INTRODUCTION: IMPORTANCE AND IMPACT

Public View

The science of mathematics education is not a very sexy science if publications in the popular media are taken as an indicator. In many countries one could also use the amount of money spent on this science as a good measure. But sometimes the science of mathematics education hits the media in a big way or becomes even a hot issue that will not disappear from the headlines for years to go.

Recently Germany delivered a prime example when the results of the PISA 2000 study were published. The date of the data release was December 5, 2001, and the cover of Der Spiegel (a magazine comparable to the American publications TIME and Newsweek) of 10 December 2001 tells it all: PISA-Study: The New Education Catastrophe. The 16-page-long article in small print makes the message more than clear: Compared internationally the German schools fail. The students are poor readers, are insufficient in arithmetic, and find problem solving completely beyond their possibilities. For Germans Pisa is not an Italian city with a leaning tower; Germans are suffering from the PISA “shock.”

The words and adjectives used in the article tell a story of their own: disaster, catastrophe, shock, hopeless, emergency plan. What makes the story extra difficult to accept for Germans is the fact that all neighbors of Germany score significantly better, and that the Germans are at the same level as the U.S.A. Embarrassing for the Germans, but the Americans do not see any reflection on the PISA results in their TIME and Newsweek: No media attention at all is probably the fair picture of how the results of PISA 2000 were received in the U.S.A. No news is good news seemed to be prevailing judgment here.

It is worth noting that, according to Der Spiegel and many others, PISA had a winner: Finland. The horse race was back, and whatever the organizers of international comparative studies have in mind, it seems inevitable that the horse race component always reaches the front pages. Finland has known it: Flocks of experts went to Finland to see the educational miracle in action. According to Der Spiegel the Finns are proud of their comprehensive system. Policy issues are quite often in the focus of the interest: the comprehensive system in Finland, the brilliance of the Japanese “average” student in an otherwise miserable system.

There seems to be no interest in the content of the study: There was no discussion of the items that were used, the competencies needed, the quality of the instruments, or the relation to the curricula. And there certainly was no discussion about the whole study being comparative, or norm-referenced, and the fact that it does not indicate any “absolute” quality. It is definitely not criterion referenced. This often-ignored fact can cause serious interpretation problems, especially when considering policy measures. According to the national Dutch report on PISA, the Netherlands was first in Mathematical Literacy (“The Winner”). In the official OECD publication one cannot find the Netherlands because the country did not
meet the statistical standards. The conclusion of the Dutch Ministry of Education was one of being proud of the system, the schools, and the teachers. The system was above all “effective” because the Netherlands is “underspending” on education in the OECD. At the same time math education experts (not excluding the present author) were trying to make a point about the very low level of math education in the Netherlands.

This paradox also runs among experts: In 1997 Lyle Jones argued in his William H. Angoff Memorial Lecture that NAEP (National Assessment of Educational Progress) has raised a high level of controversy because the NAEP report of 1996 shows only 18% of U.S. fourth graders “proficient” and only 2% “advanced” in math (Reese, Miller, Mazzeo, & Dossey, 1997) whereas recent results from TIMSS show that the average math performance for U.S. fourth graders is significantly above the international average (Mullis et al., 1997). Jones asked his audience, “When U.S. fourth-graders perform reasonably well in an international comparison, isn’t it unreasonable that only 20 percent are reported to be ‘proficient’?” Of course the answer is not the seemingly suggested “it is unreasonable,” but there can be very good reasons that this is completely normal—although not always easy to explain to policymakers or journalists.

Mathematics Education View

Not only are international comparative studies popular with the media and policymakers, they are also a subject for much controversy within the math education community itself. One of the key issues is about the preoccupation with competition (the horse race). Another is the assumptions on which comparative studies in mathematics are predicated. Keitel and Kilpatrick (1999) questioned in particular the treatment of the mathematics curriculum as unproblematic. The next is the assumption that a single test can give comparable measures of curriculum effects across countries. Jablonka (2003) discussed the fundamentally situated nature of mathematical literacy (like in PISA, but also in any assessment in context). She argued that contexts will be familiar to some students and not others. And thus any attempt to use a single instrument to assess mathematical literacy beyond the most local context would appear to be self-defeating. Cultural differences exist between countries and will play a major role. Freudenthal found this argument the most important one in discarding international comparative studies. Another problem is the fact that the single scale instrument does not mean that when two countries have (more or less) the same score, the consequences for policymakers are the same. As an example we analyzed the situation in two neighboring countries, the Netherlands and Belgium, who score almost the same on TIMSS but show big differences in curricula, didactics, culture, and policies (De Lange, 1997a). So if a policymaker wants to learn from these relatively high-scoring countries, where should she go? Probably to Finland.

Politics

One of the concerns that comes up after every large-scale assessment is whether or not such tests are really about measurement. Especially people that look more at the policy side of such studies argue that performance measures are not really about measurement, but about political communication. Political scientists have expressed this view on many occasions, and there is certainly reason to take this view seriously.

Clearly politicians make up their own interpretation of the facts, to their own liking. This can be observed in most countries, but a precondition to make this observation is knowing the facts in some detail. This is most often not the case and beyond what journalists are interested in. If one looks at the large number of national reports that tend to come out parallel with the international report, it is amazing to see how the data were selected. It takes some effort, but it is worth it: large-scale international assessments are indeed about politics, about policy, about political communication.

In the discussions, whether scientific or more public, these politics play a large role. The chain of reasoning is quite often influenced by political views. This may lead to chains of reasoning in the arguments like the ones mentioned earlier that are open to some kind of criticism because of the weak reasoning, not being based on the actual data, or taking the data too seriously. Quite often the points made are mere opinions, not seldom without proper knowledge of the methodology of the study involved. Where appropriate I will deal with a number of controversial issues, hopefully without falling into the trap of weak causal reasoning and mere opinion.
from giving a definition at all, and those who even try come up with rather unsuccessful propositions. In the publication *Reinventing Assessment* (1998) Randy Elliott Bennett proposed the following:

Large-scale educational assessment consists of those tests administered to sizable numbers of people for such purposes as placement, course credit, graduation, educational admissions, and school accountability. It includes group-administered, standardized tests used most often in the secondary through post secondary years. (p. 1)

Bennett is clever enough to use words as “for such purposes as,” leaving room for other purposes. The second sentence seems a bit redundant. But there are few definitions around, so I start with this one and look to improve on this definition, if only for this chapter of the book.

The National Research Council (NRC) publication *Knowing What Students Know* (2001) saw two components in educational assessment: The first context in which educational assessment occurs is the classroom. Here assessment is used by teachers and students mainly to assist learning, but also to gauge students’ summative achievement over the longer term. Second is large-scale assessment, used by policymakers and educational leaders to evaluate programs or obtain information about whether individual students have met learning goals.

The committee adds that large-scale, standardized assessments can communicate across time and place, but by so constraining the content and timelines of the message, they often have limited utility in the classroom. Thus the contrast between classroom and large-scale assessments arises from the different purposes they serve and contexts in which they are used. And large-scale assessments are further removed from instruction but can still benefit learning if well designed and properly used. Finally, large-scale assessments not only serve as a means for reporting on student’s achievement but also reflect aspects of academic competence societies consider worthy of recognition and reward. Thus large-scale assessments can provide worthwhile targets for educators and students to pursue.

The problem with the definition and description just given is that it seems to fit seamlessly to the American situation but ignores the realities in many other countries. International comparative studies may be blooming; international cooperation and recognition still has a long way to go.

Paul Barton made the point of the relatively isolated position in the international large-scale assessment landscape clear in his 1999 publication *Too Much Testing*. In one area there is very little (large-scale) testing whatsoever in the United States, and a lot in other developed countries: extensive examinations at the exit point for secondary education. In effect, in many countries large-scale testing is quite often identified with these exit-examinations, which are often very closely related to the curriculum. This has as very desirable side-effect that the distinction between classroom testing and large-scale testing, as clearly identified by the committee that is responsible for *Knowing What Students Know*, becomes a fluent connection, especially in those countries where a classroom assessment is part of the large-scale examination. If asked for a definition of large-scale assessments, these countries would probably suggest quite different definitions than the ones above.

### Large-Scale Assessment as Entrance/Exit

The isolation of the United States in this area has been properly phrased by Eckstein and Noah (1993):

The United States is unique among the countries we have studied (U.S., China, Japan, Germany, England, Wales, France, Sweden, and the former Soviet Union) in having no coordinated, public, national system for assessing student achievement at the end of secondary school. (p. 238)

A more recent study (Stevenson & Lee, 1997) focused on entrance and exit examinations in Japan, Germany, France, and the United Kingdom. The authors observed:

Entrance and exit examinations in these countries are based on a curriculum established by ministries of education at the local, regional or national level. Rather than imposing some arbitrarily defined standards of achievement, the examinations are closely tied to what students have studied in high school. Because teachers are aware of what students are expected to know in examinations, it becomes their responsibility to equip students with the information and skills needed to pass the examination. (p. 47)

This last observation is important because it points to the fact that large-scale assessments need not to be completely different from classroom assessment, as a matter of principle. On the contrary: For students in many countries the exit examination fits seamlessly with their classroom assessment. This fact surely complicates the task to come up with a fitting definition of large-scale assessment. Especially because of the fact that these exit examinations also do not always fit the pencil-and-paper format that people tend to link to
large-scale assessments, not to mention the emphasis on the multiple-choice format.

As Stevenson and Lee observed, the exit examinations typically include open-ended questions that require organization and application of knowledge, and oral examinations that require students to express themselves verbally. Thus exist large-scale assessments that do have more than one measure, use different formats, and try to measure different competencies.

Large-Scale Assessment As Proficiency

Finally I need to mention the developments in large-scale assessments that are the results of information technology. Several examples in existence now illustrate how technology can help infuse ongoing formative assessment into the learning process. The intelligent tutors form a category that makes a difference. Knowing What Students Know observes that these intelligent tutoring systems are powerful examples of the use of cognitively based classroom assessment tools blended with instruction. The systems have in common that when the student makes a mistake the system offers effective remediation.

In relation to large-scale assessment caution is necessary when such intelligent tutoring systems are being used nationwide or even made compulsory—as is being considered in the Netherlands. If all schools in a country or state are using a (diagnostic) student tutoring system that also keeps track of student progress over time (maybe over a somewhat narrow band of competencies), can one then conclude that this is large-scale assessment, or should one follow Knowing What Students Know and classify this as formative or classroom assessment, which would make it distinct from large-scale assessment?

Definition

Clearly trying to formulate a definition of large-scale assessment, even if only for this paper, is not a simple task. For the scope of this article I will use the following formulation:

Assessment is a process by which educators use student’s responses to specially created or naturally occurring stimuli to draw inferences about the students’ knowledge and skills (Popham, 2000). Large-scale assessments are assessment instruments intended to be administered to a large number of students for a wide variety of reasons. Three broad purposes can be identified: to assist learning, to measure individual achievement, and to evaluate programs (NRC, 2001). They can be delivered by different media, be individual or teamwork, be formative or summative, use a variety of formats, can be aligned with a curriculum or not, can be standards based or not, may operationalize competencies (in contrast with a content-oriented curriculum), and can be part of a comprehensive balanced assessment system or not.

STRUCTURING THE LARGE-SCALE ASSESSMENTS DISCUSSION

The definition gives way to a variety of ways to structure the discussion. One class of assessments seems a very natural one, and it is also a very visible one: the class of international comparative studies. In mathematics this class is usually restricted to the IEA (International Association for the Evaluation of Educational Achievement) studies like SIMSS, TIMSS, TIMMS-R and the recent TIMSS—with the meaning of the T changed from Third to Trends—and the OECD (Organisation for Economic Co-operation and Development) studies like PISA 2000, PISA 2003, and so on. But there are numerous other studies, especially in the area of adult mathematical literacy. I will discuss the TIMSS and PISA studies in some detail and make a relation with some of the other studies. All these studies have as a common denominator the international comparative component.

At the national level we run already into problems because of the large spread in formats and uses of large-scale assessments. In many countries exit examinations would be the first assessment that people think of. In others, most notably the United States, the one that comes closest to a national test is the NAEP (National Assessment of Educational Progress) project. This is a national survey intended to provide policymakers and the public with information about academic achievement of students across the nation. Its goal is of a similar nature as the IEA and OECD studies, albeit on a national level. Trend studies at national level include for instance the PPON (Periodieke Peiling van het Onderwijsniveau, Periodic Sounding of the Education Level) study in the Netherlands at elementary level. Of course aside from the NAEP study that has a reputation in regard to the use of standardized tests and the industry that comes with them.

I will discuss in some detail at the national level both national examinations and national trend studies, and to some extent, standardized tests both more traditional and more innovative.

In some countries it is relevant to go down one more level: to that of the states. Of course this is especially relevant in the United States, but other countries have state tests as well, for instance Australia, where Victoria has been very visible when introducing their quite innovative and balanced assessment system.
will not discuss some consequences of the No Child Left Behind Act in the United States as this issue is addressed in Wilson’s chapter (this volume).

It seems reasonable to devote a discussion on the future of large-scale assessment in light of evolving information technology. The trend towards more problem-solving skills and more complex competencies like knowledge organization, problem representation, strategy use, and metacognition in mathematics education seems very challenging, as the underlying assessment principles in these systems are traditional.

The same trend towards more complex competencies in mathematics education also leads to renewed interest in teamwork in real-world problem solving. I will mention some examples that have been in place for quite some time and certainly can be regarded as large-scale assessments.

This leaves a structure with three key tenets: international, national, and state; technology; and assessing real-world problem solving. Of course this is no real taxonomy by any measure but seems a way of clustering that might help the reader to grasp this complex subject. Within each of the five clusters I will discuss some examples, problems and benefits, influence or results, and some suggestions for the future.

---

**REVIEW OF ISSUES IN LARGE-SCALE ASSESSMENTS**

**International Comparative Studies**

**TIMSS and PISA**

The IEA (International Association for the Evaluation of Educational Achievement) has conducted studies of cross-national studies since 1959. TIMSS, originally named the Third International Mathematics and Science Study, from 2003 onwards Trends in International Mathematics and Science Study, is the most recent one of the IEA series to measure trends in students’ mathematics and science achievement. Additionally the countries that participate are provided with a resource for interpreting the achievement results and to track changes in instructional practices. TIMSS asks students, their teachers, and their school principals to complete questionnaires about the contexts for learning mathematics and science. TIMSS 2003 assesses the mathematics and science achievement of children in two target populations. One target population, sometimes referred to as Population 1, includes children Ages 9 and 10. It is defined as “the upper of the two adjacent grades with the most 9-year-olds.” In most countries this is the fourth grade. Population 2 includes children Ages 13 and 14 and is defined as “the upper of the two adjacent grades with the most 13-year-olds.” In most countries this is eighth grade (Mullis et al., 2001b). By assessing these grades using the same target populations as in 1995 and 1999, TIMSS 2003 will provide trend data at three points over an 8-year period. In addition, TIMSS data will complement IEA’s Progress in International Reading Literacy Study (PIRLS) being conducted at the fourth grade. According to the authors of the TIMSS Assessment Frameworks and Specifications 2003 (Mullis et al., 2001b), countries participating in PIRLS and TIMSS will have information at regular intervals about how well their students read and what they know and can do in mathematics and science. TIMSS also complements another international study of student achievement, the OECD’s Programme for International Student Achievement (PISA), which assesses the mathematics and science literacy of 15-year-olds. (p. 6)

According to the OECD (2000):

The OECD’s Programme for International Student Assessment (PISA) is a collaborative effort among the member countries of the OECD to measure how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today’s knowledge societies. The assessment is forward looking, focusing on young people’s ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum. This orientation reflects a change in the goals and objectives of curricula themselves, which are increasingly concerned with what students can do with what they learn at school, and not merely whether they have learned it. The term ‘literacy’ is used to encapsulate this broader conception of knowledge and skills. (p. 12)

The first PISA survey was carried out in 2000 in 32 countries, including 28 OECD member countries. Another 13 countries completed PISA 2000 in 2002, and from PISA 2003 onwards more than 45 countries will participate “representing more than one third of the world population” (p. 22). PISA 2000 surveyed reading literacy, mathematical literacy, and scientific literacy, with the primary focus on reading. In 2003 the main focus was on mathematical literacy (published in 2004), and in 2006 scientific literacy will be highlighted.

It will be clear that both studies have a lot of similarities resulting in improper identification of the two series of studies in the media, which is undesirable and confusing. But the descriptions of the organizations that are responsible show that they both claim similar relevance for the studies. It seems safe to say that, for both TIMSS and
PISA, countries participating in their studies will be given information at regular intervals about how well their students read and what they know and can do in mathematics and science. Both studies provide such information and they provide it, methodologically speaking, in a very similar way (based on Item Response Theory, IRT). Even the reporting tables in the respective reports look very similar. We refer to the Technical Aspects Reports for both studies for the methodological aspects (OECD, 2001; Yamamoto & Kulick, 2000).

If there is a problem that both studies share it is the design of the measuring instrument in relation to the validity of the outcomes. Traditionally, validity concerns associated with tests have centered about test content, meaning how the subject domain has been sampled. Typically evidence is collected through expert appraisal of alignment between the content of the assessment tasks and the curriculum standards (in case of TIMSS) and “subject matter” assessment framework (PISA). Nowadays, empirical data are often used before an item is included in a test.

Traditionally validation emphasized consistency with other measures, as well as the search for indirect indicators that can show this consistency statistically. More recently is the recognition that these data should be supplemented with evidence of the cognitive or substantive aspect of validity (Linn, Baker, & Dunbar, 1991; Messick, 1993). Or as Knowing What Student Knows summarized: “The trustworthiness of the interpretation of test scores should rest in part on empirical evidence that the assessment tasks actually tap the intended cognitive processes.” (p. 7). One method to do this is a protocol analysis in which students are asked to think aloud as they solve problems; another is an analysis of reasons in which students are asked to provide rationales for their responses; and a third method is an analysis of errors in which one draws inferences about processes from incorrect procedures, concepts, or representations of problems. Although some of these methods are applied only after the test is administered, there is a trend that large-scale assessments like TIMSS and PISA use these methods as well. The use of cognitive laboratories to gauge whether students respond to the items in ways the developers intended has become a new instrument in the developmental process. The use of double-digit coding is another sign of interest in the process of problem solving instead of just judging whether an answer is incorrect or correct. A “correct” or “partly correct” score given not only to each work of the student, but also to which strategy was used or where in the process the students “lost track.”

A point of critique about the validity that remains to be discussed is the way of assessing, and more specifically the choice of item formats. Clearly tension exists between the frameworks (what educators want to measure) and the choice of item format (what educators can afford in a practical and economical way).

Frameworks

The original Curriculum Frameworks for Mathematics and Science of TIMSS did not mention item formats (Robitaille et al., 1993). In practice TIMSS relied heavily on the multiple-choice format. That the alignment between the framework and test items leaves some room for improvement may be illustrated by the example shown in Figure 25.1.

![Figure 25.1 TIMSS-item, 1996 (Beaton et al., 1996, p. 63).](image)

This item is from the Population 2 cycle of 1996 (Beaton et al., 1996, p. 63). This rather straightforward problem, in multiple-choice format, is intended to show the use of “complex (mathematical) procedures.” If this is true it shows clearly that TIMSS is not criterion referenced but very normative and that the alignment between framework and test items leaves some room for improvement, especially given the ambitious performance expectations.

The more recent TIMSS Assessment Frameworks and Specifications 2003 (Mullis et al., 2001b) made major improvements by including a “Question Types and Scoring Procedures” chapter. This made clear that the instrument heavily relies on the multiple-choice formats. The Framework left it to the reader to do the arithmetic to find out how many questions still are multiple-choice:
Two question formats will be used—multiple-choice and constructed response. Each multiple-choice question will be worth one point. Constructed-response questions generally will be worth one, two or three points. However, extended-problem solving and inquiry-items may be worth up to five points. Up to two-thirds of the total number of points will come from multiple choice items. (p. 88)

The wording in the PISA Frameworks is different, and this study seems to be more open to open questions. The PISA 2000 Framework (OECD, 1999) explicitly mentioned the somewhat problematic status of multiple-choice items:

Travers and Westbury (1989) state when discussing the second IEA mathematics study (SIMSS) that: “The construction and selection of multiple choice items was not difficult for the lower levels of cognitive behavior—computation and comprehension.” But they continue: “difficulties were presented at the higher levels.” There is place for the use of multiple-choice formats but only to a limited extend and only for the lowest goals and learning outcomes. For any higher-order goals and more complex processes, other test formats should be preferred, the simplest being open questions. (p. 54)

In practice, PISA 2000 ended up with around 50% of the items being multiple-choice. What is interesting in the 2000 Framework is appendix 2: “Considerations for future survey cycles of OECD/PISA.” In fact it is a plea for a balanced assessment instrument in which students can better show what they can do in a more constructive way: extended-response essay tasks, oral tasks, two-stage tasks, and production items.

For the PISA 2003 Framework, the test instrument consisted of “a combination of items with open-constructed response types, closed constructed-response types and multiple-choice types. About equal numbers of each of these types will be used in constructing the test instruments for OECD/PISA 2003.” (p. 50).

In practice the amount of multiple-choice items is higher than indicated in the Framework. None of the recommendations of the appendix of 2000 were realized.

The validity of the test instrument remains a complex issue. It goes without saying that there is an inherent tension between the traditional choice of item formats, usually with very restricted time (1–2 minutes per item), and the rather ambitious definitions of what the instrument is intended to measure. As Barton (1999) stated:

In the 1990s, there have been constructive attempts to improve the testing enterprise. Serious efforts have been made to broaden tests beyond multiple-choice questions, and to include open-ended questions, performance assessments, and portfolios. However, the assessment reform movement has been slowed over issues of reliability and measurement error. (p. 4)

But not only the concern about “errors” plays an important role in relying so much on multiple-choice, it is also an economic issue: Many countries participating in these large cooperative studies are unwilling or unable to fund much more expensive multiple marker studies, even if such studies have demonstrated their efficacy.

PISA 2003 also had a problem-solving component. Many of the items (Figure 25.2) would fit the

![Figure 25.2 PISA-item 2003 (OECD, 2004, pp. 70–72).](image-url)
mathematics Framework, and given the fact that the instrument for problem solving (PS) had much more open “constructive” items, a study relating the math items and the PS items could be very helpful in advancing the discussion on item instruments and their restrictions in large-scale international studies.

This example (see Figure 25.2) from PISA Problem Solving shows not only the relation to mathematics, but also that students are given the opportunity to engage in a more open problem, like question 2. According to the PISA report on problem solving

[T]he items for problem solving give a first glimpse of what students can do when asked to use their total accumulated knowledge and skills to solve problems in authentic situations that are not associated with a single part of the school curriculum. (p. 18)

One can easily argue that this is always the case in a curriculum: For mathematical literacy, mathematics as taught at school will not suffice. Students need to read, need to interpret tables and graphs (seen by many as belonging to reading literacy), and, indeed, need problem-solving strategies. But seen from the perspective of promising developments on item formats and item quality, the problem-solving component of PISA is interesting, at least. And if TIMSS implements their intent to “place more emphasis on questions and tasks that offer better insights into students’ analytical, problem-solving, and inquiry skills and capabilities,” innovation in large-scale assessments could materialize.

**PISA versus TIMSS**

The main differences between TIMSS and PISA seem to be the following:

- curricular emphasis for TIMSS versus functional aspect (literacy) for PISA;
- grade-specific structure of TIMSS versus age-specific structure of PISA.

TIMSS uses the curriculum as the major organizational aspect. The TIMSS curriculum model (see Figure 25.3) has three aspects: the intended curriculum, the implemented curriculum, and the attained curriculum. These represent, respectively, the mathematics and science intended for students to learn, and how the education system should be organized to facilitate this learning: what is actually taught in the classrooms, who teaches it, and how is it taught; and finally, what it is that students have learned, and what they think about those subjects.

International curricular diversity was a serious point of concern to the TIMSS study. The goal was to develop an international test that would be equally fair to all participating countries. Therefore subject-matter specialists from all countries were consulted and asked to contribute to the process of test development (Garden & Orpwood, 1996). Most countries participating in TIMSS had an intended mathematics curriculum that matched with more than 90% of the items (Beaton et al., 1996; Kuiper, Bos, & Plomp, 1997). The outliers were the United States and Hungary with 100% matching, and the Netherlands, with 71% matching.

Insiders have discussed the procedure and its validity of this equally unfair analysis. The question not satisfactorily answered is how the mathematics education communities in the different countries were involved, and how representative they were. But if these numbers are accepted, in this context it is worth looking at the minimal matching result of the Netherlands.

As Vos (2002) observed, it was expected that students of other countries would outperform Dutch students. However, contrary to expectations, in 1995 Dutch Grade 8 students performed well on the TIMSS test. Their score was significantly above the international average, just below the four Asian top-scoring countries (Beaton et al., 1996). After some additional research by Beaton it was concluded that somehow the Dutch students were knowledgeable about the 29% of test items that were remote from their intended curriculum. In the end it was concluded that the students had the abilities for transfer of their knowledge and skills to items that did not match with their intended curriculum (Kuiper, Bos, & Plomp, 2000). As Heubert and Hauser (1999) explained it can be very appropriate to test students on material they have not been taught, if the test is used to find out whether the schools are doing their job.

PISA takes this point even further: It is based on a dynamic model of lifelong learning in which new knowledge and skills necessary for successful adaptation to a changing world are continuously acquired.
throughout life. It focuses on young people’s ability to use their knowledge and skills to meet real-life challenges, rather than on the extent to which they have mastered a specific school curriculum.

Given the observation of the TIMSS researcher about the transfer skills of Dutch students, combined with goals of PISA, it may come as no surprise that Dutch students outperformed all other countries in mathematics but did not appear in the official report because “the response rate was too low” (OECD, 2001, p. 236).

The two different approaches can both be critiqued: What does it mean that the Netherlands scored so high with the minimal relation with its curriculum? What does it mean if PISA will not constrain itself to any national curricula? It is clearly not true, as Clarke (2003) suggested as a proposition and as Freudenthal supported in 1975 already, that international studies of student achievement may be unintentionally measuring little more than the degree of alignment between the test instrument and the curriculum. What it does measure is still a question open to interpretation.

Another indication that shows how difficult it is to make statements that go beyond well-intended opinions can be found in the observation of Westbury in 1992, in relation to SIMS, when he observed that the lower achievement of the United States is the result of curricula that are not as well matched to the SIMS test as are the curricula of Japan. But in TIMSS the match was 100% (see earlier), and still the United States did not perform very well.

The Real World

Another concern that the math education community has is about the implicit imposition of a global mathematics curriculum. Keitel and Kilpatrick (1999) have problematized the assumptions on which international comparative studies are predicated. In particular, they questioned the treatment of the mathematics curriculum as unproblematic and the associated assumption that a single test can give comparable measures of curriculum effects across nations. They further suggested that the concept of an “idealized international curriculum” lies behind even the most sophisticated research designs. The problem with this judgment is that no one knows which part of the learning process as demonstrated in the tests can be attributed to schooling and the curriculum.

Without defining what a curriculum is, it is difficult to have a proper discussion. And the question can even be stated whether the whole question about analyzing curricula is a relevant one, given the outcomes of the Netherlands in TIMSS and the PISA conception of learning in society with schools being one component of the learning process. This of course is a correct observation.

Keitel and Kilpatrick’s criticism seems to focus on the underlying assumption of the international “common” curriculum. The TIMSS model seems to invite this kind of criticism, but PISA much less, if at all. However, the problem of creating such an international curriculum seems very much alive as PISA expands to about 70 countries in 2006. Many countries will look at the PISA items as something worth teaching, if alone to get better results next time. This is indeed a very serious problem as many politicians lack the knowledge and skills to realize that PISA only measures part of mathematics, more precisely: literacy. This has led to communication problems in several countries, and by many mathematicians, who tend to see PISA mathematics as representing mathematics as a discipline.

Another concern expressed in Clarke’s (2003) article “International Comparative Research in Mathematics Education” is that differences in societies—like Ethiopia and the United States—automatically lead to differences in skills required for effective participation in these societies. This point seems very straightforward, but many of the competencies identified in the PISA Framework or item-collection intend to operationalize competencies that are of some relevance in any society. To state that “to assess the extent to which Ethiopian students possess the mathematical skills required for effective participation in American society seems both futile and uninformative” (p. 153) seems futile itself.

Another point that should be discussed, as it refers to an important aspect given the trend towards more “real-world” problem solving, is the link between mathematical literacy and social practice. Jablonka (2003) argued that even within a single school system there will be contexts familiar to some students and not others. Any attempt to use a single instrument to assess mathematical literacy beyond the most local of contexts would appear to be self-defeating. Again the chain of reasoning is not evidence based. The choice of context in relation to item construction and validation is a very complex one, as shown by many examples from previous experiments in designing curricular materials (e.g., Feijs, 2005). These experiments indicate that one cannot say anything firm about the relationship “context familiarity: success rate.” As an example, due to the fear of experts that an item for an international study about the tides of an ocean should not be included because of the participation of land-locked countries, the item was rejected forcefully by the field trials. The influence of contexts should be studied much more systematically than is presently the case, and we researchers should
refrain from strong statements that have proven to be of disputable quality until we have firmer evidence. For the present studies all countries had to agree on the items, including the role of the context.

In a working paper (Nohara, 2001) from the (U.S.) National Center for Education Statistics in which TIMSS-R(repeat), PISA, and NAEP were compared, remarks were made about the inclusion of real-life situations, defined as items not presented strictly in the language of mathematics. According to the paper, this characteristic is significant because connecting mathematics to the world outside of school is a major goal of many mathematics education reform initiatives. It is also significant because it means that students have to choose for themselves the operations and solutions most appropriate for the problem and figure out how they relate to the information provided, thereby “adding to the difficulty of the item.” One may comment here that clearly such contexts require different cognitive demands from the students, but whether one can conclude that this leads to more difficult items is a matter to be studied in more detail, and especially PISA 2003 might offer some direction. The reason for that somewhat optimistic expectation can be found in the same working paper when the author stated that both TIMSS-R and PISA (and NAEP) contained many items situated in real-world contexts: 44% of items for TIMSS-R, 48% for NAEP, and 97% for PISA. The author also noted that in reviewing PISA items, panel members noted that several items in real-life situations presented students with significantly more challenges than others. These contexts were highly unique, that is, not typically encountered in mathematics instruction or textbooks, or required significantly more thought regarding how the nature of the context affects the mathematics involved in the problem. This type of item can be contrasted with the standard word problems typically used in mathematics classes, which can be described as “proxies for reality.” Only a few such items can be found in TIMSS-R or NAEP.

Meaning of Measures

Via the real or perceived influence of context I now come back to the relative meaning of outcomes of large-scale international assessments. The use of context is or could be one factor contributing to the relative difficulty of the assessments. Other factors contributing to the relative difficulty are (Nohara, 2001) the number of items that use the extended-response format, the number that require multistep reasoning, and the amount of computation as “the presence of a computation requirement does not present an additional degree of difficulty as it would in an item classified in another content strand” (p. 14).

Looking only at these four factors PISA appears to be the most difficult: It has the highest percentages in all four categories.

If it is true that one can attribute item difficulty to these four factors and conclude that PISA is the most difficult large-scale international assessment, it might give some additional insight into what one can say about the absolute level in relation to the relative level. In addition to the TIMSS item presented before in this contribution it might be interesting to look at a PISA item that was considered “too difficult” to be included in the final item collection.

In my opinion, Figure 25.4 shows, as part of a wider pattern if one analyzes the items of large-scale assessments, that the “cognitive competency level” operationalized by these tests leaves lots of room for improvement. This implies that the “results” of such studies should be interpreted much more carefully than is usually the case. The horse race outcomes will be differently interpreted if policymakers understand that the metaphor itself is quite inappropriate if the horses do not have to meet a certain qualifying time over a fixed distance but merely are allowed to underperform.

Indeed, as Clarke (2003) summarized the feeling of most of the mathematics education community, it is “the current preoccupation with competition” that always seems to make it into the headlines of the media and always seems to support the present politicians and policymakers. The “Mathematics Olympics” under the banner of PISA 2000 made a media uproar in Germany, as did TIMSS in the United States in 1997.

What is missing quite often is a discussion by somewhat informed mathematics educators about what is being measured from a “content” view and what it means. The quoted working paper from NCES is a laudable attempt to shed some light on this aspect of international comparative studies. Keitel and Kilpatrick (1999) observed correctly the striking fact that almost all of the people with the primary responsibility for conducting the study (TIMSS) have been empirical researchers in education, psychometricians, or experts in data processing. The problem of the content of mathematics education, Keitel and Kilpatrick (1999) argued, has been dealt with as no more than a technical question:

The most monumental and most accurately treated feature of the studies has been the handling of data once they have been collected. Particularly notable has been the fashion in which problems of methodological validity, reliability and quality have been re-
I agree with this point and also note that in relation with this point several countries carried out national options, in part because content-related matters gave concern about the validity of the assessment as appropriate for the specific country. An impressive effort in relation to PISA 2000 was carried out by Germany. Although the report (Baumert et al., 2001) was in general positive about PISA 2000, the Germans opted for a National Option because of the fact that there was a need (in Germany) for more “inner-mathematical problems,” for more “basic-skills problems” and “facts.” This was done to make the PISA test “closer” to the German curriculum.

A similar action, but then in the opposite direction, was taken by the Netherlands in reaction to TIMSS. As I reported in 1997, the Netherlands (here meaning some influential math educators) concluded that TIMSS did not fit in a satisfying way to the intended/perceived/implemented curriculum, not on the basis of the earlier mentioned relatively low “alignment-quotient,” but because of too many multiple-choice items operationalizing too many lower
level thinking skills (De Lange, 1997). Therefore a National Dutch Option was carried out. The international test and the national option had an intersection of carefully selected “anchor items.” Methodologically speaking, the results on the anchor items within the national option or in the international option should be the same. In reality the students did much better on the TIMSS items when they were part of the national option. This outcome supports the arguments to first think what is really being measured and with which instruments before making policy statements regarding the horse race. In fact, the TIMSS score in reality was 541 in Grade 8. But based on the national option as a more reliable measurement, the same TIMSS items would lead to a score of 585. The only explanation brought up by the national experts, including the IEA group, was that the “national” context made children feel better and more motivated. Or, to put it slightly differently, perhaps the fact that the anchor items were culturally correctly embedded made students perform better.

The content is a point of concern that receives relatively little attention, contrary to the question of the influence of the different cultures. It goes without saying that we measure “cultures.” Bracey’s (1997) observation that “international achievement differences become similarities if we compare the performance of students of Asian cultural affiliation, no matter what school system they attend” is true to a certain extent. One does not have to be a rocket

Figure 25.5  Graph showing countries participating in TIMSS 1996 (De Lange, 1997).
scientist, let alone a math education researcher, to see a certain socio-economic-cultural structure in the listing of the horses of TIMSS.

In my opinion one can see a certain clustering of countries, starting with the Pacific Rim countries (referred to as “Asian Pacific Tiger” countries in Figure 25.5), via small European countries, to France, to (former) commonwealth countries, to Germany and the United States, to Scandinavia, and finally Southern Europe (De Lange, 1997). The mathematics education community has the responsibility not only to contribute to such studies (and to influence them in a positive way), but also to critique them when necessary, to draw attention to the content part of such studies, and to inform policymakers what the study might mean beyond the horse race.

**Missing Content**

In a review of the PISA 2000 report Bonnet (2002) underscored the lack of interest—in content matters—in trying to understand in pedagogical terms how and why pupils gave the responses, or absence of responses, that they actually gave. Instead, he observed, the focus of the report is on trying to analyze results in the light of socioeconomic data and the organization of education systems and schools. One would have wished for more space to be devoted to trying to understand what happens in the classroom in terms of the teaching and learning process. This exactly is what the title of the earlier mentioned IEA/STUDY promises but fails to materialize. Both studies, TIMSS and PISA, should and could do much better in this respect. It is difficult to point to the real reason behind this phenomenon. Keitel and Kilpatrick (1999) suggested looking at who conducts the study in order to better understand the results.

Taking into account that PISA is organized by the OECD, Bonnet (2002) observed that it is illuminating that chapter 3 of the PISA 2000 report ends at page 91 with tables showing the relationships between the average country performance in the three combined scales (reading, mathematics, and science) and the GDP (Gross Domestic Product) per capita. The true worth of such correlations becomes apparent with the inspiring conclusions that “countries with higher income per capita tend to perform better, on average, but some countries do better or worse than their income would predict” (p. 92). In my opinion, Bonnet was right in concluding that the bias in the report in favor of assessing education from the point of view of economic efficiency is overshadowing another conception of education, prevailing in Europe, whereby education is just as much about social cohesion, personal fulfillment, and cultural development. It is very frustrating for those who care about children if a minister of education proudly states that the Netherlands has a very efficient educational system, because we outperform most other countries, while spending well below the OECD norm. In this way international comparative studies can be very destructive to the improvement of the teaching and learning process.

**Impact of Studies**

Consider TIMSS. It is very tempting to use as a source the publication by Robitaille, Beaton, and Plomp for IEA/TIMSS with the title: The Impact of TIMSS on the Teaching and Learning of Mathematics and Science (2000). The title of the book is a bit misleading as it deals only in a minor way with the “impact”; the book reports also on national education systems, and the results and findings per country. Although the title of the book stresses—as it should do—the “learning and teaching,” this aspect is difficult to find in the book. This is also to some extent the conclusion of the authors as they state: the results seem, at least until now, to have made a greater impact on the general public and in political circles through the mass media than they have on educators. On the other hand, it is also true that many of the reports refer to the impact of the study on the process of curriculum reform or on national evaluation programs. Those kind of processes tend to take a long time to bring to fruition, and they are then followed by an equally lengthy and perilous process of implementation. So it is probably unrealistic to expect that the results would have made more of an impact at the classroom level since only three years have elapsed since the publication of the first set of international results. (p. 168)

Without trying to be exhaustive a rather typical example of a specific country’s reaction under the chapter “Impact” will resemble something like the following (based on Switzerland):

TIMSS showed primarily where Switzerland stands in comparison to other countries with respect to achievement in mathematics . . . [and] gave a detailed picture of current levels and patterns of achievement. This is a valuable basis for discussion of needed improvements. TIMSS also gives some indications of possible reasons for such differences in achievement. . . . However, given the way TIMSS was designed, such statements of causality are only valid to a limited extent. It is not easy to deduce directly from TIMSS how improvements might be made. While TIMSS had a direct effect on debate and reflections concerned with educational policy, educational planning, pedagogy, educational science, and teacher preparation, it had only an indirect effect on corresponding practice. (p. 158)
One is tempted to comment that these observations cannot come as a big surprise. Having some experience and involvement with both TIMSS and PISA (see note at end), I find it striking to note the relatively little interest that IEA and OECD have in both content and teachers and students, the key factors in math education. As the studies are usually announced and described as an educational system analysis, and as relations are made between the performance of educational systems and the economy, it is hard to expect much interest from the “real” stakeholders in mathematics education. The introductory remarks for the “impact study” from the executive director of IEA, Hans Wagemaker, make clear where the emphasis lies:

International comparative studies of educational performance . . . can therefore play a central role in assisting policy-makers, curriculum experts, and researchers in judging the extent to which an education system meets national expectations and the extent to which it is likely to contribute to the nation’s overall economic and social well-being. (p. 7)

I now turn my attention to examples of the possible impact of the PISA studies. They are both not about system changes, but about content-oriented matters, taking teachers, children, and content as key variables. In the examples I will discuss two completely different ways the outcomes of PISA 2000 have been used—one of them being by Germany. This cannot come as a surprise after the introduction of this article focusing on the PISA shock; but Germany is also a good starting point for quite another point that I will only mention here, but that is worth further research and analysis. The Germans produced a national PISA report of 550 pages, the international OECD report was 330 pages, and the Dutch report a mere 65 pages. Most countries had something around 150 pages. It is not the statistics that are interesting here, but the message from the report and what has been selected to be included. Even a superficial analysis, which was carried out for this article with the reports mentioned and the one from the United States, makes significant differences visible. As Keitel and Kilpatrick (1999) pointed out very correctly, there is a common myth that “numbers do not lie.” It is now widely accepted that data can be gathered, processed, mathematized, and interpreted in a variety of ways. So a key issue is the question of who influences this process, for what reasons, and through what means. The four studies just mentioned underscore this concern apart from the fact that even numbers can lie.

Back to the very gründliches German report. Not only did the German PISA Konsortium do an excellent and thoughtful job, it also made recommendations for immediate improvement, including ones that directly affect the content. According to Blum (2000), the changes should include

- more integration of inner- and outer-mathematical “networks”;
- fewer calculations;
- more thinking activities and student mental “constructions”;
- more reflection;
- more flexible use of schoolbooks.

These goals can be reached when the recommendations that were formulated after TIMSS are implemented:

- development of a different math-problems culture: more open-ended, more “real-world”;
- a new teaching-and-learning culture, with a more exiting cognitive school environment (Klieme, Schümmer, & Knoll, 2001);
- more and different professionalization of teachers, emphasizing teamwork.

PISA adds to these recommendations a “very different conceptualization” of mathematical concepts and emphasis of modeling and mathematization, situated in contexts. And, argued the report, the Germans have definitely not reached the optimum in using different representations as a tool to build better conceptual understanding.

Mathematics education is in a state of transition, in part because of the fact that both TIMSS and PISA were taken seriously. Surprisingly the shock and catastrophe that struck Germany as some kind of natural disaster, if one had only the popular media as a resource, has resulted in a government-supported nationwide action-plan with a very strong content part that will result in a different mathematics education culture at schools. Of course, the success of these changes will be measured by PISA 2003, 2006, 2009, and so on. At least in part.

A very different way of using the “very different conceptualization” of PISA was reported by Dekker and Feijs (2005). Classroom teachers (in Philadelphia and Milwaukee, WI) were invited to reflect on their own classroom assessment, initially in their own way using their principles. Teachers were questioned why they asked the questions they asked, why they ordered the questions as they did, what were the presumed learning goals that the items tried to operationalize, and so on.

After getting familiar with the different competency clusters (computation, connections, reflections) and discussing PISA items and other “good” items that
were more of an open kind and needed more student construction, the teachers were again confronted with their own classroom assessments. This Classroom Assessment as a Tool for Teacher Change (CATCH) project was considered powerful because it not only changed the teacher’s attitudes towards the teaching-learning-assessing process, but also changed their actual classroom practices—not just their classroom assessments. It seems that using an international comparative study by focusing on the concept of the study from a mathematical point of view, in combination with the available instruments, and by inviting teachers to make this instrument better and more fit for the classroom practice should be strongly encouraged as it can be a powerful tool from the bottom-up in combination with top-down approaches.

In the CATCH project a competency model was used that was quite similar to that of the PISA framework (2003). The three different competency clusters were visualized in the Assessment Pyramid (De Lange, 1994, 1999) as competency levels. This pyramid turned out to be very effective in changing teachers’ attitudes and practices in the classroom. Some of this effect can be attributed to the fact that PISA had adapted a similar model, which is seen as adding validity to the Pyramid.

As indicated before, I would like to close this “international comparative studies” part with a short discussion about other large-scale assessments in mathematics education other than the TIMSS and PISA studies that I have discussed in some detail.

Numeracy, Quantitative, and Mathematical Literacy Assessments

Studies like NAEP and TIMSS are connected to some kind of “idealized” curriculum. PISA is not, as it claims to test mathematical literacy, a term described as an individual’s capacity to identify and understand the role that mathematics plays in the world, to make well-founded mathematical judgments and to engage in mathematics in ways that meet the needs of that individual’s current and future life as a constructive, concerned and reflective citizen (OECD, 2001, p. 22).

If the TIMSS publication on the impact of the study expressed some concern about the relative lack of interest in the results, this lack of interest is even more evident in studies about numeracy, quantitative literacy, and mathematical literacy, especially if these are about adults and not, like PISA, for 15-year-olds.

An important event in this respect was the Young Adults Literacy Skills (YALS) project in the United States, in which young adults (12–15) were assessed on tasks related to three literacy scales, among them quantitative literacy. The description of quantitative literacy was somewhat limited compared to the more recent PISA definition: “The knowledge and skills required to apply arithmetic operations, either alone or sequentially, using numbers embedded in printed material (e.g., balancing a checkbook, completing an order form)” (Dossey, 1997, p. 46; Gal et al., 1999, p.12). The results of this assessment led to conclusions like “performances on the scale were less than what the nation required in human talent for long-term international competitiveness” (Dossey, 1997, p. 48).

The follow-up of YALS was the National Adult Literacy Survey (NALS), carried out in 1992. Gal (1993) concluded that roughly 50% of American adults would either have major difficulty with or be fully unable to handle real world tasks such as:

- using a bus schedule to determine departure time;
- identifying a trend on a simple graph about sales figures;
- using a calculator to find the difference between the regular price and the sale price;
- estimating the unit price of a grocery item;
- understanding a table summarizing a school survey;
- calculating interest charges related to a home loan.

In the technical report Issues and Challenges in Adult Numeracy (1993) Gal observed that there are two “extreme” views on numeracy. One extreme is to see numeracy as the lower end of mathematics, or whatever math educators attempt to achieve in the early grades (Van Groenestijn, 2002) whereas the other end is viewed as “encompassing a broad set of skills, knowledge, strategies, beliefs and dispositions that people need to autonomously engage in and effectively manage situations involving numbers, quantitative on quantifiable data, or information based on quantitative data” (Gal, 1993, p. 2).

More recently the definition has broadened even more, as in PISA, to reflect that literacy in the area of mathematics goes well beyond the strictly numerical (numercy) or quantitative aspects (quantitative literacy). Confronted with this problem, I proposed in Madison and Steen’s publication on quantitative literacy (2003) to see numeracy as part of quantitative literacy. This includes understanding of and mathematical abilities concerned with certainties (quantities), uncertainties (quantity as well as uncertainty), and relations (types of, recognition of, changes in,
and reasons for those changes) and defines spatial literacy, most likely the most basic and neglected aspect of mathematical literacy, as understanding of the (three dimensional) world in which we live and move (De Lange, 2003). Shaeffer (2003) supported this view because of the indisputable fact that much more in mathematics is useful besides numbers. Indeed, many aspects of statistical thinking (uncertainty) are not about numbers as much as about concepts and habits of mind, he argued.

This discussion is relevant to the problems of large-scale assessments as it is not clear at all what role school curricula play in setting up large-scale literacy assessments, although clearly some are very much tuned to traditional curricula and others are more society oriented. An example is the International Numeracy Survey (INS) in which only a limited number of countries took place: U.K., France, the Netherlands, Sweden, Denmark, Japan, and Australia, organized by The Basic Skills Agency in London in 1997. Two typical examples (out of only 12 problems) are:

The total of 4.25, 6 and 7.74 = 18.99
Work out 15% of 700.

Van Groenestijn (2002, p. 34) observed that “though in this study ‘numeracy’ was seen as doing mathematical operations on school-based tasks, it does indicate that the performance on such school-related tasks is disappointing and presumably undermines numerate behavior.” This observation seems somewhat questionable as it directly relates literacy to school mathematics in a restricted basic skills sense and fails to clarify how such skills “undermines” numerate behavior. I cannot but agree on the observation that the performance on the task is disappointing. What can be learned from that is quite another matter.

The NALS was followed by an international study, set up by the OECD: the International Adults Literacy Skills Survey. This survey was carried out in 1997 in OECD countries. Apart from the United States and Canada 21 countries including Germany, Sweden, Switzerland, and the Netherlands took part. This study defined quantitative literacy quite narrowly, trying to connect traditional arithmetic with the problems of daily life: “To apply arithmetic operations to numbers embedded in printed materials, such as balancing a check-book, figuring out a tip, completing an order form, or determining the amount of interest on a loan from an advertisement.”(NRC, 1989, p. 7–8)

Finally I mention the follow-up survey of IALS, the Adult Literacy and Life Skills Survey (ALL) organized by the National Center for Education Statistics (NCES) and Statistics Canada, which uses a more elaborated numeracy domain. An interesting “Policy Information Report” about this study was published in 2002 (Sum, Kirsch, & Taggart, 2002) for the United States. The report provided a good view of the role of international large-scale assessments as the goal of the report is to call attention to important findings from the IALS and their implications in terms of educational and workforce development policies. The report also stated that it is interesting that the results of IALS are becoming available at a time when the role of human capital in influencing the fate of individuals and nations is receiving increased attention. According to a recent report released by the OECD (2001), research shows that the development of human capital is correlated with better health, lower crime, political and community participation, and social cohesion. Some studies, they reported, even suggest that the social impacts of acquiring these knowledge and skills could be as large as their impacts on economic productivity. Elsewhere in this article some doubt is cast on this statement, or more precisely on the relation between achievement in education and economic performance in a causal way.

One of the more interesting results is the fact that 46% of the respondents in the United States belonged to Level 1 and 2, “possessing a very-limited to limited mathematical literacy proficiency.” Also noteworthy is the fact that there is no “horse race” ranking in the report: The United States is compared only with all high-income countries, but only as a group.

From the summary I quote the following remarks, clearly identifying the purpose of the report as the title is Summary of Findings and Policy Recommendations.

- The U.S. educational system is clearly less productive in raising literacy skills of students per dollar spent.
- The U.S. appears to be living off its past as higher educational investments and will inevitably lose ground in the coming decade.
- The nation’s changing demographics will likely exacerbate the literacy skill deficit in coming years.
- Although skills influence labor market success everywhere in the world, this is especially true in the U.S. and other English-speaking countries.
- Mediocre skills and inequality in the distribution of skills may have worked reasonably well in recent years but are not likely to do so in the future.

Given these results it may hardly come as a surprise that mathematics educators pay relatively little attention to these studies, as the report gave no indication...
whatever about the instruments used and the reasons behind them. It is my opinion that more balanced reports, not only this one, showing the links between policy implications and the assessment instruments are in the interest of everyone.

In 2000 the OECD carried out the earlier discussed PISA study, also with a much wider ranging definition of mathematical literacy. It seems not very reasonable that the difference in definition of literacy can be attributed to the difference in target population: adults or 15-year-olds. Hopefully this change—by OECD—can contribute to newer insights into the importance of mathematics in society and the relevance of seeing mathematics function in the real world.

---

**NATIONAL LARGE-SCALE ASSESSMENTS**

**Introduction**

Two major clusters comprise large-scale assessments on a national level: a discussion of NAEP and the TIMSS Benchmarking Study for the United States at one hand and at the other side the national exit examinations as they exist in many countries around the globe. It goes without saying that, like this whole chapter, I can touch on the subjects discussed and then only in a quite selective way.

**NAEP**

In 1969 the Carnegie Corporation of New York appointed the Exploratory Committee on Assessing the Progress of Education (ECAPE) with Ralph W. Tyler as chair. According to Tyler (1969) a major purpose of the national assessment program was to provide the lay public with census-like data on the educational achievements of children, youth, and adults—data that furnish dependable background information about educational attainments, the progress being made, and the problems still faced in achieving educational aspirations.

This was the first description for the program that is presently known as the National Assessment of Educational Progress. NAEP is a national survey intended to provide policymakers and the public (definitely lay people) with information about the academic achievement of students across the nation. It serves as a source of information for policymakers, school administrators, and the public for evaluating the quality of their educational programs. Just like TIMSS, NAEP is not tied to a specific curriculum. Whereas TIMSS tried to find a set of instruments that best fit the existing curricula of all participating countries, NAEP is based on a set of frameworks that describe the knowledge and skills to be assessed in each subject area. The performances assessed are intended to represent the leading edge of what all students are learning.

For over 30 years, the NAEP has served as the United States’ only ongoing gauge of student achievement across time. One has to understand that just like with the other large-scale assessments discussed before, the studies do not report or even intend to report on student’s individual achievements. This is caused by the multiple-matrix sampling method, the design selected by the founders of NAEP. Not all students will be sampled. On the contrary, a strategically selected sample can support the targeted inferences about groups of students with virtually the same precision as the familiar approach of testing every student. Moreover, like in PISA and TIMSS, not all students are administered all items. It can hardy come as a surprise that one of the current debates is about this aspect of large-scale assessments of this kind: the impossibility to provide student-level reports, something that exit examination do provide, sometimes in great detail. In NAEP, measurement at the level of individual students is poor, and individuals cannot be ranked, compared, or diagnosed (NRC, 2001, p. 224). Worse even, students who have similar performances on one particular item can have different competencies—something known not only at the student level, but also at country level in TIMSS or PISA.

Silver, Alacaci, and Stylianou (2000) have demonstrated some limitations of scoring methods used by NAEP for capturing the complexities of learning. For a particular item NAEP concluded that 11% of the students gave satisfactory or better responses but can give no further detail about the important differences in the quality of the reasoning demonstrated. In this particular case the reasoning varied from surface-level reasoning to analytic reasoning and to even more sophisticated reasoning.

Dossey, Jones, and Martin (2002) confirmed that indeed a “correct” answer is only part of the story. They did so in the context of the TIMSS study, which used the double-digit coding to find out more information about the strategies used by students in an effective way. In a case that involves three distinct different ways of reasoning one can code a correct answer with 1.1, 1.2, or 1.3, the 1 standing for “correct,” the 1, 2, and 3 for the strategy used. In this way important information about the process or strategy of the student can be made visible to a certain extent. But the data offer more for those interested in the actual solution process, in addition to double-digit coding. This can be interesting at the national level, but also at the international level. In
one country most students could be using one—often computational—technique, whereas in another country students use a variety of strategies.

A national analysis of the actual process the students have gone through is done by studying and investigating actual students’ work, in combination with double-digit coding and PISA scores—for instance in the Netherlands. Indeed, it seems that stakeholders are increasingly interested in better—which often means more complex—assessments, even on a large scale and even if this means higher costs and less reliability because of intersubjective scoring. I will come back to this issue later, when discussing future developments.

Over the past thirty years (since 1973) three distinct NAEP projects have evolved: the Main NAEP, the Long-Term Trend NAEP, and the State NAEP. The Main NAEP periodically assesses students’ achievement in reading, mathematics, science, writing, U.S. history, civics, geography, the arts, and other subjects at Grades 4, 8, and 12.

Trial assessments with voluntary state-by-state assessments were conducted in 1990 and 1992 and were considered so successful that the U.S. Congress authorized regular state assessments starting in 1996 (Allen, Jenkins, Kulick, & Zelenak, 1997).

The content of both the Main and State NAEP programs follows curriculum frameworks developed by the National Assessment Governing Board (College Board, 1996). Kane and Webb (2004) argued that test-item types for the Main and State NAEP assessments are consistent with the current state-of-the-art in achievement testing and have evolved over time with curriculum changes. In contrast, both the students sampling frame and the content of the Long-Term NAEP have remained essentially unchanged.

This Long-Term NAEP, or the Long-Term Trend Mathematics Assessments (LTT Mathematics), has provided a measure of students’ achievement over time by administering similar assessments periodically since 1973 (Galindo, Caulfield, Mohr, & McCormick, 2004). It should be noted that the LTT has a computational focus, measuring students’ knowledge of basic facts, their ability to carry out paper-and-pencil calculations, their knowledge of basic measurements formulas, and their ability to apply mathematics to daily-life situations such as those dealing with time and money. This focus reflects the school mathematics in the United States in the 1970s. The content areas of the LTT are represented in the Main NAEP. The three types of items used on Main NAEP—multiple-choice, short constructed response, and extended response—are reduced to only multiple-choice and short constructed response in the LTT.

Not only in content and item-type format is the LTT different from the Main NAEP, but also the reporting scales. The Main NAEP is more comparable in many respects with the international comparative studies like TIMSS and PISA. The Main NAEP has three achievement levels: Basic, Proficient, and Advanced. The levels are based on collective judgments of experts about what students are expected to know and do at each grade level (NCES, 2001). The performance is based on a 0 to 500 scale, just like with LTT (and PISA and TIMSS), but in the Trend Study the students’ performance is reported in terms of percentages of students, which corresponds to five points on the scale: 150, 200, 250, 300, and 350, with level 150 described as “Simple Arithmetic Facts” to Level 350 as “Multi-step Problem Solving and Algebra.” The five LTT performance levels were arbitrarily set at 50-point increments.

Although the results of the LTT study indicate an almost continuous upward—though very slow—trend, the gains document the fact that students’ ability to complete traditional mathematics tested by means of a multiple-choice or short-answer format questions has improved across the 30 years, and that the gains were much smaller than the gains on Main NAEP during the same period (Galindo et al., 2004). The “deeper” meaning of this phenomenon is somewhat unclear, as it is unknown how much influence familiarity with the test instrument has on the outcomes. It is a fact that the longer a test exists the more familiar students and other stakeholders become with the nature of the test items. Experts can already recognize a NAEP, TIMSS, or PISA item from a distance, although the items seem to converge, which may be an undesirable development. This familiarity might itself be responsible for the observed almost continuous, but slow, upward trend. Policymakers even go so far as to speak of a TIMSS curriculum, indicating that they think one can identify a framework behind these studies that is reflected in the items and item formats. A familiar request that reaches persons or organizations responsible for these studies is “Please help me to prepare my students for the next study.”

Linking the LTT NAEP with the Main NAEP is no easy task because of different content, different scaling, and different item formats. Also linking State NAEP results to those administered by states or local schools is difficult, if not impossible, to achieve (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999); Linn (2000) observed that comparability of state assessments with NAEP scores is sometimes compromised when purposes differ. Add to this the evidence suggested by Kenney and Silver (1997) and Linn, Koretz, Baker, and Burstein (1991) that variation in item for-
mat may result in different estimates of student knowledge and skills across content topics and content dimensions. This point was also made in an explorative study in assessing mathematics problem solving for 16–17-year-olds (De Lange, 1987).

Linking NAEP with TIMSS and PISA is also difficult: The three assessments are targeted towards slightly different student populations, place different emphases on content areas within mathematics, include questions requiring different types of responses and thinking skills, and report results in different ways.

From the content point of view the most commonly addressed NAEP mathematics Content Strand on both NAEP and TIMSS was number sense, properties, and operations, addressed by 32% of NAEP items and 46% of TIMSS-R items, compared to only 9% of PISA items (Nohara, 2001). PISA did pay a lot of attention to Data Analysis, Statistics, and Probability: 31% of the items belonged to this cluster, with 14% in NAEP and 11% in TIMSS. (Note that these data are for PISA 2000; PISA 2003 has a more balanced distribution matrix.) Probably more important than attribution of items to content strands is the fact that one fourth of the items of Data Analysis, Statistics, and Probability were related to the subcategory “read, interpret, and make decisions using tables and graphs,” which may come as no surprise given the fact that PISA aims at measuring mathematical literacy. The more curriculum oriented studies, NAEP and TIMSS, have far fewer items in this subcategory: 4% and 7%, respectively.

In the figure above, what fraction of rectangle ABCD is shaded?

- $\frac{1}{6}$
- $\frac{1}{5}$
- $\frac{1}{4}$
- $\frac{1}{3}$
- $\frac{1}{2}$

Figure 25.6 Example item from NAEP (NAEP, 2004, p. 29).

Given these observations it may come as no surprise that NAEP and TIMSS address similar sets of subcategories within each of the content strands. This becomes also clear in the percentage of items that present students with real-life situations or scenarios as settings for the problem: NAEP scores 48%, TIMSS 44%, and PISA again is very different: 97%. One should keep in mind that these numbers give some indication of the emphasis placed on real-world situations but do not tell anything about the artificiality, relevance, role, mode of representation, familiarity, and so on, of the context.

A weak spot in any existing large-scale assessment is addressed by many authors and summarized very well in Knowing What Students Know. The authors commented on NAEP by noting that NAEP is beginning to be influenced by the call for more cognitively informed assessments of educational programs:

Recent evaluations of NAEP (National Academy of Education, 1997; NRC, 1999) emphasize that the current survey does not adequately capitalize on advances in our understanding of how people learn particular subject matter. These study committees have strongly recommended that NAEP incorporate a broader conceptualization of school achievement to include aspects of learning that are not well specified in the existing NAEP frameworks or well measured by current survey methods. The National Academy of Education panel recommended that particular attention be given to such aspects of student cognition as problem representation, the use of strategies and self-regulatory skills, and the formulation of explanations and interpretations, contending that consideration of these aspects of students achievement is necessary for the NAEP to provide a complete and accurate assessment of achievement in a subject area. (pp. 244–245)

The NRC agreed with this position and added that NAEP should include carefully designed, targeted assessments administered to smaller samples of students that could provide in-depth descriptive information about more complex problem solving over longer periods of time. For instance, smaller data collections could involve observations of solving mathematical problems in groups.

It should be noted that the recommendations suggested in the above publications are not new: Both TIMSS and PISA designers realized this problem and have expressed intentions to address the issues of more complex processes. But the realization of these intentions is no simple task: We have already noted the intentions for more complex assessments that were mentioned in the PISA 2000 framework but have not been realized so far. One can easily guess the reasons: The required “careful design” is extremely difficult, expensive, and not easy to score even if just settling for qualitative descriptions. Time is another big issue.
As I will discuss later there are already rather large-scale assessments in existence, even on problem solving in group work, but they take at least one full day to complete. On the other hand, there are also examples of national large-scale (exit) examinations that consist of more than one paper-and-pencil session of 50 minutes or so and do address more complex processes. To discuss these national tests will be the next part of this chapter. But before going to this interesting area I would like to mention an interesting study that tries to connect an international comparative study (TIMSS) with state or local assessments: The Mathematics Benchmarking Report TIMSS 1999 (Mullis et al., 2001a).

TIMSS 1999, also known as TIMSS-R, focused on mathematics achievement of eighth-grade students. The United States and 37 other countries participated, and important for the United States, 13 states and 14 districts or consortia participated in a voluntary Benchmarking Study. The study provided those states and consortia an opportunity to assess the comparative international standing of their students’ achievement and to evaluate their mathematics programs in an international context. And, of course, compare themselves with the score of the United States as a whole.

The Benchmarking Study confirms the extreme importance of looking beyond the averages (horse race) to the range of performance found across the nation. It may come as no surprise that performance across the participating school districts and consortia reflected nearly the full range of achievement internationally: Although achievement was not as high as Singapore, Korea, and Chinese Taipei, the top-performing Benchmarking jurisdictions of Naperville School District and the First in the World Consortium (both Illinois) performed similarly to Hong Kong, Japan, Belgium (Flemish), and the Netherlands. The problem is, as mentioned before, that the international context consists of four very different socioeconomic cultural entities. And one wonders what to do with the finding that the Chicago public schools, the Rochester school district, and the Miami-Dade county public schools are at par with Thailand, Macedonia, and Iran.

This item from the TIMSS Benchmark study (Figure 25.7) illustrates an item from the content area Fractions and Number sense, 8th grade mathematics. Singapore scored 89% correct, the American First in the World Consortium scored 71%, England and the Russian Federation 52%, Maryland 42%, and the Dade County Public Schools scored 20% correct.

Another positive development regarding TIMSS is the inclusion of videotaping in classrooms and Case Studies. In doing this it follows the recommendations of the authors of Knowing What Students Know in an unprecedented way. In his overview of the Case

Figure 25.7 TIMSS Benchmark Study-item (Mullis et al., 2001a, p. 72).

Studies, Stevenson (1998) argued that the questionnaires, as used in the main studies of TIMSS and PISA, have both strengths and weaknesses. Their major strength—as with multiple choice—is to cheaply gather large amounts of data on a wide variety of topics from many individuals. On the other hand, and this causes serious problems, respondents may provide answers that are difficult to interpret. Deprived of the opportunity to question the respondents about the meaning of certain replies, to probe for more complete answers, or to ascertain the respondents’ understanding of the questions, the investigator can only conjecture about what an answer might have meant. And the opposite is a problem too: What is the real meaning of the question for the respondent?

Case studies, which allow for one-to-one interactions between investigator and respondent, can help to understand the contexts and relationships that lie behind more traditional quantitative surveys of beliefs, attitudes, and practices, especially when seen in combination with the more quantitative data.

The five-volume report from TIMSS on Japan, Germany, and the United States offers examples from standards to homework, from teacher training to response to individual differences. From these examples Stevenson (1998) concluded that the education systems of the different countries are embedded within the culture of each country. As a result of this close relationship, one cannot hope to learn how education systems can be improved or academic achievement increased without understanding the actions, beliefs, and attitudes related to education in that culture. However it is impossible to make causal conclusions about the relation between scores on a test and the culture.
National Exit Examinations, a Personal Experience

Reaching this part of this chapter I am tempted to reflect on this subject in a very personal way. As student who underwent a national exit examination, as a teacher who was responsible to prepare students for this event, and as a researcher interested in finding out if different tests measure different competencies. As a student I was perfectly aware of the goal of our 11-year-long educational process: to pass the national examination, a national event for all students at exactly the same time, somewhere in May. The final year was merely used to prepare us for the main event: to do three 3-hour long written tests—by no means with multiple-choice but by solving large and complex problems, although not in context. To be prepared was quite easy, to pass difficult. Easy to prepare because there were numerous examples of examination problems from previous years, as all examination test problems are in the public domain after the examination. Passing the exam was difficult because students needed high technical skills and preciseness. Apart from the written tests (9 hours total) most students had to do an oral session as well, under the principle that different cognitive demands could be operationalized under different formats. Some deeper conceptual understanding should be measured in an oral session carried out by the teacher and an outside “expert,” quite often a university mathematician. But if a student had very good grades for the written exams, one did not have to do the orals. And after passing the exam, one had a feeling of absolute freedom: One could enter any university to study an almost unlimited number of disciplines.

As a teacher I felt that the final national examination was a rather threatening event, the moment of truth whether I was a good teacher or not. The only good thing about the last year of schooling was that I had the undivided attention of most students, as they were all very interested in passing the exam. But nightmares for beginning teachers were common: All students failing and the teacher being fired was a dreaded scenario. Of course the teacher was also responsible for scoring the test, but regrettably a randomly appointed colleague had to check the work, and the scores were expected to be within a very narrow band. Reflecting on this practice it is quite surprising that parents completely trusted this process, even today in the twenty-first century. Nowadays the oral has been discontinued because of economical reasons—a problem that again and again overshadows good assessment practices. It was replaced by a school exam, to be designed and scored by the teacher. It accounted for 50% of the total score of the final examination. There was a large degree of freedom for the teacher resulting in many teachers’ trying to copy the “real thing” at the end of the same school year. But more adventurous teachers tried something very different like essays and open-ended problem-solving items.

Although the examinations included quite extensive open-ended problems, the researcher in me was interested in trying to find other ways to operationalize the so-called higher order thinking skills, with expected but still surprising results. I used the liberty of the 50% school examination to find out differences to test out the restricted-time written test, a take-home test, and an oral test. The most striking observation I was able to measure was that the correlation between the different formats was low, indicating that indeed one test cannot measure all cognitive demands or, in this case, that different formats measure different competencies (De Lange, 1987).

Exit Examinations

School exit examinations are not a very popular object of study. One of the few studies, Secondary School Examinations: International Perspectives on Policies and Practices by Eckstein and Noah (1993), was announced as filling an informational and conceptual vacuum. The study observed that most (industrialized) countries require students to take extensive examinations at the end of secondary education to determine whether they have satisfactorily met national criteria and are qualified for further education. These examinations are also used to assess individual schools (and teachers, as indicated in my personal note) and a nation’s educational program as a whole. It is remarkable that the United States does not have high school tests at the national level: “The United States is unique among the countries we have studied in having no coordinated, public, national system for assessing student achievement at the end of secondary school” (Eckstein & Noah, 1993, p. 238). In Barton’s paper (1999) the author commented that the examinations in the countries studied by Eckstein and Noah (U.S.A., China, Japan, Germany, England and Wales, France, Sweden, and the former Soviet Union) are closely related to the curriculum. In the United States, he continued, it is hard to conceive of any national exam being closely related to the actual curriculum for such a high-stakes examination because of the decentralized control over the curriculum. According to Eckstein and Noah the governmental control of the school curriculum in the United States has been extraordinarily weak.
But this argument of decentralization cannot be used to not introduce decentralized exit examinations. In several countries there are no national exit examinations but state examinations (like in Germany and Australia).

Japan, the United Kingdom, France, and Germany were the subjects of a more recent study: Comparisons of Entrance and Exit Examinations (Stevenson & Lee, 1997). The authors observed also the relation between curriculum and examination, noting that:

rather than imposing some arbitrarily defined standard of achievement, the examinations are closely tied to what the students have studied in high school. Because teachers are aware of what students are expected to know in examinations, it becomes their responsibility to equip students with the information and skills needed to pass the examination. (p. 47)

It is amazing to see how this fits with the personal experiences described before.

One of the observations in the introduction of the Eckstein and Noah study is the relative lack of systemic, cross-national comparative work, and the fact that, with only a few notable exceptions, the comparative study of examinations has been a largely neglected aspect of comparative education. Although the study is somewhat dated (1993), many observations and conclusions are still very informative if only as sketching possibilities for future development. Quite often things of the past are neglected just for that reason, without looking at the benefits from a strictly educational point of view instead of a practical or economical perspective. A nice example is the disappearance of the oral part of the mathematics examination in the Netherlands: Officially one argument was the lack of validity and reliability with two persons examining a student. In reality is was an economical and logistical argument that did away a format that quite often was highly regarded by students. It is hopeful to observe that the new rules allow for a variety of formats in the school examination part of the exam, bringing back a more balanced assessment system.

The Eckstein-Noah study observed that mathematics is a subject that all must take in some countries, with accommodation for more and less advanced studies. In others, it is differentiated according to specializations in the subject, all at an advanced level, whereas in yet other countries it may not be required at all. Substantial differences among countries in the knowledge required of candidates in mathematics exist to a surprising extent. But what is even more remarkable, especially in light of the (inter)national large-scale assessments, is the observation that format and content are inseparable, an ever-returning point in any discussion about valid and fair assessment when the designers of large-scale assessments point out that multiple-choice and short-answer questions quite often offer enough possibilities to test the students. It is commonplace that the format of examinations strongly influences both the content and style of teaching. As a consequence, the format of examinations, as much as their content, is a target of the debate over the proper objectives of instruction and over what teachers and schools are supposed to do. The United States and Japan continue to rely heavily on multiple-choice formats. In China as well, in response to the large number of candidates, the high cost of university entrance examinations, and concern about their reliability and objectivity, the examination authorities are increasingly substituting short-answer and multiple-choice questions for the traditional extended-answer format. Concern for reliability and validity thus influences the choice of formats, which in turn influences the content selected for examination, which in turn influences the content and style of classroom instruction.

These remarks are even more remarkable as they were made more than a decade ago, and one assumes that some progress has been made in this area. But this progress fails to show up in the international large-scale assessments to a convincing extent. TIMSS still has a large amount of multiple-choice and short-answer questions, and although PISA may have less than 50% multiple-choice, it still limits itself to very short restricted time-open questions. I noted before that Knowing What Students Know recommended strongly that NAEP should include carefully designed, targeted assessments that could provide in-depth descriptive information about more complex activities that occur over longer periods of time (longer than 50 minutes). Having participated in several discussions about the trade-off between “fair-and-valid” assessments and “practical-and-cheap” assessments I have noticed that the decision makers often tend to favor the “practical-and-cheap,” which fits in nicely with the earlier observation that these stakeholders show little interest in the content domain of large-scale assessments.

“Broader” Assessment Systems

That many countries and states have shown that a broader “assessment system” can work in reality has been indicated by the study about the examinations. The variety of formats used for this form of “high-stakes testing” is almost unlimited. An example from the United States that gives indications of future types of large-
scale assessments is the New Standards Project. The choice for this example is not completely at random.

**VCE**

One of the quite interesting new approaches to something like the New Standards Project was undertaken in the state of Victoria, Australia. One of the principles behind this examination of the Victoria Certificate Assessment Board was that teacher would like to teach to the test. So the test had four different formats, testing very different cognitive demands. I have described this daring approach earlier (De Lange, 1996), and many articles are around by people directly involved. In one of their earlier publications, Money and Stephens stated that applications and modeling are central features of the Mathematics Study Design in the Victorian Certificate of Education (VCE) in the state of Victoria, Australia (Money & Stephens, 1993). All students undertaking mathematics in the final 2 years, that is Years 11 and 12, engage in substantial applications and modeling, with content options up to 50% to provide for the widely varying abilities and interest of students. In my opinion it is fair to say that the new test requirements lead to different practices in teaching (test driven innovation) or, as Money and Stephens saw it: “The new arrangements have had a significant impact on the teaching and learning of mathematics.” (p. 336). Much responsibility is given to the teachers because the Mathematics Study Design for the VCE provides a framework within which teachers develop courses, each normally 1 year in length.

Three work requirements, detailed by the teacher but each taking up between 20% and 60% of the course time, provide the basis for determining satisfactory completion and certificate credit for all semester units within VCE Mathematics:

- **Skills Practice and Standard Applications**: the study of aspects of the existing body of mathematical knowledge through learning and practicing mathematical algorithms, routines, and techniques, and using them to find solutions to standard problems.
- **Problem Solving and Modeling**: the creative application of mathematical knowledge and skills to solve problems in unfamiliar situations, including real-life situations.
- **Projects**: extended independent investigations involving the use of mathematics.

For teachers, these work requirements provide the framework for a shift in the balance of instructional patterns and a shift in role from dispenser of information towards learning facilitator, catalyst, and coach. For students, the work requirements provide an accessible pathway towards formal recognition of the learning processes they have undertaken.

It seems rather trivial to observe that one can only introduce a modeling and applications curriculum when the assessments do really operationalize the goals of such a curriculum (De Lange, 1987), a point repeatedly underlined by many authors (e.g., Blum & Niss, 1991; Money & Stephens, 1993). To achieve this, the Common Assessment Task (CAT) initially consisted of four parts:

- **CAT 1**, an Investigative Project, is based on a centrally set theme and is intended to represent 15 to 20 hours of students’ work;
- **CAT 2**, a Challenging Problem, is chosen from four centrally set problems for each course, representing about 6 to 8 hours of students’ work;
- **CAT 3**, a Facts and Skills Task, is a multiple-choice test of 90 minutes duration, covering the full range of core and optional content;
- **CAT 4**, an Analysis Task, is a test, also of 90 minutes, consisting of about four structured questions that lead from routine to nonroutine aspects of a problem.

As the reader can judge, my earlier qualification of a “daring” approach seems quite appropriate, especially for real large-scale assessments on a state level.

It cannot come as a surprise that not all CATs survived, but this does not affect the message: To design appropriate large-scale assessments, incorporating the more complex competencies, one is confronted with a daunting and challenging task.

I conclude with some examples from the Mathematical Methods Written Examinations 1 and 2 from November 2003. Written Examination 1 (Facts, Skills, and Applications) consists of two parts: Part I and Part II. Part I has 27 multiple-choice items, Part II has 6 short-answer questions. The 27 items of Part I are good for a score of 27 points, the 6 questions for Part II are worth 23 marks. Figure 25.8 shows an example of both parts:

Another 1 ½ hour is given to the students to complete four questions in Written Examination 2: Analysis Task. The four questions have several subquestions (see Figure 25.9), and the total is worth 55 marks.

**New Standards**

As the United States is often seen as “the mother of multiple-choice and other bad habits in assessment,” in part because of the existence of a commercial testing
industry and its collaborators (Kaplan Test Preparation Courses), it is remarkable that a project like the New Standards Project actually made an impact in the United States. The project illustrates a way to approach a number of issues related to large-scale assessments that differ from the earlier described international and national large-scale assessments.

The program was designed to provide clear goals for learning and assessments that are closely tied to those goals. A combination of on-demand and embedded assessments was to be used to tap a broad range of learning outcomes. Development of the program was through collaboration between the Learning Research and Development Center of the University of Pittsburgh and the National Center on Education and the Economy, in partnership with states and urban school districts. Challenging standards were established for student performances at Grades 4, 8, and 10, along with large-scale assessments designed to measure attainment of those standards.

The New Standards Project consists of three interrelated components: performance standards, a portfolio assessment system, and an on-demand exam. The performance standards describe what students should know and the ways they should demonstrate the knowledge and skills they have acquired. The performance standards include samples of student work that illustrate high-quality performances, accompanied by commentary that shows how the work sample reflects the performance standards (New Standards, 1994, 1996, 1997). According to Knowing What Students Know they go beyond most content standards by describing how good is good enough, thus providing clear targets to pursue (NRC, 2001, p. 251).

**Balanced Assessment Project**

Another influential project in the United States that clearly falls under the definition of large-scale assessment is the Balanced Assessment Project. The collaboration in-
volves not only the U.S. partners (initially Berkeley, Harvard, Michigan State) but also the University of Nottingham, U.K. The project “takes off where multiple-choice ends” (Figure 25.10; MARS website: http://www.nottingham.ac.uk/education/MARS/). Multiple-choice is seen as falling into the range of transforming and manipulating; what the project wants to do is performance assessment that cuts across ranges of mathematical processes, including modeling and formulating, inferring and drawing conclusions, checking, evaluating, and reporting.

**T-Shirt**

The aim of this assessment is to provide the opportunity for you to:

- systematically communicate about geometric shapes;
- locate shapes on a grid;
- give a clear set of directions.

The design below, including the 10 by 10 grid, is going to be used on a math team T-shirt. You accidentally took the original design home, and your friend Chris, needs it tonight. Chris has no fax machine, but has a 10 by 10 grid just like yours. You must call Chris on the telephone and tell him very precisely how to draw the design on his grid.

Prepare for the phone call by writing out your directions clearly, ready to read over the telephone.

![Figure 25.10 Grade 8 sample task. From MARS-website: http://www.nottingham.ac.uk/education/MARS/tasks/g8_1/.

**NCEM**

An interesting point that permeates almost all discussions about exit examinations is whether examinations or other large-scale assessments that can promote and stimulate worthwhile educational goals and can function as a high-stakes assessment at the same time. Another point is the fact that in many countries the examination tool becomes a tool for selection of the “best and the brightest.” In their article on reform on the national examination, Amit and Fried (2002) observed the struggle between society’s ambition for high achievement and its simultaneous ambition for mathematical literacy for all. On the one hand, because of the prominent place assumed by mathematics in professional and everyday life, it has become a critical filter, via higher education, for social and economic advancement (Sells, 1978). At the same time, and for the same reason, mathematical literacy is recognized as a basic societal goal, so that good mathematics education, therefore, becomes a basic right for all members of society. In their article Amit and Fried argued that it is possible to align the role of completion examinations in promoting the high quality of advanced studies with the original role as an agent of democracy. They based this conclusion on the new model for the National Completion Examinations in Mathematics (NCEM) in Israel.

The original NCEM consisted of three levels: Ordinary, Advanced, and Highly Advanced. The progression from one level to the next can be characterized by an increasing range of mathematical subjects, but also by an increasing demand for depth in the approach to mathematical ideas. But in the early 1990s Israel’s educational policymakers, mathematics educators, and researchers called for reform of the system because of the rising numbers of students not opting to take the ordinary level examination and therefore not getting the examination certificate.

The answer to these problems came in 1996 when both a new curriculum and a new ordinary level NCEM were introduced. The new examination comprised two separate questionnaires: the basic and the supplementary questionnaire. The first is oriented towards basic mathematical literacy, and the second towards a more complete mastery of mathematical concepts and techniques. The basic part can be taken and retaken independently of the supplementary part. This feature allows schools to develop special tracks geared towards students who will take only the basic questionnaire. It is this basic part that makes the new system work in merging the different objectives. Figure 25.11 gives an example of problems from the basic component (that can be taken when students are in 10th, 11th or 12th grade).

An evaluation by the Ministry of Education showed an increase in the number of students taking the basic examination, with the increase in the basic part being significantly higher than the number of students tak-
students are now enabled to continue to study mathematics throughout high school due to a rather simple intervention: more attention for the functionality of mathematics in the curriculum—and as noted before Israel is certainly not alone in this development—and consequential changes in the assessment system.

Let me make a concluding remark or two before moving on to technology in large-scale assessments. Large-scale and high-stakes assessments clearly can go together in an appropriate way. Regrettably, some

Example 1
The graph shows the price of books according to the number purchased (see figure above). Answer the questions according to the graph.

(a) How much do you pay when you purchase 25 books?
(b) How much does each book among the first 20 cost?
(c) How much does each book after the 21st book cost?

Example 2
Total 15 people were tested and the following scores were obtained: 8, 7, 7, 9, 8, 8, 9, 5, 7, 6, 6, 4, 6, 10, 8.

(a) Calculate the average score.
(b) What is the probability that the score of one of the people chosen at random will be higher than the average?

Example 3
Given a right triangle ABC in which angle D is right. The length of side AD is 40 cm. Angle BAD is 41º. C is a point on BD such that the angle CAD is 38º (see figure below).

(a) Calculate the length of the leg BD.
(b) Calculate the length of the segment BC.
large-scale assessments are “misused” as certain jurisdictions have adopted policies that rely exclusively on achievement test scores to make high-stakes decisions. Many standardized tests do not meet desired criteria, or, as Knowing What Students Know stated, current assessment practices are the cumulative product of theories of learning and models of measurement that were developed to fulfill the social and educational needs of a different time. The wish to refashion assessments to meet current and future needs for quality information is recognized beyond doubt, in many countries, and even implemented to a certain extent in some.

**LARGE-SCALE ASSESSMENTS, PROBLEM SOLVING, AND TECHNOLOGY**

As pointed out before one of the key issues in assessment and also large-scale assessment is not only the need to pay more attention to the mathematical content domain, but also in realizing the idea that assessments are based on modern knowledge of cognition and its measurements. One of the major problems in large-scale assessments is that it should be seen not only as a policy evaluation, but more as a tool for practitioners. Therefore assessment instruments should relate clearly to intended learning outcomes and can be used to improve students’ achievement. These points relate to two major threats to validity in high-stakes tests (Wilson, this volume). The first is construct underrepresentation and refers to a situation in which the concepts and skills that are measured in the assessment fail to include important dimensions or facets of the intended construct, in this case the content domain in mathematics. The second is construct-irrelevant variance (Brualdi, 2002). This can take the form of extraneous clues in the item or the task format that may permit some students to respond correctly or appropriately in ways that are irrelevant to the construct being assessed. Another possibility is that extraneous aspects of the task make it irrelevantly difficult for some individuals or groups.

Various technologies have been applied to bring greater efficiency, timeliness, and sophistication to multiple aspects of assessment design and implementation. Examples include technologies that generate items; immediately adapt items on the basis of the examinee’s performance; analyze, score, and report assessment data; allow learners to be assessed at different times and in distant locations; enliven assessment tasks with multimedia; and add interactivity to assessment tasks. In many cases, these technology tools have been used to implement conventional theories and methods of assessment, albeit more effectively and efficiently (NRC, 2001). It can even be argued that access to modern technologies has hindered the development of more valid tasks in respect to the earlier made points. Therefore the focus of the research relating assessments and technologies should shift towards these goals and challenges. As a matter of fact it seems that a precondition to develop valid and interesting tasks is the availability of modern technologies.

Knowing What Students Know provided a number of promising new developments varying from theory-based item generation, developing concepts maps (individual or in groups), complex problem solving in a variety of ways, analysis of complex solution strategies (using complex neural network technology), facilitating formative assessment, technology-based or technology-enhanced learning environments, and more. Many developments seem to be happening at very different levels, and many of the implications are unclear. One of the dilemmas we educators are facing is that we have to reinvent the definition of learning goals in these new environments: Are we abiding with the traditional curricular-based learning goals, or do we see quite new and unexpected horizons? At one side we see successful mathematical modeling competitions in which students work in teams for an extended period of time using technology in a variety of ways on quite a large scale. This might be considered as an existence proof for the merging of large-scale assessments and complex problem solving in groups in a somewhat traditional setting (just curricular mathematics) but where the role of technology is crucial.

An example of large-scale assessment in groups, among others, is the math A-lympiad (see Figure 25.12). This is a modeling competition for students from Germany, Denmark, and the Netherlands (with guests from other countries as well) that has been in existence since 1989 (De Haan & Wijers, 2000). Thousands of students work in many hundreds of teams on an assignment—a very open-ended problem—in which mathematical problem solving and higher order thinking skills must be used to solve a real-world problem. The result is a written report, just like in most other modeling competitions.

The A-lympiad has two rounds: the preliminary round competing at school for a whole day at their own schools, and the international final for the best 16 teams. This final takes most of 2 days. The aforementioned publication has an interesting chapter on the “Assessment of the Papers.” The conclusion is in line with earlier research on problem-solving assessment in mathematics: It takes time, effort and therefore is expensive, but the intersubjective scoring is quite reliable (De Lange, 1987). This is in line with the intermarker reliability found in the PISA study.
**Introduction**

A museum for modern art will organize a big exhibition with pieces of a number of great artists in the field of modern art. The organization is busy with preparations for the exposition, and meets with difficulties concerning security.

**A new security system**

The present security system with video cameras doesn’t satisfy the demands of an exhibition this size. The cameras aren’t movable enough (to show another corner), but the main objection is, that the present system cannot cover the complete space in the museum. To organize the exposition, it’s necessary to get a new security system.

The new type of camera that will be used is already chosen. This camera is so fast in moving (in all directions) and focusing, that one can say that this camera really secures the complete surrounding space (see the figure below).

All walls in the museum go from floor to ceiling.

A disadvantage of this camera is its high price. That’s why it’s necessary to investigate carefully where the cameras should be placed, for you want the least possible amount of cameras.

Enclosed is a map of the museum. The museum has a triangular formed top view. The grey part contains the entrance, bathrooms, cloakroom and office rooms. It’s not necessary to replace the old cameras in this part. The new security system is just needed for the part that is not grey.

**Exercise 1**

Think of a placement of cameras, in such a way that you use the smallest possible amount of cameras to secure the exhibition room. Describe the system you used to come to this placement of cameras. Mark where the cameras should be placed and show that the whole exhibition room will be secured this way.

The exhibition contains a little less than 100 paintings. To show these paintings in a proper way (with enough space in between), at least 280 meters of wall capacity is needed, but: the more wall capacity, the better.

On the map, you can see a number of bold walls. These are carrying walls; the other walls can be removed.

Because the new security system is very expensive, the organization wants to reduce the expenses. To do this, there is the possibility to reduce the number of cameras, but there is also the possibility to remove non-carrying walls. Removal of wall costs £ 500, per meter. A video camera costs £ 10,000, each.

**Exercise 2**

How is it possible through removing walls (keep the minimum amounts of capacity you need in mind!) to save expenses on the cameras? In other words: how is it possible, by removing as little wall capacity as possible, to make as many cameras as possible superfluous? And yet, all walls stay secured?

Make a proposal and calculate the savings that this solution will give you.

After this exhibition, the inside of the museum will be changed drastically. First, all non-carrying walls will be removed and just the carrying walls will stay. Because the new inside of the museum will be used for a longer period of time, and for several more exhibitions, an architect agency will be hired to create the new inside. The assignment will be to add 150 meters of wall capacity to the existing carrying walls. Furthermore, 6 cameras must be sufficient for security, yet, there must be an attractive partition of space.

**Exercise 3**

A number of architect agencies is being asked to hand in proposals in which the conditions are met. Your team works for one of the architect agencies, and your agency really wants the assignment. Of course you won’t be a strong rival if you just meet the minimum demands.

Make a proposal for the museum board (including working drawing and short ‘covering letter’).

---

**Figure 25.12** Assignment in preliminary round of Mathematics A-lympiad, 1996–1997 (De Haan & Wijers, 2000, pp. 59–61).
At the other side is complex problem solving in a less traditional setting: a MashpeeQuest assessment task gives students an opportunity to put complex problem-solving skills to use in a web-based environment that structures their work (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2000). In this example, technology plays at least two roles. The first is conceptual as the information analysis skills to be assessed, and the behaviors that serve as evidence are embedded within the web-based environment. The second is more an operational issue: Because the actions take place in a technological environment, some observations of students’ performance can be made automatically. Staying with the area of complex problem solving with technology, some interesting experiments were carried out in Germany by Klieme, Schümer, and Knoll in 2001.

Technology and Assessment: Thinking Ahead (NRC, 2002) argued that technology can help to merge the advances in cognitive sciences and measurement theories, which then should result in a significant leap forward in the science and practice of assessment. In the first paper Gitomer and Bennett (2002) illustrated how researchers at the Educational Testing Service use computer technologies to address a long-standing and very valid criticism of standardized tests. They tend to consist of certain types of traditional test items and have lost sight of the underlying constructs, or cognitive competencies. The authors argued that computer technologies can help “unmask” the constructs underlying traditional assessments, and to make the constructs more visible and explicit in the design of new assessments.

Another paper in this publication by Means and Haertel (2002) described, according to the introduction, the effort to develop computer-based, quality assessments of scientific inquiry. One of the key points they made is that in contrast to standardized tests and the more conventional paper-and-pencil tests used in most (science) classrooms, the new technology-based learning environments reflect the richness and complexity of scientific inquiry. It is at least interesting to make the observation that the authors and their colleagues saw very exciting opportunities in technology-based assessments in comparison with what they identify as ‘traditional standardized achievement test,” but failed to observe that most of the constructs or competencies can be measured with pencil-and-paper, with or without the use of technology. It is my opinion that the perceived contrast between “Innovative Technology-Supported Assessments” and “Traditional Tests” (page 15) is a construct of the assessment industry. Technology should not drive innovations in assessments, as has been the case so often in the past. The drive should come from advances in mathematical cognition, from advances in measurement theory, and should only be facilitated by the use of technology.

Eva Baker (2000) stated that the previous discussion has focused on the use of technology to make substantial improvements in our ability to assess and understand educational quality, principally measured by student achievement. But there is plenty of resistance to even this level of technology application. (p. 13)

Again, nothing seems new: There has always been plenty of resistance against innovations in assessments—looking back some 30 years shows that many efforts have been made to introduce more valid and fair assessments, with or without technology. Central should be what to measure and how to measure taking the kids and their competencies in the area of mathematics seriously.

Large-scale assessments have committed themselves to measure a wider variety of mathematical competencies. They have indicated that they will include more and more extended formats, and include technology use, starting with the calculator, but that they are also offering technology-based items (like Science in PISA 2006). Finally, high-stakes examinations can be done, although still on a voluntary basis, in a computer environment (Netherlands). There is mention of evaluating or measuring group work as this is regarded widely as an essential competency, and there are working examples in existence. Resistance exists, especially with policymakers, the people who pay for the large-scale assessment studies. But some progress has been made, and with educators’ combined efforts, further advances are around the corner.

REFERENCES


Garden, R. A. & Orpwood, G. (1996). Development of the TIMSS achievement tests. In M. O. Martin, & D. L. Kelly (Eds.), Third International Mathematics and Science Study


**AUTHOR NOTE**

In a chapter like this it seems important to make visible some of the author’s activities as a professional in this area. I have been on the National Dutch Committee for TIMSS and member of the Subject Matter Item Replacement Committee for TIMSS.

I am Chairman of the International Mathematics Expert Group of OECD/PISA, an advisory body that feels itself philosophically responsible for the mathematics framework. Furthermore I was Principal Investigator of the CATCH-project, and founder of the Math A-lympiad.

I would like to thank my colleague Truus Dekker and the reviewers John Dossey and Ken Travers for their valuable and constructive comments on earlier versions of this article. It is good to know that there are always colleagues willing to spend their time on work of others in order to get a better product.