principles

At the turn of the Century, an incredible number of changes were taking place in mathematics education, although not necessarily in the same direction. As Black and Wiliam (1998) correctly observe, the sum of all these reforms has not added up to an effective policy because something is missing: direct help with the teacher's task of managing complicated and demanding situations and channeling the personal, emotional, and social pressures of a group of 30 or more youngsters in order to help them learn and make them even better learners in the future.

Teachers need to know about their students' problems while learning, their progress, and the level of formality they are operating at so that they can adapt their teaching strategies to meet the pupils' needs. A teacher can find this information out in a variety of ways that range from observations and discussions to multi-step tasks and projects, from self-assessment and homework to oral presentations.

When the results of those activities are used in this way—to adapt the teaching and learning practice—we speak of *formative classroom assessment*.

A fundamental component of this feedback process is imparting information to students, assessing and evaluating the students' understanding of this information, and then matching the next teaching and learning action to the present understandings of the students (Hattie & Jaeger, 1998).

Some identify classroom assessment with formative assessment. We agree with Biggs (1998) that formative assessment and summative assessment are not mutually exclusive, as suggested by Black and Wiliam (1998). Their argument is that feedback concerning the gap between what *is* and what *should be* is regarded as formative only when comparison of actual and reference levels yields information that is then used to alter the gap. But if the information cannot lead to appropriate action, then it is not formative. .

Summative assessment in the form of end-of-year tests gives teachers the proof of how well they handled the formative assessment, assuming that the underlying philosophy is coherent and consequent. The differences in formative and summative assessment within the classroom are more related to timing and the amount of cumulation than anything else. Needed for both, of course, is that the assessment is criterion-referenced, incorporating the curriculum and resulting in aligned assessment.

The principle that the first and main purpose of testing is to improve learning (Gronlund, 1968; de Lange 1987) is widely and easily underestimated in the teaching and learning process. The reasons are multiple (e.g., the design of fair, rich, open and creative tasks is very difficult; the way the feedback mechanism operates; the organization and logistics of an opportunity-rich classroom). But Black and Wiliam's 1998 literature review on classrooms, *Assessment and Classroom Learning*, states very clearly that improvement in classroom assessment will make a strong contribution to the improvement of learning. So there is a strong need for a framework that takes this principle as its starting point.

But other principles and standards need to be considered, too. Standards published by the National Council of Teachers of Mathematics (NCTM, 1989) had a great influence in the discussion on reform in the U.S., and the NCTM recognized that “assessment standards” were needed as well (NCTM, 1995). But Standards will not be enough: “A focus on Standards and accountability that ignores the processes of teaching and learning in classrooms will not provide the directions that teachers need in their quest to improve” (Schmidt, McKnight, & Raizen, 1996). Nevertheless the NCTM Assessment Standards offer an excellent starting point for a discussion on principles and standards in classroom assessment. The Standards are about (a) the mathematics, (b) the learning of mathematics, (c) equity and opportunity, (d) openness, (e) inferences, and (f) coherence. The following sections discuss each of these standards in turn.

Standard 1. Mathematics

Few would argue with the assertion that useful mathematics assessments must focus on important mathematics. Yet the trend toward broader conceptions of mathematics and mathematical abilities raises serious questions about the appropriateness of the mathematics reflected in most traditional tests because that mathematics is generally far removed from the mathematics actually used in real-world problem solving. Nevertheless, there is still much debate over how to define important mathematics and who should be responsible for doing so.

This, of course, is a key issue. School mathematics is defined by long traditions resulting in a set of separate and often disconnected sub-areas that have little relation with the phenomenology of mathematics. Not only is that subdivision in strands rather arbitrary, but the timing of each of them in the learning process is also without any reasonable argument. Furthermore, we do not attempt to give a full picture of mathematics by any standard, but there is no discussion about which subject in school mathematics should be covered: for example, take the long discussion and the slow progress on the introduction of discrete mathematics in school curricula. Traditional assessment practices have emphasized this compartmentalization of school mathematics. Common features of teachers’ formative assessment focuses on superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge that students soon forget (Crooks, 1988, and Black, 1993, as summarized by Black and Wiliam, 1998). It is for this reason that we have chosen to focus on “big ideas” in mathematics (a cluster of related fundamental mathematical concepts ignoring the school curricula compartmentalization) and that we try to assess broader mathematical ideas and processes.

Standard 2. Learning

New views of assessment call for tasks that are embedded in the curriculum, the notion being that assessment should be an integral part of the learning process rather than an interruption of it. This raises the issue of who should be responsible for the development, implementation, and interpretation of student assessments. Traditionally both standardized and classroom tests were designed using a psychometric model to be as objective as possible. By contrast, the alternative assessment movement affords teachers much more responsibility and subjectivity in the assessment process. It assumes that teachers know their students best because teachers have multiple, diverse opportunities for examining student work performed under various conditions and presented in a variety of modes. When teachers have more responsibility for assessment, assessment can truly become almost seamless with instruction (Lesh & Lamon, 1992).

It will be clear from our introduction that we see classroom assessment as an integral part of the teaching and learning process, there should be a mutual influence. It is actually so trivial that one is surprised to see that the actual practice is so different. The main cause for this situation is the standardized test system. The ironic and unfortunate result of this system is that teachers resist formal evaluation of all kinds, given the intellectual sterility and rigidity of most generic, indirect, and external testing systems. But because of that resistance, local assessment practices are increasingly unable to withstand technical scrutiny: Teacher tests are rarely valid and reliable, and "assessment" is reduced to averaging scores out (Wiggins, 1993). Biggs (1998) blames psychometricians who, although through no fault of their own, have done enough damage to educational assessment. The result is that in most classrooms assessment is no longer a part of the teaching and learning process.

We should and will try, by means of this Framework, to offer teachers a wide array of instruments and opportunities for examining work performed under various conditions. Teachers need to be aware about the connections between the tests tools and the curricular goals and how to generate relevant feedback from the test results.

Standard 3. Equity and Opportunity

Ideally, assessments should give every student optimal opportunity to demonstrate mathematical power. In practice, however, traditional standardized tests have sometimes been biased against students of particular backgrounds, socioeconomic classes, ethnic groups, or gender (Pullin, 1993). Equity becomes even more of an issue when assessment results are used to label students or deny them access to courses, programs, or jobs. More teacher responsibility means more pressure on teachers to be evenhanded and unbiased in their judgment. Ironically, the trend toward more complex and realistic assessment tasks and more elaborated written responses can raise serious equity concerns because reading comprehension, writing ability, and familiarity with contexts may confound results for certain groups (Lane, 1993).

Clearly, teachers have a very complex task here. As Cobb et al. (1991) argued, we do not assess a person objectively, but we assess how a person acts in a certain setting. Certain formats favor boys more than girls, others are more equal; boys do better under time pressure than girls (de Lange, 1987); girls seem to fare better when there is more language involved; certain contexts are more suited for boys, others for girls (van den Heuvel-Panhuizen & Vermeer, 1999); and cultural differences should be taken into account. For these reasons, we discuss the role of context in some detail, the effect of and the need to use different formats, and the need for a variety of representations. For similar reasons, we advocate the assignment of both individual and group work as well as the use of both time-restricted and unrestricted assessments. Only if we offer that wide variety do we have a chance at "fair" classroom assessment.

Standard 4. Openness

Testing has traditionally been quite a secretive process, in that test questions and answers were carefully guarded, and criteria for judging performance were generally set behind the scenes by unidentified authorities. By contrast, many today believe that students are best served by open and dynamic assessment—assessment where expectations and scoring procedures are openly discussed and jointly negotiated.

Students need to know what the teachers expect from them, how their work will be scored and graded, what a 'good explanation' looks like, etcetera. Teachers should have examples of all the different tests that are possible or to be expected, with scoring rubrics and possible student work. They need to know why these tests are given, and what will be done with the results. Again tradition and existing practice have done much damage. Secrecy was a key issue when testing—secrecy as to the questions being asked, how the questions would be chosen, how the results would be scored, what the scores mean, and how the results would be used (Wiggins, 1993). According to Schwarz (1992), standardized tests can be given on a wide scale only if secrecy can be maintained because this testing technology requires a very large number of questions that are expensive and difficult to generate. Yet according to Schwarz, this is an undesirable situation. He proposes new approaches to the filing, indexing, and retrieving of previously used problems. Publicly available, richly indexed databases of problems and projects provide opportunity for scrutiny, discussion, and debate about the quality and correctness of questions and answers. It seems that we have a long way to go, but openness and clarity are prerequisites for any proper classroom assessment system.

Standard 5. Inferences

Changes in assessment have resulted in new ways of thinking about reliability and validity as they apply to mathematics assessment. For example, when assessment is embedded within instruction, it becomes unreasonable to expect a standard notion of reliability to apply (that a student's achievement on similar tasks at different points in time should be similar) because it is actually expected that students will learn throughout the assessment. Similarly, new forms of assessment prompt a re-examination of traditional notions of validity. Many argue that it is more appropriate to judge validity by examining the inferences made from an assessment than to view it as an inherent characteristic of the assessment itself. Nevertheless, it is difficult to know how new types of assessment (e.g., student projects or portfolios) can be used for decision making without either collapsing them into a single score (thereby losing all of their conceptual richness) or leaving them in their raw, unsimplified, and difficult-to-interpret form.

Reliability and validity are concepts from an era when psychometricians made the rules. These terms have taken on a specific and narrow meaning, have caused much damage to the students and society, and more specifically have skewed the perception of what constitutes good school mathematics. More important, especially in classroom assessment, is authenticity of the tasks (i.e., performance faithful to criterion situations). "Authentic" means that the problems are "worthy" and relate to the real world, are non-routine, have "construction" possibilities for students, relate to clear criteria, ask for explanations of strategies, and offer possibilities to discuss grading.

In order to do justice to the students (which entails freedom from distortion and letting the object speak [Smaling, 1992]) and add validity in the traditional sense, we need a sample of authentic tasks to get a valid picture. And, indeed, reliability in the traditional sense is something to be avoided at all times if we really want assessment as part of the teaching and learning process. If we offer the students the same tests at different moments, we should note differences in levels of formality, different strategies, even different answers in some cases. If the tests yield the same results (and thus are reliable), then our teaching has failed. It is exactly for this reason that a longitudinal study on the effects of a new middle school curriculum has four different operationalizations of the “same” problem to find out about students’ growth over time in Grades 5, 6, 7, and 8 (Shafer & Romberg, 1999).

Smaling (1992) defined “reliability” in a more ecological way: Reliability refers to the absence of accidental errors and is often defined as reproducibility. But here it means virtual replicability. The emphasis is on “virtual,” because it is important that the result be reported in such a way that others can reconstruct it. The meaning of this is aptly expressed by the term “trackability” which, according to Gravemeijer (1994), is highly compatible with Freudenthal’s conception of developmental research because “trackability” can be established by reporting on “failures and successes,” the procedures followed, the conceptual framework, and the reasons for the choices made.

Standard 6. Coherence

The coherence standard emphasizes the importance of ensuring that each assessment is appropriate for the purpose for which it is used. As noted earlier, assessment data can be used for monitoring student progress, making instructional decisions, evaluating achievement, or program evaluation. The types of data appropriate for each purpose, however, may be very different. Policymakers and assessment experts often disagree on this issue. Policymakers may have multiple agendas in mind and expect that they can all be accomplished by using a single assessment while assessment experts warn against using an assessment for purposes for which it was never intended.

Coherence in classroom assessment can be accomplished quite simply if the teaching and learning process is coherent and the assessment is an integral part of it. Teachers have a wide variety of techniques and tools at their disposal to “design” their own classroom assessment system that fits with the didactical contract they have with the classroom. Depending on their teaching and learning practice and style, they will present the students with their “balance” within the classroom assessment system. Coherence with colleagues will be achieved by sharing the same criteria and possibly by designing and discussing common tasks and tests. Together with designing and using the same “end-of-year test” for students in the same grade, “fairness” for all students in the same year and over the years is ensured because the end-of-year tests are not secret although they change over the years.

Coherence in relation to external assessment is also essential. For this reason this framework is somewhat related to the recently published framework for mathematics (OECD, 1999) which is being used in a comparative international Assessment Study. Several key components of this framework and OECD's framework are aligned in order to ensure more coherence between classroom assessment and a very visible form of external assessment.

Reflecting on these standards and the existing literature, we make the following list of principles for classroom assessment.

Principles for Classroom Assessment

1. The main purpose of classroom assessment is to improve learning (Gronlund, 1968; de Lange, 1987; Black & Wiliam, 1998; and many others).
2. The mathematics is embedded in worthwhile (engaging, educative, authentic) problems that are part of the students' real world.
3. Methods of assessment should be such that they enable students to reveal what they know, rather than what they do not know (Cockroft, 1982).
4. A balanced assessment plan should include multiple and varied opportunities (formats) for students to display and document their achievements (Wiggins, 1992).
5. Tasks should operationalize all the goals of the curricula (not just the "lower" ones). Helpful tools to achieve this are performance standards, including indications of the different levels of mathematical thinking (de Lange, 1987).
6. Grading criteria should be public and consistently applied; and should include examples of earlier grading showing exemplary work and work that is less than exemplary.
7. The assessment process, including scoring and and grading, should be open to students.
8. Students should have opportunities to receive genuine feedback on their work.
9. The quality of a task is not defined by its accessibility to objective scoring, reliability, or validity in the traditional sense but by its authenticity, fairness, and the extent to which it meets the above principles (de Lange, 1987).

These principles form a "checklist" for teachers who take their classroom assessment seriously. But the journey from principles to practice can be long. So we will now turn to a discussion about several key issues in designing and implementing a classroom assessment system.

In the list of principles, the content was mentioned in different ways (relevant, real-world mathematics) and at several levels of mathematical thinking and reasoning because our goal for mathematics education is to enable individuals to deal with the mathematics involved in real-world problems. This is needed for each individual's current and future private life, occupational life (work or education), and understanding and appreciation of mathematics as a scientific discipline. In other words: We want our students to become mathematically literate. So first, we will elaborate on mathematical literacy. This definition is based on the one used in the OECD's framework for mathematics (OECD, 1999), which draws heavily on the work of Niss and others from the mathematics functional expert group for the same study.[1]

Mathematical Literacy

"Mathematical literacy" is an individual's ability to identify, understand, exert well-founded judgment about, and act toward the roles that mathematics plays in dealing with the world (i.e. nature, society, and culture)—not only as needed for that individual's current and future private life, occupational life, and social life with peers and relatives but also for that individual's life as a constructive, concerned, and reflective citizen.

Some explanatory remarks are in order for this definition to become transparent.

1. In using the term "literacy," we want to emphasize that mathematical knowledge and skills that have been defined and are definable within the context of a mathematics curriculum do *not* constitute our primary focus here. Instead, what we have in mind is mathematical knowledge put into functional use in a multitude of contexts in varied, reflective, and insight-based ways. Of course for such use to be possible and viable, a great deal of intra-curricular knowledge and skills are needed. Literacy in the *linguistic* sense cannot be reduced to—but certainly presupposes—a rich vocabulary and a substantial knowledge of grammatical rules, phonetics, orthography, and so forth. In the same way, *mathematical* literacy cannot be reduced to—but certainly presupposes—knowledge of mathematical terminology, facts, and procedures as well as numerous skills in performing certain operations, carrying out certain methods, and so forth. Also, we want to emphasize that the term "literacy" is not confined to indicating a basic, minimum level of functionality only. On the contrary, we think of literacy as a continuous, multidimensional spectrum ranging from aspects of basic functionality to high-level mastery. In the same vein when we use the word "needed" we do not restrict ourselves to

what might be thought of as a minimum requirement for coping with life in the spheres that are at issue. We also include what is “helpful,” “worthwhile,” or “desirable” for that endeavor.

2. The term “act” is not meant to cover only physical or social acts in a narrow sense. Thus the term includes also “communicating,” “taking positions toward,” “relating to,” and even “appreciating” or “assessing.”
3. A crucial capacity implied by our notion of mathematical literacy is the ability to pose, formulate and solve intra- and extra-mathematical problems within a variety of domains and settings. These range from purely mathematical ones to ones in which no mathematical structure is present from the outset but may be successfully introduced by the problem poser, problem solver, or both.
4. Attitudes and emotions (e.g., self-confidence, curiosity, feelings of interest and relevance, desire to do or understand things) are not components of the definition of mathematical literacy. Nevertheless they are important prerequisites for it. In principle it is possible to possess mathematical literacy without possessing such attitudes and emotions at the same time. In practice, however, it is not likely that such literacy will be exerted and put into practice by someone who does not have some degree of self-confidence, curiosity, feeling of interest and relevance, and desire to do or understand things that contain mathematical components.

Mathematical Competencies

Again, in defining Mathematical Competencies we follow the Mathematics Literacy framework published by the OECD Program for International Student Assessment (PISA). Here is a nonhierarchical list of general mathematical competencies that are meant to be relevant and pertinent to all education levels.

- **Mathematical thinking**
  - ♦ Posing questions characteristic of mathematics—Does there exist...? If so, how many? How do we find...?
  - ♦ Knowing the kinds of answers that mathematics offers to such questions.
  - ♦ Distinguishing between different kinds of statements (e.g., definitions, theorems, conjectures, hypotheses, examples, conditioned assertions).
  - ♦ Understanding and handling the extent and limits of given mathematical concepts.

- **Mathematical argumentation**
    - ♦ Knowing what mathematical proof is and how it differs from other kinds of mathematical reasoning.
    - ♦ Following and assessing chains of mathematical arguments of different types.
    - ♦ Possessing a feel for heuristics (what can happen, what cannot happen, and why).
    - ♦ Creating mathematical arguments.
- **Modelling**
    - ♦ Structuring the field or situation to be modelled
    - ♦ Mathematizing (i.e., translating from "reality" to "mathematics").
    - ♦ De-mathematizing (i.e., interpreting mathematical models in terms of "reality").
    - ♦ Tackling the model (working within the mathematics domain).
    - ♦ Validating the model.
    - ♦ Reflecting, analyzing, offering critique of models and model results.
    - ♦ Communicating about the model and its results (including the limitations of such results).
    - ♦ Monitoring and control of the modelling process.
- **Problem posing and solving**
    - ♦ Posing, formulating, and making precise different kinds of mathematical problems (e.g., pure, applied, open-ended, closed).
    - ♦ Solving different kinds of mathematical problems in a variety of ways.
- **Representation**
    - ♦ Decoding, interpreting, and distinguishing between different forms of presentations of mathematical objects and situations, and the interrelations between the various representations.
    - ♦ Choosing and switching between different forms of representation according to situation and purpose.
- **Symbols and formal language**
    - ♦ Decoding and interpreting symbolic and formal language and understanding its relations to natural language.
    - ♦ Translating from natural language to symbolic or formal language.
    - ♦ Handling statements and expressions that contain symbols and formulas.

  - ♦ Using variables, solving equations, and performing calculations.
- **Communication**
  - ♦ Expressing oneself in a variety of ways on matters with mathematical components, in oral as well as in written form.
  - ♦ Understanding others' written or oral statements about such matters.
- **Aids and tools**
  - ♦ Knowing about and being able to make use of various aids and tools (including information technology tools) that may assist mathematical activity.
  - ♦ Knowing about the limitations of such aids and tools.

Competence Levels

We do not propose development of test items that assess the above skills individually. When doing real mathematics, it is necessary to draw simultaneously upon many of those skills. In order to operationalize these mathematical competencies, it is helpful to organize the skills into three levels. They were successfully operationalized in the National Dutch option of TIMSS (Boertien & de Lange, 1994; Kuiper, Bos, & Plomp, 1997) and the ongoing longitudinal study on the effects of a middle-school curriculum and have also been adapted for the OECD study.

The three levels are—

1. Reproduction, definitions, computations.
2. Connections and integration for problem solving.
3. Mathematization, mathematical thinking, generalization, and insight.

We will elaborate on these levels next.

Level 1. Reproduction, procedures, concepts, and definitions

At this first level, we deal with the matter dealt with in many standardized tests, as well in comparative international studies, and operationalized mainly in multiple-choice format. In TIMSS, the performance expectation aspects of *knowing* and *using routine procedures* would fit this level. It deals with knowledge of facts, representing, recognizing equivalents, recalling mathematical objects and properties, performing routine procedures, applying standard algorithms, and developing technical skills. Dealing and operating with statements and expressions that contain symbols and formulas in "standard" form also relate to this level.

Items at Level 1 are often in multiple-choice, fill-in-the-blank, matching, or (restricted) open-ended questions format.

Level 2. Connections and integration for problem solving

At this level we start making connections between the different strands and domains in mathematics and integrate information in order to solve simple problems in which students have a choice of strategies and a choice in their use of mathematical tools. Although the problems are supposedly nonroutine, they require relatively minor mathematization. Students at this level are also expected to handle different forms of representation according to situation and purpose. The connections aspect requires students to be able to distinguish and relate different statements such as definitions, claims, examples, conditioned assertions, and proof.

From the point of view of mathematical language, another aspect at this level is decoding and interpreting symbolic and formal language and understanding its relations to natural language. This level relates somewhat to the TIMSS *investigating and problem-solving* category, which included formulating and clarifying problems and situations, developing strategy, solving, predicting, and verifying. Judging by these items, however, one has to bear in mind that *problem solving* and *using complex procedures* in TIMSS are competencies that are actually very close to those in our proposed Level 1. Examples therefore play an important role in making our levels of competencies and skills clear and workable.

Items at Level 2 are often placed within a context and engage students in mathematical decision making.

Level 3. Mathematization, mathematical thinking, generalization, and insight

At Level 3, students are asked to mathematize situations (recognize and extract the mathematics embedded in the situation and use mathematics to solve the problem). They must analyze, interpret, develop their own models and strategies, and make mathematical arguments including proofs and generalizations. These competencies include a critical component and analysis of the model and reflection on the process. Students should not only be able to solve problems but also to pose problems.

These competencies function well only if the students are able to communicate properly in different ways (e.g., orally, in written form, using visualizations). Communication is meant to be a two-way process: students should also be able to understand communication with a mathematical component by others. Finally we would like to stress that students also need insight competencies—insight into the nature of mathematics as a science (including the cultural and historical aspect) and understanding of the use of mathematics in other subjects as brought about through mathematical modeling.

As is evident, the competencies at Level 3 quite often incorporate skills and competencies usually associated with the other two levels. We note that the whole exercise of defining the three levels is a somewhat arbitrary activity: There is no clear distinction between different levels, and both higher- and lower-level skills and competencies often play out at different levels.

In the TIMSS framework, Level 3 relates best to the *mathematical reasoning* performance expectation: developing notation and vocabulary, developing algorithms, generalizing, and conjecturing.

Level 3, which goes to the heart of mathematics and mathematical literacy, is difficult to test. Multiple-choice is definitely not the format of choice at Level 3. Extended response questions with multiple answers (with [super-] items or without increasing level of complexity) are more likely to be promising formats. But both the design and the judgment of student answers are very, if not extremely, difficult. Because Level 3 is at the heart of our study, however, we should try, as much as practice permits, to operationalize these competencies in appropriate test items.

The three levels can be visually represented in a pyramid (Figure 1; de Lange, 1995). This pyramid has three dimensions or aspects: (a) the content or domains of mathematics, (b) the three levels of mathematical thinking and understanding (along the lines just defined), and (c) the level of difficulty of the questions posed (ranging from simple to complex). The dimensions are not meant to be orthogonal, and the pyramid is meant to give a fair visual image of the relative number of items required to represent a student’s understanding of mathematics. Because we need only simple items for the lower levels, we can use more of them in a short amount of time. For the higher levels we need only a few items because it will take some time for the students to solve the problems at this level.

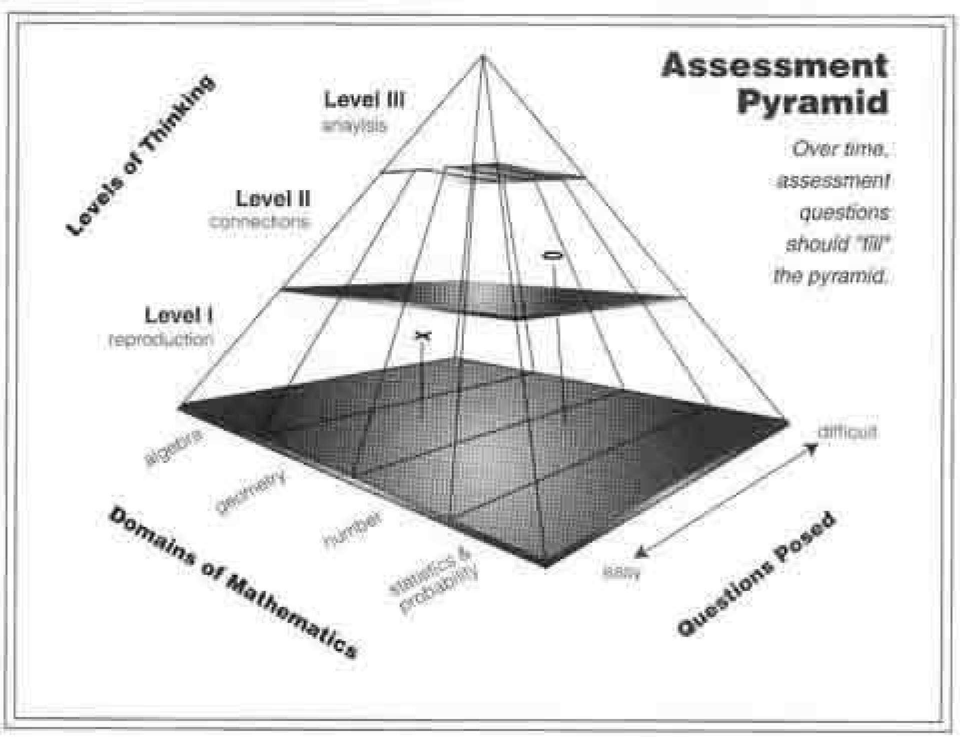


Figure 1. Assessment pyramid

The *easy* to *difficult* dimension can be interchanged with a dimension that ranges from *informal* to *formal*.

All assessment questions can be located in the pyramid according to (a) the level of thinking called for, (b) mathematical content or big ideas domain, and (c) degree of difficulty. Because assessment needs to measure and describe a student's growth in all domains of mathematics and at all three levels of thinking, questions in a complete assessment program should fill the pyramid. There should be questions at all levels of thinking, of varying degrees of difficulty, and in all content domains.

Essential to mathematical literacy is the ability to mathematize a problem. This process of mathematization will therefore be described in a little more detail:

Defining mathematization. Mathematization, as it is being dealt with here, is organizing reality using mathematical ideas and concepts. It is the organizing activity according to which students used acquired knowledge and skills to discover unknown regularities, relations and structures (Treffers & Goffree, 1985). This process is sometimes called horizontal mathematization (Treffers, 1987) and requires activities such as—

- Identifying the specific mathematics in a general context.
- Schematizing.
- Formulating and visualizing the problem.
- Discovering relations and regularities.
- Recognizing similarities in different problems (de Lange, 1987).

As soon as the problem has been transformed to a more-or-less mathematical problem, it can be attacked and treated with mathematical tools. That is, mathematical tools can be applied to manipulate and refine the mathematically modeled real-world problem. This is the process of vertical mathematization and can be recognised in the following activities:

- Representing a relation in a formula.
- Proving regularities.
- Refining and adjusting models.
- Combining and integrating models.
- Generalizing.

Thus the process of mathematization plays out in two different phases. The first is horizontal mathematization, the process of going from the real world to the mathematical world. The second, vertical mathematization is working on the problem within the mathematical world (developing mathematical tools in order to solve the problem). Reflecting on the solution with respect to the original problem is an essential step in the process of mathematization that quite often does not receive proper attention.

One can argue that mathematization plays out in all competency classes because in any contextualized problem one has to identify the relevant mathematics. The varying complexity of mathematization is reflected in the two examples below. Both are meant for students of 13–15 years of age and both draw upon similar mathematical concepts. The first requires simple mathematization whereas the second requires more complex mathematization.

Example 1. (Level 2) A class has 28 students. The ratio of girls to boys is 4:3.

How many girls are in the class?

Source: TIMSS Mathematics Achievement in the Middle Years, p.98

Example 2. (Level 3) In a certain country, the national defence budget is $30 million for 1980. The total budget for that year is $500 million. The following year the defence budget is

$35 million, while the total budget is $605 million. Inflation during the period covered by the two budgets amounted to 10 percent.

a. You are invited to give a lecture for a pacifist society. You intend to explain that the defence budget decreased over this period. Explain how you could do this.
b. You are invited to lecture to a military academy. You intend to explain that the defence budget increased over this period. Explain how you would do this.

Source: de Lange (1987)

The Mathematics: Strands and Big Ideas

Mathematics school curricula are organized into strands that classify mathematics as a strictly compartmentalized discipline with an over-emphasis on computation and formulas. This organization makes it almost impossible for students to see mathematics as a continuously growing scientific field that continually spreads into new fields and applications. Students are not positioned to see overarching concepts and relations, so mathematics appears to be a collection of fragmented pieces of factual knowledge.

Steen (1990) puts it somewhat differently: School mathematics picks very few strands (e.g., arithmetic, algebra, geometry) and arranges them horizontally to form a curriculum. First is arithmetic, then simple algebra, then geometry, then more algebra, and finally—as if it where the epitome of mathematical knowledge—calculus. This layer-cake approach to mathematics education effectively prevents informal development of intuition along the multiple roots of mathematics. Moreover, it reinforces the tendency to design each course primarily to meet the prerequisites of the next course, making the study of mathematics largely an exercise in delayed gratification.

“What is mathematics?” is not a simple question to answer. A person asked at random will most likely answer, “Mathematics is the study of Number.” Or, if you’re lucky, “Mathematics is the science of number.” And, as Devlin (1994) states in his very successful book, “Mathematics: The Science of Patterns,” the former is a huge misconception based on a description of mathematics that ceased to be accurate some 2,500 years ago. Present-day mathematics is a thriving, worldwide activity, it is an essential tool for many other domains like banking, engineering, manufacturing, medicine, social science, and physics. The explosion of mathematical activity that has taken place in the twentieth century has been dramatic. At the turn of the nineteenth century, mathematics could reasonably be regarded as consisting of about 12 distinct subjects: arithmetic, geometry, algebra, calculus, topology and so on. The similarity between this list and the present-day school curricula list is amazing.

A more reasonable figure for today, however, would be between 60 and 70 distinct subjects. Some subjects (e.g., algebra, topology) have split into various subfields; others (e.g., complexity theory, dynamical systems theory) are completely new areas of study.

In our list of principles, we mentioned content: Mathematics should be relevant, meaning that mathematics should be seen as the language that describes patterns—both patterns in nature and patterns invented by the human mind. Those patterns can be either real or imagined, visual or mental, static or dynamic, qualitative or quantitative, purely utilitarian or of little more than recreational interest. They can arise from the world around us, from depth of space and time, or from the inner workings of the human mind (Devlin, 1994). For this reason, we have not chosen traditional content strands as the major dimensions for describing content. Instead we have chosen to organize the content of the relevant mathematics around "big ideas" or "themes."

The concept of big ideas is not new. In 1990, the Mathematical Sciences Education Board published *On the Shoulders of Giants: New Approaches to Numeracy* (Steen, 1990), a book that made a strong plea for educators to help students delve deeper to find the concepts that underlie all mathematics and thereby better understand the significance of these concepts in the world. To accomplish this, we need to explore ideas with deep roots in the mathematical sciences without concern for the limitations of present schools of curricula.

Many big ideas can be identified and described. In fact the domain of mathematics is so rich and varied that it would not be possible to identify an exhaustive list of big ideas. It is important for purposes of classroom assessment, however, for any selection of big ideas that is offered to represent a sufficient variety and depth to reveal the essentials of mathematics and their relations to the traditional strands.

The following list of mathematical big ideas meets this requirement:

- Change and growth.
- Space and shape.
- Quantitative reasoning.
- Uncertainty.

Change and Growth

Every natural phenomenon is a manifestation of change. Some examples are organisms changing as they grow, the cycle of seasons, the ebb and flow of tides, cycles for unemployment, weather changes, and the Dow-Jones index. Some of these growth processes can be described or modeled by some rather straightforward mathematical functions (e.g., linear, exponential, periodic, logistic, either discrete or continuous). But many processes fall into different categories, and data analysis is quite often essential. The use of computer technology has resulted in more powerful approximation techniques, and more sophisticated visualization of data. The patterns of change in nature and in mathematics do not in any sense follow the traditional content strands.

To be sensitive to the patterns of change, we follow Stewart (1990), who states that we need to—

- Represent changes in a comprehensible form.
- Understand the fundamental types of change.
- Recognise particular types of changes when they occur.
- Apply these techniques to the outside world.
- Control a changing universe to our best advantage.

These competencies relate nicely to both our definition of mathematical literacy and the competencies as defined earlier in this framework.

Many different sub-strands of traditional content strands emerge in this main mathematical domain of change and growth. The obvious ones are relations, functions and their graphical representations. Series and gradients are also heavily intertwined with functions. Considering rates of growth for different growth phenomena leads to linear, exponential, logarithmic, periodic, logistic growth curves, and their properties and relations. These, in turn, lead to aspects of number theory, such as Fibonacci-numbers and the Golden-ratio. The connections between these ideas and geometrical representations can also play a role here.

In geometry, one can also explore patterns in nature, art or architecture. Similarity and congruence might play a role here, as would the growth of an area in relation to the growth of the perimeter or circumference.

Growth patterns can be expressed in algebraic forms, which in turn can be represented by graphs. Growth can also be measured empirically, and such questions arise as which inferences can be made from the growth data and how the growth data might be represented. Aspects from the data analysis and statistics content strands also naturally emerge here.

Space and Shape

Patterns are encountered not only in growth and change processes, but also they occur everywhere around us: in spoken words, music, video, traffic, constructions, and art. Shapes are patterns: houses, churches, bridges, starfish, snowflakes, city plans, cloverleaves, crystals, and shadows. Geometric patterns can serve as relatively simple models of many kinds of phenomena, and their study is possible and desirable at all levels (Grünbaum, 1985). Shape is a vital, growing, and fascinating theme in mathematics that has deep ties to traditional geometry (although relatively little in school geometry) but goes far beyond it in content, meaning, and method (Senechal, 1990).

In the study of shape and constructions, we are looking for similarities and differences as we analyse the components of form and recognize shapes in different representations and different dimensions. The study of shapes is closely knitted to “grasping space” (Freudenthal, 1973). That is learning to know, explore, and conquer in order to improve how we live, breathe, and move through the space in which we live.

This means that we must be able to understand relative positions of objects. We must be aware of how we see things and why we see them this way. We must learn to navigate through space and through constructions and shapes. It means that students should be able to understand the relation between shapes and images or visual representations (e.g., the relation between the real city and photographs or maps of the same city). They must also understand how three-dimensional objects can be represented in two dimensions, how shadows are formed and must be interpreted, and what “perspective” is and how it functions.

Described in this way, the study of Space and Shape is open-ended and dynamic, and it fits well into both mathematical literacy and mathematical competencies as defined for this framework..

Quantitative Reasoning

Karl Friedrich Gauss’ (1777–1855) teacher had asked the class to add together all the numbers from 1 to 100. Presumably the teacher’s aim was to keep the students occupied for a time. But Gauss was an excellent quantitative reasoner and identified a shortcut to the solution. His reasoning went like this:

You write down the sum twice—once in ascending order, then in descending order, like this:

$$1+2+3+\ldots+98+99+100$$

$$100+99+98+\ldots+3+2+1$$

Now you add the two sums, column by column, to give:

$$101+101+101+\ldots+101+101+101$$

As there are exactly 100 copies of the number 101 in this sum, its value is

$$11\times 101=10{,}100$$

This product is twice the answer to the original sum, so you can halve it to obtain the answer: 5,050.

Talking about patterns: we might elaborate a little further as the formula that gives the general situation for Gauss' problem looks like this:

$$1+2+3+\ldots+n=\frac{n(n+1)}{2}$$

This formula also captures a geometric pattern that is well known: Numbers of the form $\frac{n(n+1)}{2}$ are called triangular numbers because they are exactly the numbers you can obtain by arranging balls in an equilateral triangle. The first five triangular numbers—1, 3, 6, 10, and 15—are shown in Figure 2.

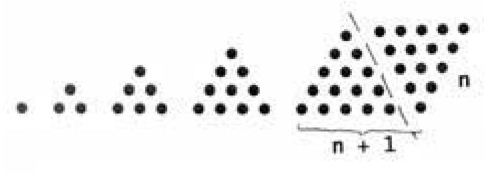


Figure 2. The first five triangular numbers (1, 3, 6, 10, and 15)

Quantitative reasoning is more than being excellent at reasoning in mathematical situations. It includes number sense: meaning of operations, feel for magnitude of numbers, smart computations, mental arithmetic, estimations. And it comes close to being mathematically literate if we accept a broader definition (Howden, 1989).

Given the fundamental role of quantitative reasoning in applications of mathematics, as well as the innate human attraction to numbers, it is not surprising that number concepts and skills form the core of school mathematics. In the earliest grade, we start children on a mathematical path designed to develop computational procedures of arithmetic together with the corresponding conceptual understanding that is required to solve quantitative problems and make informed decisions.

Quantitative literacy requires an ability to interpret numbers used to describe random as well as deterministic phenomena, to reason with complex sets of interrelated variables, and to devise and critically interpret methods for quantifying phenomena where no standard model exists.

Quantitatively literate students need a flexible ability to (a) identify critical relations in novel situations, (b) express those relations in effective symbolic form, (c) use computing tools to process information, and (d) interpret the results of these calculations (Fey, 1990). Here we border on the next big idea: uncertainty.

We also like to stress that in the seeming order of elementary school arithmetic, there is a place for quantitative reasoning similar to that of Gauss, as described earlier. Creativity, coupled with conceptual understanding, is often ignored at the elementary level of schooling. Students know how to execute a multiplication but have no idea how to answer when asked, “What is multiplication?” Students are very poorly educated in recognizing isomorphic problems, or problems that can be solved using the same mathematical tools. For example, they often fail to recognize that the following three problems can all be solved using the ratio table.

1) Tonight you’re giving a party. You want about a hundred cans of Coke. How many six-packs are you going to buy?
2) A hang glider with glide ratio of 1 to 23 starts from a sheer cliff at 123 meters. The pilot is aiming for a spot at a distance of 1,234 meters. Will she reach that spot?
3) A school wants to rent minivans (with 8 seats each) to transport 78 students to a school camp. How many vans will the school need?

Uncertainty

Uncertainty is intended to suggest two related topics: data and chance. Neither is a topic in mathematics but both are phenomena that are the subject of mathematical study. Rather recent recommendations concerning school curricula are unanimous in suggesting that statistics and probability should occupy a much more prominent place than has been the case in the past (Mathematical Sciences Education Board, 1990; NCTM, 1989). Because these recommendations emphasize data analysis, however, it is particularly easy to view statistics as a collection of specific skills. David S. Moore, the well-known statistics educator, pointed out for us what the big idea “uncertainty” is really all about. We follow his ideas as presented in *On the Shoulders of Giants* (Steen, 1990).

The ability to deal intelligently with variation and uncertainty is the goal of instruction about data and chance. Variation is a concept that is hard to deal with: Children who begin their education with spelling and multiplication expect the world to be deterministic. They learn quickly to expect one answer to be right and others to be wrong, at least when the answers take numerical form. Variation is unexpected and uncomfortable, as Arthur Nielsen (1987) from the famous marketing research firm noted:

> [Business people] accept numbers as representing Truth…. They do not see a number as a kind of shorthand for a range that describes our actual knowledge of the underlying condition.
>
> …I once decided that we would draw all charts to show a probable range around the number reported; for example, sales are either up 3 percent, or down 3 percent or somewhere in between. This turned out to be one of my dumber ideas. Our clients just couldn’t work with this type of uncertainty. (p. 8)

Statistical thinking involves reasoning from uncertain empirical data and should therefore be part of the mental equipment of every intelligent citizen. Its core elements are the—

- Omnipresence of variation in processes.
- Need for data about processes.
- Design of data production with variation in mind.
- Quantification of variation.
- Explanation of variation.

Data analysis might help the learning of basic mathematics: The essence of data analysis is to “let the data speak” by looking for patterns in data without first considering whether the data are representative of some larger universe.

Phenomena that have uncertain individual outcomes but a regular pattern of outcomes in many repetitions are called random. Psychologists have shown that our intuition of chance profoundly contradicts the laws of probability. In part, this is due to students’ limited contact with randomness. The study of data offers a natural setting for such an experience. This explains the priority of data analysis over formal probability. Inference should be an important principle in the learning and teaching of uncertainty.

Relationship With Traditional Strands

It will be clear from our introduction about the “big ideas” that (a) we will never be able to fully grab “mathematics” in such themes and (b) not only do other themes exist but these themes may be better, depending on one’s perspective. It also requires little imagination to relate the big ideas with the traditional strands. We also realize that relatively few school text materials depart from the big-ideas approach. The visualization of competency levels in the Pyramid reflects that dilemma. But we only do justice to the discipline of mathematics and to our students if we give a more honest picture of mathematics. That means that in our classroom assessment we need to strive for broader assessment items that do not necessarily fit with the traditional strands. That has repercussions for the classroom teaching and learning process because assessment should be aligned seamlessly. But if we accept the principle that the mathematics should be important and relevant, we need at least to embed the traditional strand in the big ideas. Once more, we stress the hypothesis of Black and Wiliam that classroom assessment might be the most powerful tool to change and improve mathematics education. It might even help us paint a much more vivid and dynamic picture of mathematics.

Methods for Classroom Assessment

When engaging in classroom assessment, the teacher is confronted with many tasks, choices, and dilemmas. How can we provoke Socratic dialogues that spur learning, and how can we assess these dialogues? (Note that even during such dialogues, Hans Freudenthal warned against students’ roles being too passive [Freudenthal, 1973].) How can we organize effectual interaction, and how can we assess the resulting effects? What kind of tasks lead to fruitful arguments and how can we value these arguments? How can we observe in a proper way and keep track of what is observed?

For many reasons, none of these questions have simple, easy-to-implement answers. The most obvious reason, however, seems to be that assessment is so interwoven with the learning and teaching. It is impossible to say where the learning ends and assessment begins. Another reason is that the sociocultural context plays a major role. There are no “general” rules. We can only give the teacher some information about classroom experiments and the results of the observation, quite often resulting in “local” theories.

We can offer somewhat more practical suggestions on the field of self- and peer assessment and even more when we address the more common assessment formats; their possibilities, qualities, and drawbacks; how to choose the appropriate format; and how to score the tasks. We have chosen as our principle that mathematics needs to be relevant, which quite often means that there needs to be a link to the real world; therefore, special attention should be given to the choice, role, and function of contexts in assessment. This aspect plays an important role in any form of assessment, so we will begin with a discussion on contexts.

Contexts

It will be clear from our domain description that contexts have to play a major role as a vehicle for assessing insight, understanding, and concepts.

A variety of contexts is needed, as well as a range of roles for the contexts. The variety is needed to minimize the chance of featuring issues and phenomena that are not culturally relevant. The range of roles for the contexts needs further elaboration because of the effects on what we are measuring relates to this role. Meyer (2001) distinguishes five different roles of the context: (a) to motivate, (b) for application, (c) as a source of mathematics, (d) as a source of solution strategies and (e) as an anchor for student understanding.

Distance to students

One can think of context as being certain “distances” from the students: the context that is closest is the private life (daily life); the next closest is school life, work, and sports; next is local community and society as encountered in daily life; and beyond that are scientific contexts. In this way, one can define a more or less continuous scale that can be regarded as another aspect of the framework. It is unclear how these distances affect students’ performance on tasks. This aspect needs further study, as results so far are inconclusive in the sense that we cannot say that “closer” contexts are more attractive for students or better suited for tasks than more scientific ones. Common belief suggests that less brilliant students “prefer” contexts closer to their immediate environment because they can engage more easily through the context. This can lead to items such as—

- The ice cream seller has computed that if he sells 10 ice creams, they will be the following kinds: 2 cups, 3 cones and 5 sticks. He orders 500 ice creams for the football game. What distribution of the different kinds will he use?
- Marge is lighter than Alice. Anny is lighter than Alice. Who is lighter: Anny or Marge?
- A pack of papers containing 500 sheets is 5 cm thick. How thick is one sheet of paper?

At the primary level we often see this kind of context that is “close to the student” and taken from his or her “daily” life. According to Gravemeijer (1994) and Dekker (1993), however, familiarity with the context can be a hindrance: There should be a “certain distance.”

Fantasy worlds offer another popular context in which students’ fantasy and creativity can lead to relevant, but not authentic, mathematical activities. Of course we cannot measure the distance to each student individually, so we have to make certain assumptions. One assumption, important because it relates directly to one of our guiding principles, is that the distance for a particular context might be different for girls and boys. We need to be aware of typical boys’ and girls’ contexts. Research by van den Heuvel-Panhuizen and Vermeer (1999) suggests that boys do better on experimental knowledge on numbers and measures from the daily real world, whereas girls seem to perform better on items where a standard algorithm or procedure is needed.

At the secondary level, an assumption that the context needs to be very close to the student does not hold. We notice at least two relevant issues. First, we notice more and more new real-worlds for the students—including the scientific and political worlds. But there also seems to be a tendency for this development to be postponed somewhat for lower-ability students. The reasons for this seem to be based more on practical classroom teacher intuition than on research.

Another aspect of context use that we need to be aware of is its role in assessment items. Of course we know that many items have no context at all, and looking at our domain definition, it seems highly unlikely that we will encounter this kind of problem often, but mathematics itself is part of our real world, so we are bound to encounter this aspect.

Relevance and Role of Context

Contexts can be present just to make the problem look like a real-world problem (fake context, camouflage context, “zero-order” context). We should stay away from such uses if possible.

The real “first-order” use of context is when the context is relevant and needed for solving the problem and judging the answer.

Second order use of context appears when one really needs to “mathematize” the problem in order to solve it, and one needs to reflect on the answer within the context to judge the correctness of the answer. So the distinction between first- and second-order use of context lies in the role of the mathematization process. In the first order, we have already premathematized the problem, whereas in the second order much emphasis is placed on this process (de Lange, 1979, 1987).

For this reason, we expect first order context use in most of the shorter items (e.g., multiple-choice; open-ended, short answer), whereas second-order context use is most often restricted to formats that allow for more process-oriented activities that quite often represent second- and third-level competencies and skills.

Special mention should be made of third-order context use, in which the context serves the construction or reinvention of new mathematical concepts. A very simple example is the use of a bus ride as a model for addition and subtraction (van den Brink, 1989).

Real Versus Artificial Versus Virtual Contexts

It seems clear that when we emphasize mathematics education that will prepare our citizens to be intelligent and informed citizens, we have to deal with all kinds of real contexts. We have to deal with pollution problems, with traffic safety, with population growth. But does this mean that we have to exclude artificial and virtual contexts? The answer is no, but we need to be aware of the differences for students.

A virtual context contains elements that are not drawn form any existing physical, social, practical, or scientific reality. They are of an idealized, stylized or generalized nature. For instance, if a stylized street layout of a city C is considered for an idealized traffic problem, it is only the labels “street,” “city,” and “traffic” that are real—the city, streets, and traffic are not real or authentic.

An artificial context deals for instance with fairy tales—nonexistent objects or constructs. This class of context is easier to separate from the real context and should be used with care. Students will not always be able to co-fantasize within this artificial setting or engage in a world that is clearly not real. But sometimes the use of these situations can be justified.

For all uses of context, the conditions are that they feature mathematics and that they enable us to analyze systems and situations, sometimes even *before* they are established or put into practice and hence before resources are spent or dangers encountered.

Let us now turn our attention from this general feature that plays a role in all methods of assessment and in some large part can be decisive whether or not we get good assessment in the sense that students are willing and eager to engage in the problem we pose to them. We will first discuss aspects of the daily classroom practice that are not always considered as having valuable assessment aspects: discourse, observations, and homework.

Discourse

Discussing, explaining, justifying, illustrating, and analogizing are all features of reasoning in a mathematics classroom. Having an argument to find the appropriate mathematical solutions and definitions is generally considered as contributing both to the whole classroom's learning as well to each individual's progress (Wood, 1998; Cobb, Yackel, & Wood, 1993). Under the specific conditions of a classroom, argumentation can have a strong impact on learning. Classroom interaction is based on the assumption that the students are at different "levels of mathematical and social competencies and skills" or that there are "framing differences" (Krummheuer, 1995). So there will only be an agreement about the core of the argument. Furthermore, the core of the argument means something different for each student, depending on the framing. This in turn leads to different levels of confidence.

Important in argumentation is the ability to construct a structural similarity among several argumentation experiences in different situations. Such a "pattern" of similarly structured arguments is called a *topos*. The argumentation in classroom can contribute to the formation of a *topos* for an individual student, which leads to conceptual mathematical development.

Krummheuer gives an example of two students who know the argumentation for solving simple balancing scales problems but cannot make use of it, which means that no individual *topos* is available. Of course, this is important information for the teacher, and as such, formative assessment.

In summary, Krummheuer uses the notions of data, conclusions, warrants, and backing as a means to analyze argumentation. According to Yackel (1995), Krummheuer's approach is useful for two reasons: It clarifies the relation between the individual and the collective, and—especially relevant here—it provides a way to demonstrate the changes that take place over time.

Another example is given by van Reeuwijk (1993), showing how students' knowledge of the concept of 'average' was assessed during a classroom discussion. The question posed was whether it is possible to compute the average size of families using the following data:

# children per family	*# families (in thousands)*
0	1,176
1	810
2	1,016
3	417
4	149
5	59
6	23
7 or more	16

Student A: How do we have to do this?

Student B: Just sum up and divide.

Student C: Yes, but what?

Student B: I don’t know. OK, there are 3,650 families.

Student C: OK, divide by 7.

Student A: That doesn’t make sense.

Teacher: What is a family?

Student : Mum, Dad and the kids.

Student: So, we have to find out how many kids in a family.

Teacher: How many children are there?

Student: 28 or more. Oh no, that doesn’t make sense.

Teacher: How many families are there with no kids?

Student: 1,176

Teacher: How many kids is that?

Student: (Surprised) None!

Student: That means 810 kids in the families with no kids.

An article by Cobb (1999) provides another example of an interesting classroom discourse that gives important information about where students are in the teaching and learning process, and as such is part of the assessment process. It is about reasoning with data from the outset. The activity focused on the question of whether the introduction of a police speed trap in a zone with a speed limit of 50 miles per hour had slowed down traffic and thus reduced accidents. The data are shown in Figure 3. The bottom graph shows the speeds of 60 cars before the speed trap was introduced, and the top one shows the speeds of 60 cars after the speed trap has been in use for some time. To begin the discussion, one of the teachers asked Janice to read the report of her analysis:

> If you look at the graphs and look at them like hills, then for the before group, the speeds are spread out and more than 55, and if you look at the after graph, then more people are bunched up close to the speed limit, which means that the majority of the people slowed down close to the speed limit.

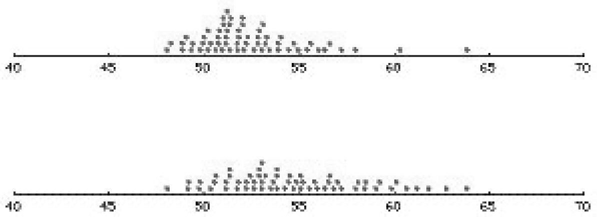


Figure 3. Graph of data from Cobb's (1999) speed trap activity

As Cobb noticed, this was the first occasion in public classroom discourse in which a student described a data set in global, qualitative terms by referring to its shape. Both teachers then capitalized on Janice's contribution in the remainder of the discussion, treating other students' analysis as attempts to describe qualitative differences in the data sets in quantitative terms. For example, Karen explained that she had organized the data sets by using a fixed interval width of five: "Like, on the first one, most people are from 50 to 60—that's where most people were in the graph." One of the teachers checked whether other students agreed with her interpretation. Karen then continued: "And then on the top one, most people were between 50 and 55 because, um, lots of people slowed down… So, like, more people were between 50 and 55."

It goes without saying that the subject of discourse lends itself for much more attention than we can give within this framework. The only point we want to make here is that, assuming we offer the students worthwhile tasks and organize the interaction and discourse in some organized way (there are infinitely many), we will not only contribute to the conceptual development of the students but also gain more insight in the failures and successes of that development and will be able to adjust our teaching and learning practice. As such, it is one of the most important aspects of classroom assessment.

Johnson & Johnson (1990) present a meta-analysis to show that collaborative discourse can produce significant gains in learning. In the science area, Rodrigues & Bell (1995), Cosgrove & Schaverien (1996), and Duschl & Gitomer (1997) report more support for learning gains by means of discourse.

Observations

The discussion about discourse blends naturally with one on observations, especially if we see observations in an interactive classroom environment. But observations include more than that discourse alone. Observations show which students perform better alone, and which perform better in groups. They give insight into how well students organize the result of a discussion on paper, how organized they are. They help teachers estimate confidence levels that are so important in order to engage in mathematical argument.

Many teachers have relegated the important information gained from observation to second-class status behind information that can be gained from a “test.” Part of the problem is that observations are hard to organize in a systematic way and the information is too informal to make hard educational decisions. The introduction of new information technology such as PDIs and spreadsheets, however, makes it somewhat easier to make this format a more serious contender, especially for classroom assessment.

And some help is available for those teachers who want to make effective use of observations. Beyer (1993) gives some suggestions that are ecologically valid: Use your reflections as teacher in order to develop your own performance indicators. Next: Try to record student performance against your indicators on a regular basis. This might sound more complicated than necessary. Performance indicators could indicate, for example, three simple levels—do not understand, on the way to understanding, and really understanding the concept—and aim only at very important mathematical concepts, limiting the scope but still getting very relevant information. These indicators can be developed in more detail using this framework or other external sources. Most teachers know very well what Webb (1995) notes: Systematic observations of students doing mathematics as they work on a project supported by their responses to probing questions are more authentic indicators of their ability to do mathematics than a test score compiled by totaling the number of correct item responses.

Homework

Homework is not very often used as or perceived to be assessment, and certainly not if we are thinking about assessment as we see it. Quite often, little thought is given to the selection of homework problems ("just do the next sums"), nor is there an elaborate discussion about the results. This is not too surprising, given that many of the homework problems generally encourage only superficial and rote learning. But the exceptions show the possibilities and strengths, as became clear within the observations of the NCISLA-RAP project. All students got the same homework. The problems were carefully selected to guarantee possibilities for different strategies in the students' solutions. The teacher first checked whether the students had successfully accomplished the homework and made notes (grades) in special cases (better than, worse than). Next, the teacher invited several students to write their homework on the blackboard, making sure those students represented different strategies and solutions. Then all solutions were discussed in a plenary classroom session that involved all students. Students could make up their minds and make revisions to their own work. During this discussion, and based upon input by individual students, the teacher could make more notes on students' understanding of the mathematics involved.

Operationalizing homework in this way brings together the best aspects of discourse and observations and also gives the students an excellent introduction into self-assessment.

Self-assessment

It is tempting to quote here a postulate by Wiggins (1993): "An authentic education makes self-assessment central." The means for dispelling secrecy are the same means for ensuring a higher across-the-board quality of work from even our most worrisome students: teaching students how to self-assess and self-adjust, based on the performance standards and criteria to be used. A practical implication of this postulate is that we should require students to submit a self-assessment with major pieces of work. Feed back from the teacher will help make clear to the students how the teacher assesses in relation to their own perception of "quality." This in turn will most likely improve the student's learning through a better understanding of the criteria and especially the different levels of mathematical competencies as they play out on tests and tasks.Wiggins's postulate is not trivial, as confirmed by a literature review conducted by Black and Wiliam (1998), in which they conclude that a focus on self-assessment by students is not common practice, even among teachers who take assessment seriously. Also remarkable is the fact that in the general literature on classroom assessment, the topic is frequently overlooked, just as it is in the otherwise comprehensive collection by Phye (1997).

Arguments to introduce self-assessment vary. We have already noted Wiggins's point of view. Norway introduced self- and peer-assessment as an intrinsic part of any program that aims to help students take more responsibility for their own learning. A further argument is that students need to reflect on their own understanding and how to change their understanding, and self-assessment is an excellent tool to achieve this. Sadler (1989) argues that students cannot change their understanding unless they can first understand the goals that they are failing to attain, develop at the same time an overview in which they can locate their own position in relation to these goals, and then proceed to pursue and internalize learning that changes their understanding. In this view, self-assessment is a sine qua non for effective learning. The research data are in general very positive: several qualitative studies report on innovations to foster self-reflection. We just mention the results that show that students can be helped by using self-assessment to realize, through feedback on their self-assessment, the lack of correspondence between their own perception of their work and the judgments of others. This leads to quality improvement in the students' work (Merrett & Merrett, 1992; Griffiths & Davies, 1993; Powell & Makin, 1994; Meyer & Woodruff, 1997).

We have already indicated that homework can play a role in self-assessment, but it can also function within the concept of peer-assessment—students judging the work of students. And both self-assessment and peer assessment can find a very effective form when the "own production" format is used. Peer assessment will be discussed next, followed by students' "own productions."

Peer Assessment

Peer assessment, like self-assessment, can take many forms. Students may be asked to grade an otherwise "traditional" test, to comment on an oral presentation by another student, or to construct test items or even whole tasks (Koch & Shulamith, 1991; de Lange et al., 1993; Streefland, 1990; van den Brink, 1987). The rate of success has not been well established because peer assessment is often introduced at the same time as other innovations such as group work (Webb, 1995). Peer assessment provokes a discussion among students about different strategies and solutions and helps them to clarify their views in a setting where they can feel safe.As a more concrete example of both self-assessment and peer assessment that relates in a positive way to our principles—more particularly, positive testing—we will next turn to "own productions" as assessment.

Own Productions

If one of our principles is that testing should be positive—which means that we should offer students an opportunity to show their abilities—and that tests are part of the teaching and learning process, then own productions offer nice possibilities. The idea of own productions is not really new. Reports about experiences go back a long time. Treffers (1987) has introduced the distinction between construction and production, which according to him is no matter of principle. Rather, free production is the most pregnant way for constructions to express themselves.

By constructions, we mean—

- Solving relatively open problems that elicit divergent production due to the great variety of solutions they admit, often at various levels of mathematization.

And—

- Solving incomplete problems that require self-supplying data and references before they can be solved.

The construction space for free productions might be even wider:

- Contriving own problems (easy, moderate, difficult) as a test paper or as a problem book about a theme or a course, authored to serve the next cohort of pupils (Streefland, 1990).

The third suggestion—have students produce test items or a test—was implemented in a study on the subject of data visualization in an American high school. The authors (de Lange & van Reeuwijk, 1993) describe in detail how a teaching and learning sequence of about five weeks was assessed. Apart from two more-or-less-traditional formats with some unusual problems, the end-of-the-period test was different. The tasks were presented to the students in the following way:

Data Visualization Task

> This task is a very simple one. At this moment, you have worked yourself through the first two chapters of the book and taken relatively ordinary tests. This task is different:
>
> Design a test for your fellow students that covers the whole booklet.
>
> You can start your preparation now: Look at magazines, books, and newspapers for data, charts, and graphs that you want to use. Write down ideas that come up during school time.

After finishing the last lesson, you will have another three weeks to design the test. Keep in mind:

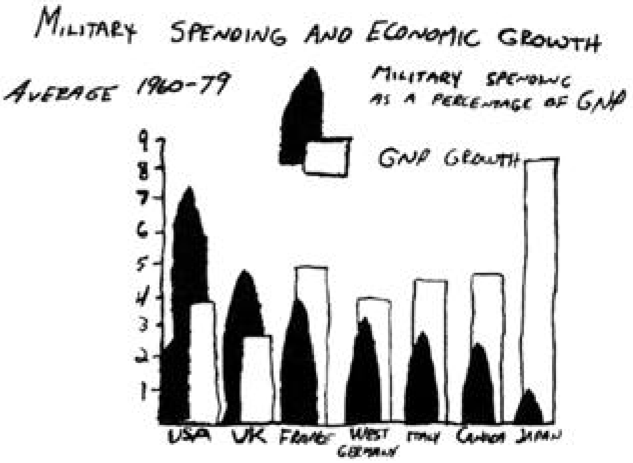
The test should be taken in one hour.

You should know all the answers.

Good luck!

The results of this test were very helpful indeed for the teacher involved. Some students showed very disappointing work: They simply copied items from the booklet, avoiding any risk-taking or creativity. Others showed that even that was too much for them: Answers to questions that were discussed in the classroom showed that the students' learning was minimal at best. Although many of the questions designed by the students were much better, they were quite often at Competency Level 1 (see Figure 4).

The graph shown here is a bar graph involving information about how much money is being spent on military supplies compared to the country's GNP.



Is this information easy to read and understandable? Why or why not?

No, this information is not easy to read because the numbers on the left have no meaning. One cannot tell if it means millions, billions, etc.

Could this information be represented better? Explain.

No, because a line graph (box-plot, pie graph, stem leaf, etc.) would not work.

Is this graph accurate? Explain your answer.

No, because this graph is based on an average between 1960 and 1979.

Figure 4. Math item written by student

The researchers concluded that if the designed tests were to reflect the perceived curriculum, that this did not meet the intended goals. Most of the exercises were rather traditional and mainly addressed issues as representing numbers in a graph or reading and interpreting graphs. Some of the items that were designed, however, operationalized the higher levels in a very naive way. These were used to provoke classroom discourse about the teaching and learning process of the previous five weeks.

It was evident that this way of collecting information about the teaching and learning process was very effective and was also well suited for giving feedback to the students.

Van den Brink (1987) suggested a slightly different “own production,” carrying out experiments with first graders. The idea was that they would act as authors of a math textbook. This idea immediately raises many questions. Van den Brink mentions:

- Are children motivated by the idea of writing down their knowledge for others?
- Should the books be written at the end of the year or developed by bits over the year?
- Should arithmetic lessons given by the teacher have a place in the books?
- Will children who come from different ethnic backgrounds create different books?
- Will writing books force children to reflect?
- Will the book give a reliable picture of the state of knowledge of the students?

From an assessment perspective, the last question and the one about reflection are the most intriguing. The experiments, which took place at two different schools, seem to point in the direction that indeed the activity forces the student to reflect on his or her own learning and represents their mathematical knowledge in a fair way.

Some other research results strongly support this “method” for formative assessment: King (1990, 1992a, 1992b, 1994) found that training that prompted students to generate their own specific, thought-provoking questions and then attempt to answer them is very effective. Koch and Shulamith (1991) reported similar results, showing that students’ own questions produced better results than did adjunct questions from the teacher.

Next we will turn our attention to other tools or formats for assessment organized in a somewhat logical order from simple multiple-choice to very complex project tasks.

Multiple-Choice

In constructing an achievement test to fit to a desired goal, the test maker has a variety of item types from which to choose. It will come as no surprise that the multiple-choice format seems to be the “best” format if we simply judge by its popularity.

Multiple-choice, true-false, and matching items all belong to the same category: the selection-type items. Officially, they are so popular because they are objective items—they can be scored objectively. That means that equally competent scorers can score them independently and obtain the same results. These equally competent scorers are usually computers. And therein lies the real popularity of selection-type items: They can be scored by a computer and are therefore very cheap to administer.

The rules for constructing a multiple-choice item are simple: A multiple-choice item will present students with a task that is both important and clearly understood and one that can be answered correctly only by those who have achieved the desired learning (Gronlund, 1968). This is not as simple as it seems, as we all know, especially if we include that the item should operationalize a specific goal.

Another frequently mentioned problem with multiple-choice items is the necessity that only those who have achieved the desired learning are able to answer the question correctly. But some who answer correctly may have just guessed or coincidentally wrote down the right letter. When using the multiple-choice format there will always remain a doubt regarding the validity of the knowledge assessed.

The task of constructing a multiple-choice item that is important for the students, can be clearly understood, can be answered correctly only by those who have achieved the desired learning, and operationalizes a specific goal or learning outcome is not simple.

Many items have flaws, and all have very limited value if we are really heading for authentic assessment. At this moment, the only area of application seems to be to operationalize the lower goals. In our opinion, open questions offer more possibilities than are usually exploited. Properly constructed open questions, with a variety of short, long and extended responses do offer some possibilities for assessment at a level higher than the lowest—whatever name we give to the lower levels. They may be called knowledge outcomes and a variety of intellectual skills and abilities, or computation and comprehension, or basic skills and facts. Whatever the words, it is generally agreed that we need other instruments (such as essay tests) that provide a freedom of response that is needed for measuring complex or higher order learning outcomes.

(Closed) Open Questions

Multiple-choice items are often characterized as closed questions. This suggests that there are open questions as well. However, we have to be careful. Sometimes the question is open by format but closed by nature. The respondent has to answer by a number, a yes or no, a definition, and maybe a simple graph or a formula. There is hardly any thinking or reflection involved. This category is mostly in close competition with the multiple-choice format. The following question provides an example: “A car takes 15 minutes to travel 20 kilometers. What is the speed of the car in kilometers per hour?”

The same question can be posed easily in the multiple-choice format.

The distinction between closed-open questions and open-open questions is rather arbitrary; however, we should still pay attention to this aspect when designing tests.

(Open) Open Questions

In our perception, an open-open question differs from the closed-open question in respect to the activities involved in getting a proper answer. This proper answer can still be just a number or formula but the process to get there is slightly more complicated or involves higher order activities (see Figure 5).

Which of the four building do you see on the TV?(29)

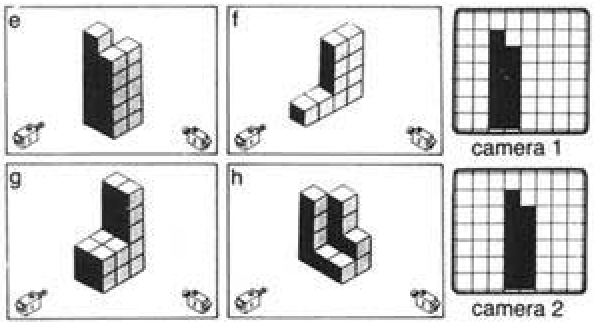
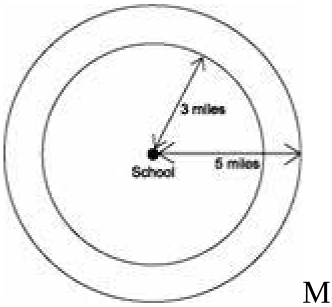


Figure 5. Example of an open-open question

Extended Response–Open Questions

Extended response–open questions give students the opportunity to get involved in a context with one or more open questions of relatively complex nature, where the student’s choice of which strategy to follow is not clear in advance. Super items fit into this category. This category differs from open-open questions in that we expect the students to explain their reasoning process as part of their answer. An example of this type of question is “Martin is living three miles from school and Alice five miles. How far apart are Martin and Alice living from each other?”

Of course the answer “they live two miles apart” is just part of the correct answer. Students may make a drawing to explain their reasoning:



Martin and Alice could live 8 miles from each other, or 2 miles or any number in between.

Super Items

Extended response “super items” (Collis, Romberg, & Jurdak, 1986) are tasks that give students the opportunity to get involved with a context or problem situation by asking a series of open questions of increasing complexity. The first few questions might be closed-open questions or open-open questions. These are followed by more difficult questions that may not have a predefinable set of possible answers. For such questions, multiple scoring and some adjudication of responses is likely.

Multiple-Question Items

This format resembles the Collis, Romberg and Jurdak super items in the sense that one context or problem situation forms the setting for the questions. Unlike the super items, however, there is no strict order necessary in the structure of the range of questions. It is likely that the first question will be simple in order to engage the students; it is also more than likely that the last question will be at Level 2 or Level 3. But in between there is quite a bit of freedom in the structure.

Essays

> The construction and selection of items is not difficult for the lower levels of cognitive behavior—computation and comprehension. The difficulties were presented at the higher levels. …….Construction of a proof is generally an analysis level behavior and is certainly something that should be tested. However, the IMC, in consultation with many mathematicians and mathematics educators, tried a variety of multiple-choice items to test the ability to construct proofs. None was satisfactory. Hence, only a few items testing the ability to analyze a proof were included.

This statement from the International Association for the Evaluation of Educational Achievement's (IEA) Second International Mathematics Study (SIMS; Travers & Westbury, 1988) clarifies many of the problems we face in mathematics education. First, as we indicated before, it is not at all easy to produce good items for the lower levels, and certainly not with the multiple-choice format. However there seems to be a general feeling, also prevalent in the TIMSS study, that it is not at all difficult. So the first problem is to show and convince specialists—and in math education there seem to be a lot of specialists—that we have a serious problem here.

Secondly, there is the problem that we have often seen a presentation of test items that are intended to measure higher order goals but fail to realize that objective.

Thirdly, everyone agrees that the higher levels should be tested. Some even state that the higher levels are the most important. It is astonishing, then, to see that because of the "lack of satisfactory" multiple-choice items, only a few items are used at the higher levels. The third problem: Why don't we explore at least some of the tools that are available to us to operationalize the higher levels?

An old-fashioned but rarely used tool in mathematics education is the essay test. As Gronlund (1968) stated: Essay tests are inefficient for measuring knowledge outcomes, but they provide a freedom of response that is needed for measuring complex outcomes. These include the ability to create, to organize, to integrate, to express, and similar behaviors that call for the production and synthesis of ideas.

The most notable characteristic of the essay test is the freedom of response it provides. The student is asked a question that requires him to produce his own answer. The essay question places a premium on the ability to produce, integrate, and express ideas.

Its shortcomings are well known: The essay task offers only a limited sampling of achievement, writing ability tends to influence the quality of the answers, and the essays are hard to score objectively.

Essays can come very close to extended response–open questions, especially in mathematics education. Another often-mentioned aspect of the essay is whether it should be written at school or at home. Although the essay task is usually seen as a take-home task, this is not necessary. One can easily think of smaller essay problems that could be completed at school. It is generally accepted that the more precise or "closed" the questions are, the more objective the scoring is.

From this perspective, one is tempted to conclude that this task can be scored with reasonable objectivity, or better, in a good, intersubjective way.

Oral Tasks and Interviews

In some countries oral assessment was usual practice, even as part of the formal national examination system. There are different forms, of which we cite three:

- An oral discussion on certain mathematical subjects that are known to the students.
- An oral discussion on a subject—covering a take-home task—that is given to the students for 20 minutes prior to the discussion.
- An oral discussion on a take-home task after the task has been completed by the students.

Quite often the oral assessment format is used to operationalize the higher process goals.

Two-Stage Tasks

Any task that combines test formats can rightfully be called a two-stage task. An oral task on the same subject as an earlier written task is a typical example. Two-stage tasks characteristically combine the advantages of the traditional, restricted-time written tests with the possibilities offered by tasks that are more open.

The characteristics of restricted-time written tests as considered here are—

- All students are administered the same test at the same time.
- All students must complete it within a fixed time limit.
- The test—
  - Is oriented more toward finding out what students *do not* know than what they *do* know.
  - Usually operationalizes the “lower” goals (i.e., computation, comprehension).
  - Consists of open questions.
- Scores are as objective as they can be given the fact that we exclude multiple-choice format.

These, then, are the characteristics of the first stage of the task.

The second stage should complement what we miss in the first stage as well as what we really want to operationalize. The characteristics of the second stage are—

- There is no time limit.
- The test emphasizes what you know (rather than what you don’t).

- Much attention is given to the operationalization of higher goals (e.g., interpretation, reflection, communication).
- The structure of the test is more open: long-answer questions and essay-type questions.
- Scoring can be difficult and less than objective; intersubjectivity in grading should be stressed.

Journals

Journal writing is one of the least used forms of assessment. This seems to be because it is time-consuming, hard to score the mathematics separate from the reading and writing skills, and unclear how to score students' work. But, like drawing schemata and graphs, writing mathematically—shaping, clarifying, and discovering ideas (Bagley & Gallenberger, 1992)—is a very important mathematical ability.

Concept Maps

White (1992) has suggested that concept mapping can be used with students to show how they see relationships between key concepts or terms within a body of knowledge. This activity, like own productions, forces students to reflect on such relationships and to develop a more integrated understanding, as opposed to learning isolated facts. Following this line of thought, it will be clear that this idea fits nicely with the introduction of big ideas instead of curriculum strands. Studies suggest that—

- Concept maps can assist students in effectively organizing their knowledge of a topic.
- Students come to understand how they learn through the use of concept maps.
- Teachers can gain valuable information about the relationship among concepts that students have already constructed.
- Concepts maps can help teachers identify misconceptions that do not come to the surface with other assessment techniques.

(Santos, Driscoll, & Briars, 1993).

According to Roth and Roychoudhury (1993), who also recommend the use of concept maps as an aid in discourse, such maps drawn by the students serve to provide useful points of reference in clarifying the points under discussion and enable the teacher to engage in "dynamic assessment."

Progress-Over-Time Tests

Progress over time has always been an implicit aspect of assessing. The next task was supposed to be more difficult than the previous one, and the curricular organization takes care of that aspect as well: everything that comes later is more complex or at a higher level. But we may need a more explicit way to measure mathematical growth. One way to do this is by using almost similar problems in tests given at different times. As an example, we refer to the Mathematics in Context Longitudinal Study, in which end-of-the-year tests were developed that contain one item (actually a super item). This item was revisited in all four end-of-the-year tests, albeit in a more complex form as the years progressed.

Reporting: Feedback and Scoring

Feedback

Designing and selecting tasks is one thing, but how to establish quality feedback is quite another, and a very important one. Without proper feedback the whole concept of assessment contributing to the learning process is endangered.

Feedback possibilities depend clearly on the “format” that is chosen. In discourse the feedback can be immediate and very differentiated in the sense that the feedback can be *direct* (giving the student information about what is wrong and why and giving a suggestion for correction) but also and probably quite often, *indirect* (just asking whether the student is “sure” and can explain his answer and comparing it with other answers given by fellow students).

Feedback possibilities with the multiple-choice format are not abundant: Usually, the only feedbacks students get is whether something was correct or incorrect; in a best-case scenario, the teacher might spend some time in the classroom highlighting some of the most common incorrect answers.

Within the common restricted-time written test, there are ample opportunities to give dedicated, individual feedback to the student. This is time-consuming and the quality of the feedback depends to a large extent on how the student’s answers are formulated. If the student fails to write down anything relevant, the question of quality feedback becomes an extremely difficult one. In such cases, oral feedback after additional questioning seems to be the only option.

Feedback can also have a very stimulating effect. Consider, for example, the homework method. Compare the experience of a student who is assigned homework but nothing is done beyond “checking” whether he or she “did” it, versus the student who gets quality feedback (as described in the Homework section). This was also pointed out in a study in Venezuela on mathematics homework (Elawar & Corno, 1985). One group of students was given specific feedback on specific errors and strategies used. Another group followed the “normal” practice of homework without comments. Analysis of the results showed a large effect of the feedback treatment on future student achievement.

A definition for feedback can be found in Ramaprasad (1983): “Feedback is information about the gap between the actual level and the reference level of a system parameter, which is used to alter the gap in some way. In order for feedback to exist, the information about the gap must be used in altering the gap.”

This definition is a little too restricted for our purposes because the “gap” need not necessarily be a gap in the strict sense. Students might be able to solve a problem at very different levels of mathematization and formalization. But they are all successful. So theoretically speaking there is no gap. But we might still use the feedback mechanism to bridge the level-of-formality “gap”: to show the students on a less formal level what is possible with some more formal mathematics. It can also be used the other way around: to show the more formal students how elegant—maybe even superior—“common sense” solutions can be.

Kluger and DeNisi (1996) identified four different ways to close the gap. The first one will come as no surprise: try to reach the standard or reference level—this needs clear goals and high commitment on the part of the learner. On the other end of the scale, one can abandon the standard completely. In between we have the option of lowering the standard. And finally, one can deny the gap exists.

Kluger and DeNisi also identified three levels of linked processes involved in the regulation of task performance: meta-task processes involving the self, task motivation processes involving the focal task, and finally the task-learning processes involving details of the task.

About the meta-task processes, it might be interesting to note that feedback that directs attention to the self rather than the task appears likely to have negative effects on performance (Siero & Van Oudenhoven, 1995; Good & Grouws, 1975; Butler 1987). In contrast to those interventions that cue attention to meta-task processes, feedback interventions that direct attention toward the task itself are generally more successful.

In 1998, Black and Wiliam were surprised to see how little attention in the research literature had been given to task characteristics and the effectiveness of feedback. They concluded that feedback appears to be less successful in “heavily-cued” situations (e.g., those found in computer-based instruction and programmed learning sequences) and relatively more successful in situations that involve “higher-order” thinking (e.g., unstructured test comprehension exercises).

From our own research (de Lange, 1987), it became clear that the “two-stage task” format affords excellent opportunities for high-quality feedback, especially between the first and second stages of the task. This is in part due to the nature of the task format: After completion of the first stage, the students are given feedback that they can use immediately to complete the second stage. In other words, the students can “apply” the feedback immediately in a new but analogous situation, something they were able to do very successfully.

Scoring

Wiggins (1992) points out, quite correctly, that feedback is often confused with test scores. This perception is one of many indications that feedback is not properly understood. A score on a test is encoded information, whereas feedback is information that provides the performer with direct, usable insights into current performance and is based on tangible differences between current performance and hoped-for performance.

So what we need is quality feedback on one side and “scores” to keep track of growth in a more quantitative way on the other side. And quite often we need to accept that we are unable to quantify in the traditional sense (e.g., on a scale from one to ten), but just make short notes when, during a discourse or during homework, a student does something special, whether good or bad.

Many of the formats described before have in common that they are in a free-response format. Analysis of students’ responses to free-response items can provide valuable insights into the nature of student knowledge and understanding and in that sense help us formulate quality feedback. With such formats we get information on the method the student uses in approaching the problem and information about the misconceptions or error types that they may demonstrate.

But as the TIMSS designers observed (Martin & Kelly, 1996), student responses to free-response items scored only for correctness would yield no information on how the students approached problems. So TIMSS developed a special coding system that can also be used in classroom assessment to provide diagnostic information in addition to information about the correctness of the student responses. It was proposed that a two-digit coding system be employed for all free-response question items. The first digit, ranging between 1 and 3, would be used for a correctness score, and the second digit would relate to the approach or strategy used by the student. Numbers between 70 and 79 would be assigned to categories of incorrect response attempts, while 99 would be used to indicate that the student did not even try. This TIMSS coding system, which was later adapted successfully for the Longitudinal Study on Middle School Mathematics (Shafer & Romberg, 1999), is demonstrated in Table 1 using a generic example of the coding scheme worth one point.

Table 1. Example coding scheme using the TIMSS coding system

Write down the reason why we get thirsty on a hot day and have to drink a lot.		
Code	Response	Example
	Correct responses	
10	Refers to perspiration and its cooling effect and the need to replace lost water.	
11	Refers to perspiration and only replacement of lost water.	• *Because when we are hot, our body opens the pores on our skin and we lose a lot of salt and liquid.*
12	Refers to perspiration and only its cooling effect.	
13	Refers to perspiration only.	• *We are sweating.* • *Your body gives away much water.* • *We are sweating and get drier.*
19	Other acceptable explanation.	
	Incorrect responses	
70	Refers to body temperature (being too hot) but does not answer why we get thirsty.	• *You cool down by drinking something cold.*
71	Refers only to drying of the body.	• *Your throat (or mouth) gets dry.* • *You get drier.* • *The heat dries everything.*
72	Refers to getting more energy by drinking more water.	• *You get exhausted.*
76	Merely repeats the information in the stem.	• *Because it is hot.* • *You need water.*
79	Other incorrect responses.	• *You loose salt.*
	Nonresponses	
90	Crossed out or erased, illegible, or impossible to interpret.	
99	BLANK	

Student responses coded as 10, 11, 12, 13, or 19 are correct and earn one point. The second digit denotes the type of response in terms of the approach used or explanation provided. A response coded as 10 demonstrates a correct response that uses Strategy 1. For items worth more than one point, rubrics were developed to allow partial credit to describe the approach used or the explanation provided.

Student responses coded as 70, 71, 76, or 79 are incorrect and earn zero points. The second digit gives us a representation for the misconception displayed, incorrect strategy used, or incomplete explanation given. This gives the teacher a good overview of where the classroom as a whole stands, as well as individual differences, which can lead to adequate and effective feedback.

Student responses with 90 or 99 also earn zero points. A score of 90 means the student attempted but failed completely, and 99 represents no attempt at all.

Another addition to the scoring system that can be very helpful is a code for the level of mathematical competency. Of course, when a teacher designs her classroom assessment system she will balance it in relation to the levels of mathematical competencies. But this will not necessarily lead to information on the levels of individual students.

A crucial and potential weak point arises when we are dealing with partial credit, as will quite often be the case. This is a difficult point for students and teachers alike. Without preparation, guidelines, exemplary student responses, or a proper “assessment” contract between teacher and students, partial-credit scoring can be a frustrating experience even though its necessity is obvious. We therefore discuss the issue of partial credits in a little more detail through the following examples.

First, we will present an example of a very simple and straightforward method for partial scoring credits in the form of an (external) examination item about a cotton cloth for a round table (National Examination, The Netherlands, 1992).

> Nowadays you quite often see small round tables with an overhanging cloth [Figure 6]. You can make such a cover yourself using—
>
> - Cotton, 90 cm wide; 14.95 guilders per meter
> - Cotton, 180 cm wide; 27.95 guilders per meter
> - Ornamental strip, 2 cm wide; 1.65 guilders per meter
>
> When buying the cotton or strip, the length is rounded to the nearest 10 cm. For instance, if you want 45 cm, you need to buy 50 cm.
>
> 1. Marja has a small, round table: height 60 cm; diameter 50 cm. On top of the table, she puts a round cloth with a diameter of 106 cm.
>
>   *3 points—How high above the ground will the cloth reach?*

2. Marja is going to buy cloth to make her own cover. She wants it to reach precisely to the ground. It will be made from one piece of cotton fabric and will be as economical as possible. There will be a hem of 1 cm.

   6 points—Compute the amount of cotton Marja will have to buy and how much that will cost.

3. Marja wants an ornamental strip around the edge of the cloth.

   4 points—Compute how much ornamental strip she will need and how much it will cost.

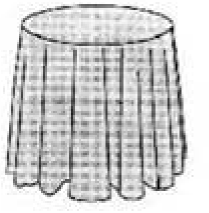


Figure 6. Round table with overhanging cloth

This example shows what some will recognize as a typical female context with questions at Levels 1 and 2. Beforehand, it shows the students clearly how many points they can earn for answering each of the questions. Next, we provide guidelines for the teachers' scoring (Table 2).

Table 2. Example of Scoring Guidelines for Teachers

No.	Max. score	Response	Points
1	**3**	Answer: 32 cm	1
		Proper explanation	2
2	**6**	Diameter: 172 cm	1
		Proper explanation	2
		Answer: 180 cm of cotton cloth	1
		Width: 180 cm	1
		Price: 1.80 × 27.95 = 50.31 (or 50.30*)	1
3	**4**	Diameter: 170 cm	1
		Circumference: 534 cm (π × 170)	1
		Answer: She has to buy 540 cm	1
		Answer: The cost will be 5.40 × 1.65 = 8.91 guilders (or 8.90*)	1

*Note: The Netherlands did not use single-unit coins at the time.

This might seem clear but of course it is not. There are still many answers possible for which the subjective judgment of one teacher might differ from another. That is why it is advisable to use intersubjective scoring with external examinations. With intersubjective scoring, at least two teachers score the test independently, and they have to come to an agreement. This is a must for high-stakes testing but can also be done on a somewhat regular basis in the classroom if teachers coordinate their classroom assessment practices.

Scores are usually on a 100-point scale and are deceptive in the sense that a score of 66 actually means a score from something like 62 to 69 and thus seems more precise than it actually is. But the advantage is that students can check the judgment of the teacher and start a discussion about a score based on clear points of departure.

If so-called "holistic" scoring is used, the clarity is less obvious because there is more room for subjective judgment. Under holistic scoring, we group the scoring systems that are quite often very meaningful for such formats as essays, journals, and projects but are nowadays also used for formats that can be scored more easily. As an example we present two higher-grade descriptors of journals by Clarke, Stephens, and Waywood (1992):

A: Makes excellent use of her journal to explore and review the mathematics she is learning. She uses mathematical language appropriately and asks questions to focus and extend her learning. She can think through the difficulties she encounters.

B: Maintains regular entries and is able to record a sequence of ideas. She uses examples to illustrate and further her understanding and is able to use formal language to express ideas but is yet to develop mathematical explorations.

And two descriptors by Stephens and Money (1993) for extended tasks (not complete):

A: Demonstrated high-level skills of organization, analysis, and evaluation in the conduct of the investigation. Used high levels of mathematics appropriate to the task with accuracy.

B: Demonstrated skills of organization, analysis, and evaluation in the conduct of the investigation. Used mathematics appropriate to the task with accuracy.

It is clear that in this latter set of descriptors, subjective judgments are a greater risk than in the previous example. But for some formats we almost have to use this kind of scoring system. One can still use numbers of course, even on a 100-point scale for very complex tasks. Exemplary student work and how the teacher judged it can be very helpful. This of course is also part of the assessment contract between teacher and students. Students need to know clearly what the teacher values—maybe not so much the correct answers but the reasoning or the solution's presentation and organization. But even without exemplary work, experienced teachers are very capable of sensible scoring on more complex tasks if we are willing to accept the uncertainty behind every scoring grade.

Our own research (de Lange, 1987) on how well teachers are able to score very open-ended tasks without any further help in the form of scoring rubrics showed that the disagreement among teachers grading the same task was acceptable for most teachers; if we assume that the average of a series of grades is the “correct” one, we noticed that 90% of the grades were within 5 points of the correct grade on a 100-point scale. Other research shows that especially the ordering of such complex tasks can be done with very high reliability (Kitchen, 1993). One example is the scoring system of the mathematics A-lympiad, a modeling contest for high school students that uses both some kind of holistic scoring (gold, silver, bronze, honorable mention) and an ordering system. Even though the commission that carried out the difficult task of scoring had many personal changes over time, agreement on the rank order was consistently high (De Haan & Wijers, 2000).

From Principles to Practice: The Process

Putting everything we have discussed so far together in an actual classroom environment is, to say the very least, a challenge. It is the aim of this part of the framework to propose a possible scenario.

Let us start with the Professional Standards for School Mathematics (NCTM, 1991). These standards envision teachers' responsibilities in four key areas:

- Setting goals and selecting or creating mathematical tasks to help students achieve these goals.
- Stimulating and managing classroom discourse so that both the students and the teacher are clearer about what is being learned.
- Creating a classroom environment to support teaching and learning mathematics.
- Analyzing student learning, the mathematical tasks, and the environment in order to make ongoing instructional decisions.

Hypothetical Learning Trajectory

These standards implicitly tell us that much of the teacher's responsibility involves planning. As Brousseau (1984) stated: "If the teacher has no intention, no plan, no problem or well-developed situation, the child will not do and will not learn anything." The consideration of (a) the learning goals, (b) the learning activities, and (c) the thinking and learning in which the students might engage is called the hypothetical learning trajectory (Simon, 1995).

Although it is necessary for a teacher to form and describe his hypothetical learning trajectory, it is also evident that this trajectory will never actually play out as planned in the classroom. A teacher might offer students nice, open-ended tasks but the teacher cannot predict a student's actual reactions, and therefore cannot predict the actual learning trajectory. So the trajectory will be modified continuously as the lesson cycle develops. And students' assessment plays a vital role in this modification process.

Of the three components in the trajectory, the teacher's learning goals seem to be the easiest to tackle, and an experienced teacher will also be able to develop a plan for learning activities (probably heavily based on the student materials available on the market). But the most difficult component is the teacher's hypothesis of the actual learning process. As Simon notes with some regret, the mathematics education literature is not abundant with research with emphasis on anticipating students' learning processes. A notable positive exception is the successful project, Cognitively Guided Instruction (Carpenter & Fennema, 1988; Carpenter et al., 1999), in which teachers learned much about research on children's thinking and thus were more capable of predicting and anticipating children's learning processes.

The design of a hypothetical learning trajectory. To a great extent, the student and teacher learning materials used will affect how complicated the design of the hypothetical learning trajectory will be. Sometimes the (textbook) materials help in a very limited way; sometimes they make more help available. As an example of the latter scenario, we have chosen the teacher guide for Looking at an Angle (Feijs, de Lange, Van Reeuwijk, Spence, & Brendefur, 1996), a unit from Mathematics in Context, a middle school curriculum funded by the National Science Foundation (NSF) and developed by the Wisconsin Center for Education Research (WCER; Madison, WI) and the Freudenthal Institute (Utrecht, The Netherlands).

In this teacher guide we find a rather explicit goal description on the three competency levels used in this framework. That by itself facilitates the design quite a bit. From the nine goals on

Level 1 (here called Conceptual and Procedural Knowledge), we quote: "understand the concepts of vision line, vision angle, and blind spot," and this goal is directly related to activities in the student materials that "offer ongoing assessment opportunities."

From Level 2, we also mention one goal: "understand the relationship among steepness, angle, and height-to-distance ratio." Again, the connection to the student materials shows that these connections are evenly spread out over the whole period of the unit (4–5 weeks), so that teachers and students can reflect on the concept several times and can work toward putting the relationship on a new and higher level of understanding.

Also, some activities are identified that refer explicitly to Level 3 competencies such as seeing the isomorphism in the different models used in this unit.

Not only are opportunities identified for ongoing formative assessment but also for "end-of-unit" assessment, which has both formative and summative aspects. Further help is available in the form of possible right and wrong answers and strategies.

In such a way, the teacher can get considerable support in formulating learning goals and planning for learning activities and can get backing for formulating the hypothetical learning process and how to relate this hypothetical learning trajectory to the assessments of student's knowledge (a particularly important aspect).

In his article on the learning trajectory, mentioned earlier, Simon (1995) correctly points out that the design of the trajectory with traditional textbooks is a difficult task. His approach represents a sharp contrast to the approach to instruction characteristic of traditional mathematics instruction and represented by traditional mathematics textbooks. Traditional instruction tends to focus on one skill or idea at a time and then provide considerable routine practice to "reinforce" that learning. Materials developed more recently differ in many ways from this approach and are more or less in line with the ideas represented by Simon, although they do not always directly represent the purely constructivist approach advocated in the article.

After forming a hypothetical learning trajectory—and the more experienced a teacher gets, the better the trajectory, assuming the flexibility of the teacher to adjust continuously—the next problem arises: where and when am I going to assess what, and how?

Our basic assumptions will be the following: there is a clearly defined curriculum for the whole year—including bypasses and scenic roads—and the time unit of coherent teaching within a cluster of related concepts is about a month. So that means that a teacher has learning trajectories with at least three “zoom” levels. The global level is the curriculum, the middle level is the next four weeks, and the micro level is the next hour(s). These levels will also have consequences for assessment: end-of-the-year assessment, end-of-the-unit assessment, and ongoing formative assessment.

Hypothetical Assessment Trajectory

Next we will see how to design a hypothetical assessment trajectory to fit the learning trajectory. Some of the ideas we describe have been suggested by Dekker and Querelle (1998).

Before. The first assessment activity should be when starting to teach a new concept or idea with some fundamental value (middle zoom level). The teacher wants to know whether the students have mastered the prior knowledge necessary to start successfully with a new unit. Already, this assessment activity will change the learning trajectory. Possible and suggested test formats for these goals are—

- **Oral test.** Questions are posed that involve basic knowledge and skills (Level 1). This format is appropriate because it enables the teacher to recapitulate important topics with the whole group in a very interactive way. Although basic knowledge and skills should be stressed, the format also allows the teacher to check Level 2 and 3 competencies in a relatively effective and fruitful way.
- **Aural test.** Questions are posed orally but answers are written down. This gives students who are not too fluent in English a second and probably fairer chance to express their ideas. This format also lends itself very well to checking whether students are able to express their informal knowledge about things to come; this is again relevant to designing the learning trajectory.
- **Entry test.** A short, written entry test consisting of open-ended questions.
- **Other test formats.** It goes without saying that the teacher is free to choose from any of the test formats described before or to design other test formats.

During. While in the trajectory, there are several issues that are of importance to teachers and students alike. One is the occurrence of misconceptions of core ideas and concepts. Because students in a socio-constructivist or interactive classroom get lots of opportunities to re-construct or re-invent their mathematics, the opportunities to develop misconceptions also abound. Because there is only one teacher but more than 30 students, the teacher needs some tools to check for student misconceptions. Dekker and Querelle (1998) recorded that cubes were mistaken for squares, Pythagoras theorem was remembered with a multiplication or “×” sign instead of a plus or “+” sign, and perimeters and areas were regularly mixed up. Possible assessment tools include:

- **Production items.** Students design a simple short-answer test. Of course answers should be provided as well (see discussion of this format), and all of the required content should be covered. The test could be graded but another, more rewarding possibility is to compose a class test using the student-designed items. Misconceptions will turn up and can then be discussed.
- **Student-generated items.** Students hand in a certain number of single-answer questions on the subject involved. These are used in a computer-based quiz for the whole group and are discussed afterwards.

As discussed previously in some detail, all assessment should result in feedback, and hopefully in feedback that goes far beyond grading a test. Feedback to the students is especially important when most students fail to solve a problem—a problem which the teacher thought fit nicely in the learning trajectory.

A very forceful way to get quality feedback is formed by the *two-stage task*. In this case, feedback on the first stage is given before the students start working on the second stage. In reality, this means that the teacher gets feedback from the students on how well the teacher’s feedback worked. Other information-rich formats include:

- **Oral questions** are asked when the topic is discussed in the classroom. In this case, the discourse is an assessment format.
- **Short quizzes**, sometimes consisting in part of one or more problems taken directly from student materials.
- **Homework** as an assessment format (if handled as described in our earlier section on homework).
- **Self-assessment**—preferable when working in small groups. Potential important difficulties will be dealt with in whole-class discussion.

Throughout the school year, the teacher will constantly evaluate the students’ individual progress and the progress of the whole classroom within the learning trajectory and thus evaluate the intended learning goals as benchmarks.

This ongoing and continuous process of formative assessment, coupled with the teachers' so-called intuitive feel for students' progress, completes the picture of the learning trajectory that the teacher builds. The problem of a strongly interactive classroom environment is that for teachers and students alike it is difficult to know whether or not they contribute to the group learning process and what they learned individually. Formats that may be helpful to evaluate students' progress include—

- **Discussions** with individual students about their understanding.
- **Observation** of students in groups and while working individually.
- **Extended-response open questions**, which require own productions, display of results for the whole group, or discussion by the whole class.
- **Peer-assessment** can be a tremendous help because students see the mistakes of their fellow students and then try to decide whether full or partial credit should be given for a certain solution.

After. At the end of a unit, a longer chapter, or the treatment of a cluster of connected concepts, the teacher wants to assess whether the students have reached the goals of the learning trajectory. This test has both formative and summative aspects depending of the place of this part of the curriculum in the whole curriculum. Different test formats are possible, while we see often that some formats with timed, written tests are the teacher's favorite—most likely because of their relatively ease of design and scoring and the limited possibility of feedback in a qualitative way

On Design

Assuming that the teacher has been able to construct a reasonable Hypothetical Learning Trajectory, the question is how to design in some more detail the assessment package that fits the corresponding trajectory. We need to take the following minimal variables into account:

- "Zoom" level.
- Content or big ideas.
- Level of competencies.
- Contexts.
- Formats.
- Feedback.

- Grading.
- Coherence and balance.

Keep in mind that we need to also consider the nine "Principles for Classroom Assessment." With these in mind, let us look briefly at each variable.

"Zoom" level. It is a good idea to start with a helicopter view of what the curriculum will look like over the whole year. This can be done in units of instruction or chapters of a textbook, or another clustering of ideas and concepts. The sequence needs to be logical and we need to pay attention to the longer lines of cognitive development that might be the results of this curriculum. For instance, it is quite possible that several concepts return to the attention months apart but at a higher and more formal level each time. If this is the case, it should be reflected in the assessment package.

From this higher or more global zoom-level, we can identify the end-of-year test and a number of end-of-cluster tests. For all practical purposes, most of these will be in a restricted-time written test format. But some of them need to be different if we want to stick to our principles. One of the tests could easily be a two-stage task. Easily, in the sense that apart from the design, these tests are relatively easy to administer. Or one of them could be a take-home task or a task to be performed in groups of two.

Content. The content can be covered in two distinct ways: cumulatively or by covering only the "unit" that has just been taught. The end-of-year test will always be cumulative, even over the years. The implication, of course, is that students should be informed about what the covered content will be far in advance. Three aspects need to be considered when looking at the content:

- How similar or dissimilar should the items be in relation to the student materials? (Are we testing reproduction or production and transfer?)
- What are the connections with other subjects and concepts? (Are we thinking big ideas, and to what extent?)
- What is the balance between more formal vs. informal mathematics?

This connects directly to the levels of mathematical competencies.

Competencies. All levels of competencies should be present in all tests but there should be more of the lower ones because they take little time. It is advisable to make an equal distribution over the three levels in terms of time rather than in terms of the number of items. It is a good idea to keep track of the distribution of the number of items on different levels and how the students perform relative to the levels in order to be able to give quality feedback both on both the classroom and the individual levels. Some support in finding higher level items and how to keep track of the distribution over the years can be found in the applications of technology to assessment. A modest contribution in this direction, consistent to a large extent with this framework is the assessment tool, “AssessMath!” (Cappo & de Lange, 1999), that offers not only a database of items but also a wide array of formats, the three levels of competencies, and the role of context.

Contexts. One should not fall into the tempting trap of designing items with the mathematics content in mind and then later adding a context. Nor should one take a context problem and simply change the context—many examples are available to show the disasters that can happen when these design strategies are followed.

The distance of the context to the students’ real world is one aspect that the teacher needs to get under control in the sense that each class environment can generate its own rules. (The assessment contract plays a crucial role here.) If the teacher takes the news from the different media as a regular discussion point in the lessons, one might expect a greater spread of distances in contexts than is the case with a teacher who facilitates contexts especially close to home and school. There is a clear trend for younger students to feel more confident with contexts closer to their life; to the surprise of some, however, context that relates to the environment, nature, and daily life in a wider sense can also motivate students very well, assuming challenging problems.

One should also be aware that in the more informal assessment formats, the freedom in context use is greater than when the tests are of a more summative character. This is because a teacher in a discussion will immediately note whether a certain context seems sensible to certain students (certain illnesses, for instance), and the teacher can correct for that on the spot.

The relevance of the context is another important point of consideration. If we take problem solving, and thus mathematization, as an important aspect of mathematics, then it is absolutely necessary to include first- and preferably second-order contexts on a regular basis. Quite often this means the use of more complex formats, although extended-response written questions offer good possibilities.

Finally, the point of “realness” of the context needs attention. Here again the teacher, in relation with the students, sets the boundaries. Although it seems sensible to stay as close to reality as possible without losing mathematical relevance, there are also good examples of not-so-real or not-so-authentic problems that have been excellent items for assessment, along with a few almost ridiculous “fantasy” problems that functioned within the classroom environment defined by teacher and students.

Formats. It is too simple to state that the choice of formats should be balanced. When we have a learning trajectory, one cannot just identify crucial assessment opportunities and choose a format. But in general, one can say certain things about the choice. From our previous discussion, it will be evident that discourse and observations are the continuous mode of operation together with homework. What is not trivial is that this has to be carried out with some structure and focus on key concepts. Also, some thought needs to be given to keeping track of the “scores” of the students on these formats. Again, technology can support us a bit with these issues: Handheld PDIs with dedicated software can be a considerable help in tracking the observations.

Regularly, some kind of restricted-time written test will be the backbone of our system. Assuming we have the right quality of items, there’s nothing wrong with that. These can vary in time from a short, 10-minute quiz to a two hour–long problem-solving task. We also need to stress more “constructive” formats of assessment with some mode of a two-stage task that can fit in well, although certainly not too often—maybe at most three times per year.

Part of the minimal requirements for a representative assessment system include that self-assessment be systemic and that homework should function, at least in part, as assessment.

It seems advisable to construct a path of incremental change in relation to more challenging assessment formats. The design, or even a proper judgment of open–open ended questions is already a very complex task. And although it seems sometimes easier to design a project task (like the question: “Is the pattern of green and red lights at this intersection the optimal in relation to the traffic flow?”), problems abound about such concerns as the execution, logistics, level of individuality, data sampling, and reporting in and out of school, not to mention how to cope with the different reports when the scoring, grading, and feedback are to be discussed. One should be careful not to fall into the hole of entering a very promising but extremely complex area of the assessment landscape without the prior experience of closely related formats.

Feedback. Feedback on the practical level relates directly to the assessment format. If we are in the discourse mode, feedback is absolutely necessary and instant. This can be challenging: Teachers need to react without much opportunity for reflection; thus they take the risk of not completely grasping the meaning of a student's remark. A sharp ear and a the eye of a hawk are the requirements for proper feedback during discourse, especially as we are viewing this as a part of the assessment system. And the better the picture of the hypothetical learning trajectory at the micro zoom level, the better this process will go.

At the other end we have the more traditional restricted-time written tests that usually only allow for a grade and some comments on errors or excellent strategies. But giving feedback this way has a large risk: It may never reach the student in an effective way. In order to learn from it, the students should internalize the feedback and reflect on it. This process sometimes becomes visible in a discussion but with written feedback on a test, there is no way to check this.

Making students aware of what feedback is and should be at all occasions, and in this way adapting the classroom environments to new socio-mathematical norms (Cobb, Yackel, & Wood, 1993), is a task that lays ahead of any teacher who wants to improve the teaching and learning process. This includes a discussion in a whole classroom of some of the students' answers and the feedback given by the teacher.

Grading. Students should be offered a clear and transparent picture of the scoring and grading for each of the assessment formats chosen. We have discussed in some detail how we can grade several of the more traditional formats. But we should also inform the students if we give certain marks for their role in a discussion, for doing excellent homework, or for suggesting a different and elegant strategy. At any given moment, the student should know which marks and grades are in the teacher's grading book. And a discussion about these should be possible, too.

Coherence and balance. Of course, we should not give the students multiple-choice throughout the year and then give a project at the end. Designing a hypothetical assessment trajectory that really is coherent and balanced, though it seems trivial, is very difficult given the variables that need to be balanced out: the competency levels, the content (from formal to informal), the contexts, the formats, and the possibilities for feedback. Teachers need concrete examples of a hypothetical assessment trajectory and imaginary curriculum for a whole year.

References

Aikenhead, G. (1997). A framework for reflecting on assessment and evaluation. In *Globalization of science education: International conference on science education* (pp. 195–199). Seoul, Korea: Korean Educational Development Institute.

Bagley, T., & Gallenberger, C. (1992). Implementing the standards: Assessing students' dispositions. Using journals to improve students' performance. *Mathematics Teacher, 85* (8), 660–663.

Beyer, A. (1993). Assessing students' performance using observations, reflections and other methods. In N. L. Webb & A. F. Coxford (Eds.). *Assessment in the mathematics classroom: 1993 yearbook* (pp. 111–120). Reston, VA: National Council of Teachers of Mathematics.

Biggs, J. (1998). Assessment and classroom learning: A role for summative assessment? *Assessment in Education: Principles, Policy and Practice*, *5,* 85–103.

Black, P. J. (1993). Assessment policy and public confidence: Comments on the BERA policy task group's article, "Assessment and the improvement of education." *The Curriculum Journal*, *4*, 421–427.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5* (1), 7–74.

Boertien, H., & de Lange, J. (1994). *The national option of TIMSS in The Netherlands*. Enschede, The Netherlands: University of Twente.

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H.-G. Steiner (Ed.), *Theory of mathematics education* (Occasional paper 54; pp. 110–119). Bielefeld, Germany: University of Bielefeld, Institut für Didaktik der Mathematik.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology, 79*, 474–482.

Cappo, M., & de Lange, J. (1999). AssessMath! [Computer software]. Santa Cruz, CA: Learning in Motion.

Carpenter, T. P., & Fennema, E. (1988). Research and cognitively guided instruction. In E. Fennema & T. P. Carpenter (Eds.), *Integrating research on teaching and learning mathematics* (pp. 2–17). University of Wisconsin–Madison, National Center for Research in Mathematics Education.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction*. Portsmouth, NH: Heinemann.

Clarke, D., Stephens, M., & Waywood, A. (1992). Communication and the learning of mathematics. In T. A. Romberg (Ed.), *Mathematics assessment and evaluation: Imperatives for mathematics educators* (pp. 184–212). Albany, NY: SUNY Press.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, *1*, 5–44.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education*, *22*, 3–29.

Cobb, P., Yackel, E., & Wood, T. (1993). Discourse, mathematical thinking, and classroom practice. In E. A. Forman, N. Minick, & C. A. Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development* (pp. 91–119). New York: Oxford University Press.

Cockroft, W. H. (1982). *Mathematics counts: Report of the committee of inquiry into the teaching of mathematics in school*. London, England: Her Majesty's Stationery Office (HMSO).

Collis, K. F., Romberg, T. A., & Jurdak, M. E. (1986). A technique for assessing mathematical problem-solving ability. *Journal for Research in Mathematics Education*, *17* (3), 206–221.

Cosgrove, M. M., & Schaverien, L. (1996). Children's conversations and learning science and technology. *International Journal of Science Education*, *18*, 105–116.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438–481.

de Haan, D., & Wijers, M. (Eds.). (2000). *Ten years of math A-lympiad: The real world mathematics team competition from 1990–2000*. Utrecht, The Netherlands: Freudenthal Institute.

Dekker, T. (1993). *Checklist toetsen met contexten* [A checklist for tests with contexts; Internal paper]. Utrecht, The Netherlands: Freudenthal Institute.

Dekker, T., & Querelle, N. (1998). [Internal publication]. Utrecht, The Netherlands: Freudenthal Institute.

Dekker, T., & Querelle, N. (in press). *Great assessment picture book*. Utrecht, The Netherlands: Freudenthal Institute.

de Lange, J. (1979). *Exponenten en Logaritmen* [Exponents and logarithms]. Utrecht, The Netherlands: Instituut Ontwikkeling Wiskundeonderwijs (IOWO; now Freudenthal Institute).

de Lange, J. (1987), *Mathematics: Insight and meaning*. Utrecht, The Netherlands: Vakgroep Onderzoek Wiskunde Onderwijs en Onderwijscomputercentrum (OW & OC).

de Lange, J. (1992). Assessing mathematical skills, understanding, and thinking. In R. Lesh & S. Lamon (Eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 195–214). Washington, DC: American Association for the Advancement of Science.

de Lange, J. (1994). Curriculum change: An American-Dutch perspective. In D. F. Robitaille, D. H. Wheeler, & C. Kieran (Eds.), *Selected lectures from the 7th international congress on mathematics education: Québec, 17–23 August 1992* (pp. 229–249). Quebec, Canada: Les Presses de l'Université Laval.

de Lange, J. (1995). Assessment: No change without problems. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 87–172). New York: SUNY Press.

de Lange, J., & van Reeuwijk, M. (1993). The tests. In J. de Lange, G. Burrill, T. A. Romberg, & M. van Reeuwijk, *Learning and testing Mathematics in Context: The case. Data visualization* (pp. 91–142). University of Wisconsin–Madison, National Center for Research in Mathematics Education.

Devlin, K. J. (1994). *Mathematics, the science of patterns: The search for order in life, mind, and the universe*. New York: Scientific American Library.

Duschl, R. D., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, *4*, 37–73.

Elawar, M. C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology*, *77*, 162–173.

Feijs, E., de Lange, J., Van Reeuwijk, M., Spence, M., & Brendefur, J. (1996). Looking at an angle. In National Center for Research in Mathematical Science Education & Freudenthal Institute (Eds.), *Mathematics in Context*. Chicago: Encyclopædia Britannica.

Fey, J. T. (1990). Quantity. In National Research Council, Mathematical Sciences Education Board, *On the shoulders of giants: New approaches to numeracy* (L. A. Steen, Ed.; pp. 61–94). Washington, DC: National Academy Press.

Freudenthal, H. (1973). *Mathematics as an educational task*. Utrecht, The Netherlands: Reidel.

Goldin, G. A. (1992). Toward an assessment framework for school mathematics. In R. Lesh & S. J. Lamon (Eds.), *Authentic assessment performance in school mathematics* (pp. 63–88). Washington, DC: American Association for the Advancement of Science Press.

Good, T. L., & Grouws, D. A. (1975). *Process product relationships in fourth grade mathematics classrooms* (Report for National Institute of Education). University of Missouri–Columbia.

Gravemeijer, K. P. E. (1994). *Developing realistic mathematics education*. Utrecht, The Netherlands: Freudenthal Institute, CD-β Press.

Griffiths, M. & Davies, G. (1993). Learning to learn: Action research from an equal opportunities perspective in a junior school. *British Educational Research Journal*, *19*, 43–58.

Gronlund, N. E. (1968). *Constructing achievement tests*. Englewood Cliffs, NJ: Prentice-Hall.

Grünbaum, B. (1985). Geometry strikes again. *Mathematics Magazine*, *58* (1), 12–18.

Hattie, J., & Jaeger, R. (1998). Assessment and classroom learning: A deductive approach. *Assessment in Education*, *5*, 111–122.

Howden, H. (1989). Teaching number sense. *Arithmetic Teacher*, *36* (6), 6–11.

Johnson, D. W., & Johnson, R. T. (1990). Co-operative learning and achievement. In S. Sharan (Ed.), *Co-operative learning: Theory and research* (pp. 23–27). New York: Praeger.

King, A. (1990). Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal*, *27*, 664–687.

King, A. (1992a). Comparison of self-questioning, summarizing, and note-taking review as strategies for learning from lectures. *American Educational Research Journal*, *29*, 303–323.

King, A. (1992b). Facilitating elaborative learning through guided student-generated questioning. *Educational Psychologist*, *27*, 111–126.

King, A. (1994). Autonomy and question asking: The role of personal control in guided student-generated questioning. *Learning and Individual Differences*, *6*, 163–185.

Kitchen, A. (1993). Coursework and its assessment in mechanics at ages 16–19. In J. de Lange, C. Keitel, I. Huntley, & M. Niss (Eds.), *Innovation in maths education by modelling and applications* (pp. 245–255). Chichester, UK: Ellis Horwood.

Kluger, A. N., & Denisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Koch, A., & Shulamith, G. E. (1991). Improvement of reading comprehension of physics texts by students' question formulation. *International Journal of Science Education, 13*, 473–485.

Krummheuer, G. (1995). The ethnography of argumentation. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning: Interaction in classroom cultures* (pp. 229–269). Mahwah, NJ: Erlbaum.

Kuiper, W., Bos, K., & Plomp, T. (1997). *TIMSS populatie 2, Nationaal Rapport* [TIMSS population 2, National Report]. Enschede, The Netherlands: University of Twente.

Lajoie, S. P. (1991, October). A framework for authentic assessment in mathematics. *NCRMSE Research Review, 1* (1), 6–12.

Lane, S. (1993). The conceptual framework for the development of a mathematics assessment instrument for QUASAR. *Educational Measurement: Issues and Practice, 12* (2), 16–23.

Lesh, R., & Lamon, S. J. (Eds.). (1992). *Assessment of authentic performance in school mathematics* (pp. 319–342). Washington, DC: American Association for the Advancement of Science Press.

Martin, M. O., & Kelly, D. L. (1996). *Third international mathematics and science study: Technical report*. Chestnut Hill, MA: Boston College.

Mathematical Sciences Education Board. (1990). *Reshaping school mathematics: A philosophy and framework of curriculum*. Washington, DC: National Academy Press.

Mathematical Sciences Education Board. (1990). *On the shoulders of giants: New approaches to numeracy* (L. A. Steen, Ed.). Washington, DC: National Academy Press.

Merrett, J., & Merrett, F. (1992). Classroom management for project work: An application of correspondence training. *Educational Studies, 18,* 3–10.

Meyer, K., & Woodruff, E. (1997). Consensually driven explanation in science teaching. *Science Education*, *80*, 173–192.

Meyer, M., Dekker, T., & Querelle, N. (2001) Context in mathematics curricula. *Mathematics teaching in the middle school, 9,* 522-527

Money, R., & Stephens, M. (1993). A meaningful grading scheme for assessing extended tasks. In N. L. Webb & A. F. Coxford (Eds.), NCTM Yearbook: *Assessment in the mathematics classroom*. (pp. 177–186). Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). Assessment standards for school mathematics. Reston, VA: Author.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*, 575–603.

Nielsen, A. C., Jr. (1987). The Nielsen panel: Statistics in marketing. In *Proceedings of the 1986 Making Statistics More Effective in Schools of Business (MSMESB) conference*. University of Chicago. Retrieved October 1, 2002, from MSMESB Web site: http://www.msmesb.org

Organization for Economic Cooperation and Development. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: OECD Publications.

Phye, G. D. (Ed.). (1997). Handbook of classroom assessment. London: Academic Press.

Powell, S. D., & Makin, M. (1994). Enabling pupils with learning difficulties to reflect on their own thinking. *British Educational Research Journal*, *20,* 579–593.

Pullin, D. C. (1993). Legal and ethical issues in mathematics assessment. In Mathematical Sciences Education Board & National Research Council, *Measuring what counts: A conceptual guide for mathematics assessment* (pp. 201–223). Washington, DC: National Academy Press.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4–13.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *TIMSS monograph no. 1: Curriculum frameworks for mathematics and science*. Vancouver, BC: Pacific Educational Press.

Rodrigues, S., & Bell, B. (1995). Chemically speaking: A description of student-teacher talk during chemistry lessons using and building on students’ experiences. *International Journal of Science Education, 17*, 797–809.

Roth, W.-M., & Roychoudhury, A. (1993). The concept map as a tool for the collaborative construction of knowledge: A microanalysis of high school physics students. *Journal of Research in Science Teaching, 30,* 503–534.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Santos, M., Driscoll, M., & Briars, D. (1993). The classroom assessment in mathematics network. In N. L. Webb & A. F. Coxford (Eds.), *Assessment in the mathematics classroom: 1993 yearbook* (pp. 220–228). Reston, VA: National Council of Teachers of Mathematics.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1996). *Splintered Vision: An investigation of U.S. science and mathematics education*. East Lansing: Michigan State University, U.S. National Research Center for the Third International Mathematics and Science Study.

Schwarz, J. L. (1992). The intellectual prices of secrecy in mathematics assessment. In R. Lesh & S. J. Lamon (Eds.), *Assessement of authentic performance in school mathematics* (pp. 427–438). Washington, DC: American Association for the Advancement of Science Press.

Senechal, M. (1990). Shape. In National Research Council, Mathematical Sciences Education Board, *On the shoulders of giants: New approaches to numeracy* (L. A. Steen, Ed.; pp. 139–181). Washington, DC: National Academy Press.

Shafer, M. C., & Foster, S. (1997). The changing face of assessment. *Principled Practice in Mathematics & Science Education, 1* (2), 1–8.

Shafer, M. C., & Romberg, T. (1999). Assessments in classrooms that promote understanding. In E. Fennema & T. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 159–184). Mahwah, NJ: Erlbaum.

Siero, F., & Van Oudenhoven, J. P. (1995). The effects of contingent feedback on perceived control and performance. *European Journal of Psychology of Education, 10*, 13–24.

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist point of view. *Journal for Research in Mathematics Education 16* (2), 114–115.

Smaling, A. (1992). Varieties of methodological intersubjectivity – The relations with qualitative and quantitative research, and with objectivity. *Quality and Quantity*, *26,* 169–180.

Stephens, M., & Money, R. (1993). New developments in senior secondary assessment in Australia. In M. Niss (Ed.), *Cases of assessment in mathematics education: An ICME study* (pp. 155–171). Dordrecht, The Netherlands: Kluwer Academic.

Stewart, I. (1990). Change. In National Research Council, Mathematical Sciences Education Board, *On the shoulders of giants: New approaches to numeracy* (L. A. Steen, Ed.; pp. 183–217). Washington, DC: National Academy Press.

Streefland, L. (1990). Free productions in teaching and learning mathematics. In K. Gravemeijer, M. van den Heuvel-Panhuizen, & L. Streefland, *Contexts, free productions, tests and geometry in realistic mathematics education* (pp. 33–52). Utrecht, The Netherlands: Vakgroep Onderzoek Wiskunde Onderwijs en Onderwijscomputercentrum (OW & OC).

Travers, K. J., & Westbury I. (Eds.). (1988). *International studies in educational achievement: Volume 1. The IEA study of Mathematics I. Analysis of mathematics curricula*.
p 22. New York: Pergamon Press.

Treffers, A. (1987). *Three dimensions: A model of goal and theory description in mathematics instruction – The Wiskobas Project*. Dordrecht, The Netherlands: Reidel.

Treffers, A., & Goffree, F. (1985). Rational analysis of realistic mathematics education – The Wiskobas Program. In L. Streefland, *Proceedings of the Ninth International Conference for the Psychology of Mathematics Education* (pp. 97–122). Utrecht, The Netherlands: Vakgroep Onderzoek Wiskunde Onderwijs en Onderwijscomputercentrum (OW & OC).

van den Brink, J. (1987). Children as arithmetic book authors. *For the Learning of Mathematics*, *7*, 44–48.

van den Brink, J. (1989). *Realistisch rekenonderwijs aan kinderen* [Realistic arithmetic education for young children; Doctoral thesis]. Utrecht, The Netherlands: Vakgroep Onderzoek Wiskunde Onderwijs en Onderwijscomputercentrum (OW & OC).

van den Heuvel-Panhuizen, M., & Vermeer, H. J. (1999). *Verschillen tussen meisjes en jongens bij het vak rekenen-wiskunde op de basisschool* [Differences between girls and boys in mathematics at primary school]. Utrecht, The Netherlands: CD-β Press.

van Reeuwijk, M., (1993) Learning and testing mathematics in context. In *Data visualization in the classroom.* National Center for Research in mathematical sciences education & Freudenthal Institute.

Verhage, H., & de Lange, J. (1997). Mathematics education and assessment. *Pythagoras, 42*, 14–20.

Webb, N. L. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis*, *17*, 239–261.

White, R. T. (1992). Implications of recent research on learning for curriculum and assessment, *Journal of Curriculum Studies*, *24*, 153–164.

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

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.

Wood, T. (1998). Alternative patterns of communication in mathematics classes: Funneling or focusing? In H. Steinbring, M. G. Bussi & A. Sierpinska (Eds.), *Language and communication in the classroom* (pp. 167–178). Reston, VA: National Council of Teachers of Mathematics.

Yackel, E. (1995). Children’s talk in inquiry mathematics classrooms. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning* (pp. 131–162). Mahwah, NJ: Erlbaum.

Endnotes

[1] The Mathematical Functional Expert Group is comprised of Jan de Lange (Chair), Raimondo Boletta, Sean Close, Maria Luisa Moreno, Mogens Niss, Kyang Mee Park, Thomas Romberg, and Peter Schuller.

[2] These levels have been developed over the last decade and find their origin at the end of the Eighties (de Lange, 1987), were made more explicit in the early Nineties (de Lange, 1992, 1994, 1995) and have been represented visually in a pyramid from then on with help from Dekker (Verhage & de Lange, 1997; Shafer & Foster, 1997).