Assessment: No change without problems

Jan de Lange
Freudenthal Institute, Utrecht University, The Netherlands

Changing mathematics education

Changing Goals

Mathematics education is changing fast - at least in a number of countries. In certain countries these changes have already taken place in the eighties - The Netherlands, Denmark, Australia, to name a few. Others are changing their mathematics education right now, in the early nineties.

One of the most visible discussions takes place in the United States. The Mathematical Sciences Education Board (1990) has summarised the changing conditions for school mathematics in terms of six points:

> *First*, as the economy adapts to information-age needs, workers in every sector - from hotel clerks to secretaries, from automobile mechanics to travel agents - must learn to interpret intelligent, computer-controlled processes.
>
> *Second*, in the past quarter of a century, significant changes have occurred in the nature of mathematics and the way it is used.
>
> *Third*, computers and calculators have changed the world of mathematics profoundly.
>
> *Fourth*, as mathematics has changed, so has American society. The changing demographics of the country and the changing demands of the workplace are not reflected in similar changes in school mathematics (MSEB, 1989).
>
> *Fifth*, learning is not a process of passively absorbing information and storing it in easily retrievable fragments as a result of repeated practice and reinforcement.
>
> *Sixth*, just as recognition of the global economy is emerging as a dominant force in American society, many recent reports have shown that United States students do not measure up in their mathematical accomplishments to students in other countries (e.g., Lapointe, Mead, & Phillips, 1989; McKnight et al., 1987; Stevenson, Lee, & Stigler, 1986; Stigler & Perry, 1988).

The changing conditions have led to changing goals. In The Netherlands, for instance, the goals (for the majority of the children) are:

1. to become an intelligent citizen (mathematical literacy);
2. to prepare for the workplace and for future education; and
3. to understand mathematics as a discipline.

These goals stated by the British Committee of Inquiry into the Teaching of Mathematics in Schools (Cockcroft, 1982). The goals reflect a shift away from the traditional practice. Traditional skills are subsumed under more general goals for problem solving, communication and critical attitude.

Changing theories

At the same time the goals of mathematics education are changing we also see the evolvement of new theories for the learning and teaching of mathematics.

At the Freudenthal Institute (formerly IOWO and OW&OC) the ‘theory for realistic mathematics education’ evolved after 20 years of developmental research which seems to be related to the constructivist approach (See Freudenthal, 1983, 1991; Treffers, 1987; de Lange, 1987; Gravemeijer et al, 1990; and Streefland, 1991). There are, however, some differences.

The social constructivist theory is in the first place a theory of learning in general, while the realistic mathematics theory is a theory of learning and instruction, and in mathematics only. One of the key components of realistic mathematics education is that students re-construct or re-invent mathematical ideas and concepts by exposing them to a large and varied number of ‘real world’ problems and situations which have a real world character or model character.

This process takes place by means of progressive schematisation, and horizontal and vertical mathematisation. Here, the students are given opportunities to choose their own pace and route in the concept building process. At some moment abstraction, formalisation and generalisation takes place - although not necessarily for all students.

The question, for instance, how far we can be successful within mathematics if our students ‘only’ master the skill of transferability instead of generalisability, is still open for discussion. We will not be surprised at all that different answers will appear for different student populations.

After the process of conceptual mathematisation the newly developed concepts are applied and used in ‘real world’ situations. This leads to reinforcement of the concepts and to adjustment of the student’s real world. lt goes without saying that (mental) construction and production play an essential role in realistic mathematics education, and it will come as no surprise that learning strands are intertwined and that student interaction is essential.

Changing content

Not only goals and teaching and learning theories have changed mathematics education. New subjects are slowly and sometimes reluctantly introduced in curricula, most clearly discrete mathematics, and there seems to be a revival at geometry. Some of these subjects enter the curriculum because of the fact that new technology has opened new possibilities. The computer has had some (limited) impact on the teaching of mathematics, but future development might have some more visible effects. A graphic calculator with a computer algebra system would outdate both personal computers and graphic calculators as we know them now.

Also, CD-I (Interactive CD) could enter ordinary households and become an important tool in education as well. Time will show if CD-I follows the development of the computer at schools or will really affect mathematics education in a dramatic way. But apart from these external factors there are internal factors too. We mentioned the revival of geometry but new insights give geometry at school level a different content as we will clearly see in our examples later on. But also the central place of calculus, the emphasis on fractions and percentages, the role of logarithms have been discussed in the last decade with mixed results.

Another internal factor can be formed by new insights how children learn and which didactical tools we have to make children understand better certain mathematical tools. Changing learning theories can definitely lead to new content subjects too.

Changing assessment

There seems to be a lot of truth in the conclusion of Galbraith (1991) when he says that the need to confront inherent contradictions that exist when constructivism drives curriculum design and knowledge construction, but positivistic remnants of the conventional paradigm drive the assessment process.

In The Netherlands we were confronted with this separation in a very hard way. Many teachers and researchers react with: ‘I like the way you have embedded your maths education in a rich context, but I will wait for the national standardised test to see if it’s been successful.’ Popper (1968) and Phillips (1987) have argued that a theory can only be tested in terms of it’s own tenets. This means that the constructivist or realistic mathematics education of teaching and learning can only be evaluated by assessment procedures derived from the same principle. Or: the assessment procedures should do justice to the goals of the curriculum and to the students - context independent generalised testing is unjust in such a case (most of the time, the context will also include the real world of mathematics itself; at least in the realistic mathematics education approach).

Therefore, an essential question is:

> Does assessment reflect the theory of instruction (and learning?)

On the other hand, not only the new notions about learning have influenced the ideas about ‘authentic’ assessment. The new goals also will have their effect. The new goals do emphasise the reasoning skills, communication and the development of a critical attitude. Together, these are popularly called ‘higher order’ thinking skills. These thinking skills were seldom or not at all present in traditional assessment and education. The change towards a ‘thinking’ curriculum forces us to focus on ‘thinking’ assessment as well. We will discuss the following points in some detail:

- Levels in assessment
- The role of the context
- Necessary and sufficient information
- Different formats of tests

In the final section we mention the following points:

- Time restrictions
- Individual or group
- Home or school
- Integratedness
- Objectivity of scoring
- Continuous or discrete.

We are aware of the fact that we can just touch the heart of the problems and have made no effort to make this article exhaustive.

Levels in assessment

Principles and Goals

lt is interesting to see that the publication that came out of the National Summit on Mathematics Assessment (*For Good Measure*, MSEB, 1991) states that their goals and principles are based on commonly held beliefs about assessment. Not only interesting but somewhat surprising if we see that the first principle is: ‘the primary purpose in assessment is to improve learning and teaching.’

Surprising because if we compare this statement with the actual school practice there seems to be hardly any relation with this first ‘commonly held belief’. The principle itself, which we share, is not new at all. Gronlund (1968) stated it clearly and we borrowed it in 1987 to formulate our principles:

- The first and main purpose of testing is to improve learning and teaching.
- Methods of assessment should be such that they enable the student to demonstrate what they know rather than what they don’t know.
- Assessment should operationalise all goals of mathematics education.
- The quality of mathematics assessment is not in the first place determined by its accessibility to objective scoring.
- The assessment tools should be practical. (de Lange, 1987)

As we indicated the first principle is easily underestimated in the teaching-learning process. All too frequently we think of it as an end-of-the-unit or end-of-the-course activity whose primary purpose is to serve as a basis for assigning course grades. A properly designed test or task should not only motivate students by providing them with short-term goals toward which they work, but also by providing them with feedback concerning their learning process.

Furthermore, more complex learning results such as levels of understanding, application and interpretation are likely to be retained longer and to have greater transfer value than the results at the knowledge level. This means that we should include measures of these more complex learning results in our tests. In this way we provide the students practice reinforcing comprehension, skills, applications and interpretations we are attempting to develop.

The second principle, sometimes referred to as positive testing, is borrowed from *Mathematics Counts* (Cockcroft, 1982). In traditional testing most of the time we check what the students do not know. Students are asked a very specific problem which has, most of the time, a single solution. If the candidate does not know the solution, there is hardly any way to show what the candidate does know. A side effect may be that the student loses confidence. And as *For Good Measure* states correctly: ‘the primary use... is to promote the development of the talents of all people’.

The third principle is that assessment tools should operationalise all goals of mathematics education. The fact that tasks that operationalise ‘higher order thinking skills’ are difficult to design and score should be no reason to restrict ourselves to the usual tests. Mathematisation, reflection, discussion of models, communication, creativity, generalisation and transfer are essential qualities of students to test. This means also that we are not always interested in the product but in the process that leads to this product. The consequence is that we need a variety of effective assessment methods - many of them are discussed in this paper.

The fourth principle is a very important one. Very often the quality of a test is derived from the accessibility to mechanical of objective scoring. This fact has caused many problems and is in part the reason for the poor state of mathematics education in the USA. lt may be difficult to score more complex tasks but experiences show that at the same time the advantages are much higher than the perceived disadvantages. In the first place: in complex context problems the problems are much clearer for the student to understand - they make the problem their own and the answers show at least really what the student is capable to do. In the traditional test we don’t even know if the student understands the question fully (examples follow), let alone that the answer gives us any indication about his understanding.

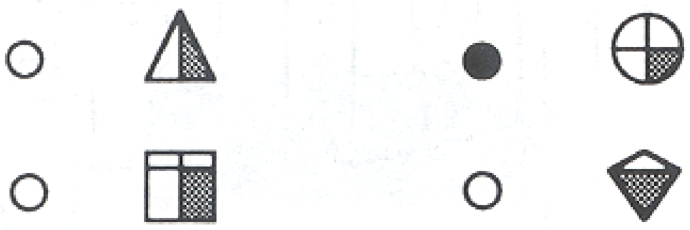
Another aspect is that the real professional mathematician is never judged by a test he has to pass but by papers he has to write. It remains unclear why we cannot use this tool at a somewhat lower level. A third point that can be made is that we will educate the mathematics education community by introducing new forms of assessment and the guidelines that will be developed for judging all forms of assessment. For the moment we leave this point here to return to it shortly in the final chapter of this article.

Our final principle is that assessment should have some degree of practicality in the school. This of course is still a very open statement but that is exactly the way we see it: at some schools a balanced package of assessment tools will be different from another school because of the physical limitations. Differences in school culture, accessibility to external sources etc. But we always should bear in mind the burden of the teacher.

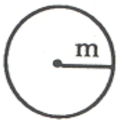
Lower level

We start at the lower level, because of the fact that here we recognise most of the traditional mathematics and the traditional tests. This level concerns ‘objects’, ‘definitions’, ‘technical skills’ and ‘standard algorithms’. Examples are abundant; we will give a few:

- Solve the equation $7x – 3 = 13x + 15$.
- What is the average of 7, 12, 8, 14, 15, 9?
- Draw the graph of $y = -x^2 – 2x + 8$.
- Which shows 1/4?



- Write 69 per cent as a fraction.



- Line m is called the circle’s...

Quite often some ‘multiple step’ problems from the real world fit into this lowest level, because in the books they are treated as standardised exercises that have no real problem meaning:

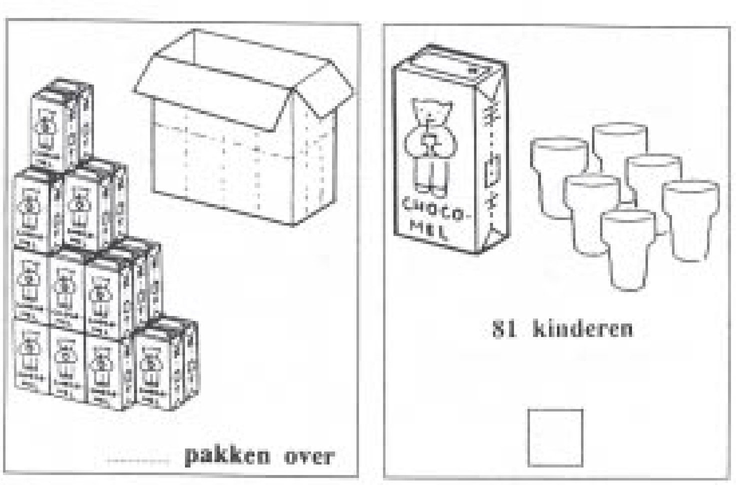
- Christine borrowed from the Friendly Finance Company US$168.00. She had to pay six per cent interest. How much is this in one year?
- We drove our car for 170 miles and used four gallons of petrol. How many miles per gallon?

Others will argue that these items belong at least at the middle level. At this moment we think it is hard to judge, because it depends on the way of instruction, the books used, the tests done before, the training for the tests and the age of the student. But the Christine exercise was part of a NAEP test for 16 year olds, and at this age level this exercise hardly fits into the middle level.

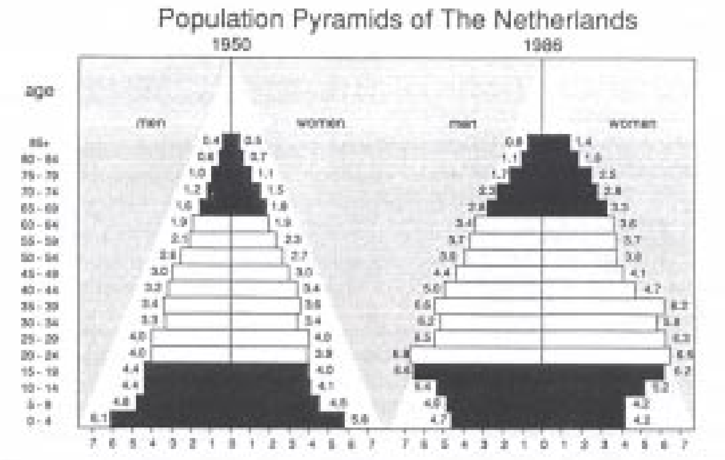
Middle level

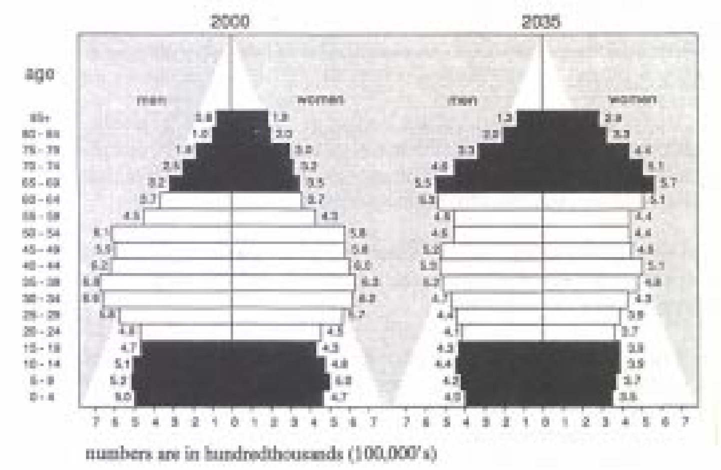
The middle level can be characterised by some key words as ‘making connections’, ‘integration’ and ‘problem solving’. lt is already harder to give examples at this level, because there is no real abundancy of good tests that operationalise this middle level. Some examples follow:

- You have driven 2/3 of the distance (in your car) and your tank is 1/4 full. Do you have a problem? (fifth grade)



- How many of the cartons will be left?
- How many cartons do we need for 81 children?

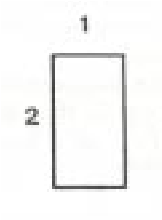




– What per cent of the age group [0,4] in 1950 will still be living in 2035?

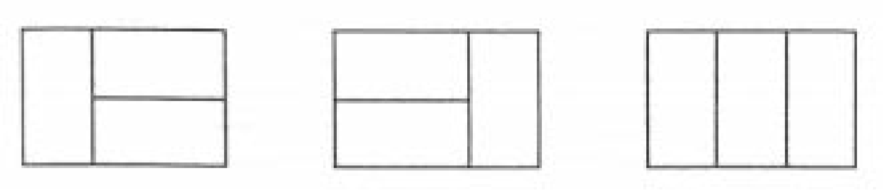
And another example:

– You have a supply of 2 × 1 rectangles like this one:



You can use these rectangles to make other rectangles which are 2 units deep and of whatever width you choose.

For example, there are some 2 × 3 rectangles:



– Describe how many 2 × n rectangles it is possible to make from 2 × 1 rectangles (where n is a natural number). Justify your conclusion.
– Extend your solution to describe how many 3 × 4 rectangles which can be made from 3 × 1 rectangles.
– Extend your solution further to describe how many m × n rectangles which can be made from m × 1 rectangles (where m and n are natural numbers).

Of course again we must stress the fact that the levels are arbitrary and so one can argue whether or not our examples really fit to the different levels. But the examples - all taken from real tests - show clearly aspects that do not belong on the lowest level, described earlier.

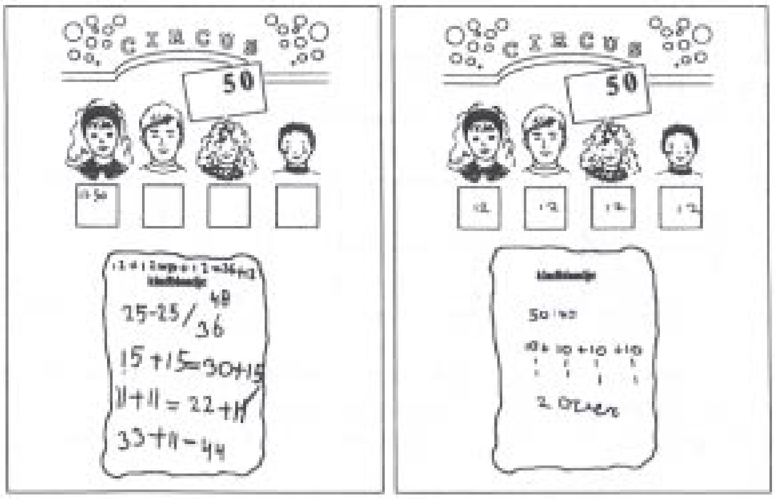
Higher level

lt is difficult to describe the highest level, even more difficult than the middle level. This, of course, is partly due to the fact that we are dealing with very complex matters: mathematical thinking and reasoning, communication, critical attitude, interpretation, reflection, creativity, generalisation and mathematising.

We will try to highlight some aspects of tests that try to operationalise some higher order thinking skills - at different school levels. A major component will be the ‘construction’ by children to complete the problem.

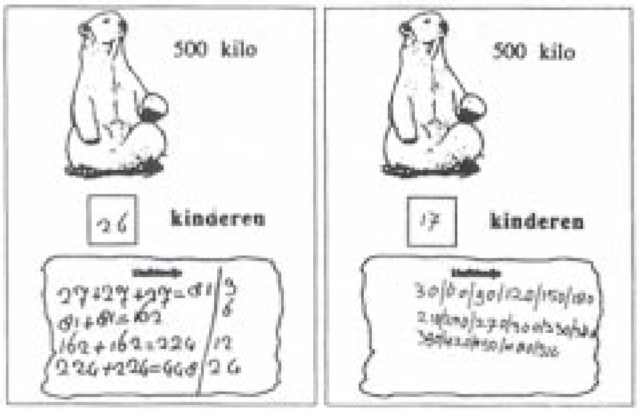
Let us first look at primary school level, with some more ‘open’ tests. Especially when we are dealing with non-algorithmic problems that relate to the student’s real world, we also need to know the procedures the children use. Or, even stronger, we are sometimes more interested in the process than in the product: the answer. Besides that, there might be multiple solutions. All of these arguments apply to the following test items:

> The first question relates to the visit of a circus: the total costs are $50. How much were the tickets?
> The second picture has the following question: how many children weigh the same as this bear?
> The third one asks the children to design as many sums as possible with the answer 100.
> Here, the children can use the piece of scrap paper pictured on the answer page. With regard to the circus item the following three figures show different strategies: (van den Heuvel, 1990).

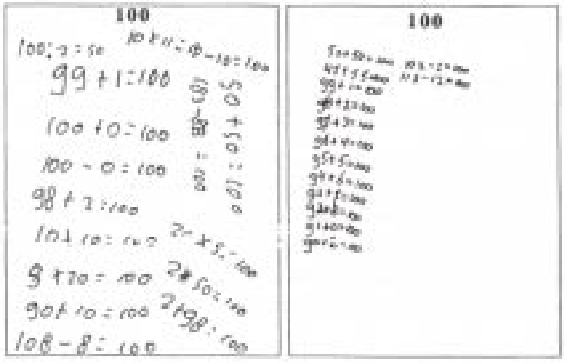


The first pupil tries to approximate the total amount of fifty as closely as possible, while the others immediately apply a formal division or less formal distribution strategy.
The bear item refers to the children’s knowledge of measures. Only the weight of the polar bear is given. It is left to the pupils to determine how much a child generally weighs. Some children, like the first pupil, stick to their own weight; others prefer a round number, or they weigh precisely

30 or 25 kilograms.



The third item (100) tries to elicit the capability of children for ‘own’ productions. The child is asked to think up rather than to solve problems (Van den Brink, 1987; Streefland, 1990). A simple way to estimate the scope of childrens’ abilities is the task to produce an easy and a difficult sum. Thanks to the latitude children are given in own productions it not only reveals what children are capable of, but also what their manner of working is. In this third item the children are asked to come up with as many sums with the result of one hundred as they can. There is not only a variety of numbers and kinds of sums, but also of working behaviour. Some record only isolated sums, whereas others proceed systematically, for instance, by always changing the first term by one unit or by applying commutativity.


100
99 + 1 = 100
100 + 0 = 100
100 - 0 = 100
98 + 2 = 100
108 - 8 = 100
100

Another item for primary school level that might fit on the highest level:

– Martin is living three miles from school and Alice five miles. How far apart are Martin and Alice living from each other?

This item might be seen as belonging to geometry. Or maybe not. Maybe just common sense reasoning. Or visualising. Multiple strategies are possible and also at different levels. But it is almost certain that the students have never seen an isomorphic exercise - or maybe one that is isomorphic but not clearly so to the students. Some 70 teachers were interviewed about the appropriateness of this item. Some of the first reactions of teachers to the question were: ‘5 – 3 = 2, so it’s a simple subtraction (lower level) and for that reason we don’t like it as a test item.’

A second reaction was: ‘You can’t say the proper answer because there is not one proper answer, and for that reason this is not a good test item.’

A third reaction was: ‘You can’t say the proper answer because there is not one proper answer, and for that reason it’s a good test item.’ A typical reaction falling in this third category: ‘You can’t tell it exactly, but you can say something. For instance that Martin and Alice cannot live farther away from each other than 8 kilometres, or not closer than 2 kilometres. You can show that with a nice picture.’

Looking at the item in this way makes it a very rich item offering many possibilities for different strategies reflecting the reasoning of the students. But the teachers reactions showed clearly that we have a long way to go if we want to implement this kind of question: of a group of teachers favouring developments towards more ‘realistic’ mathematics education only 17 per cent gave arguments like the one we just quoted.

The majority of teachers using more traditional books (57%) thought the item was unfit because of the lack of one single answer. This lack of one single answer was by far the most frequent argument under the teachers who where in favour of this item. lt shows how difficult the process of change towards ‘new’ modes of assessment will be (Gravemeijer et al., 1992).

There are many other ways to have children or students ‘produce’. A very interesting test item as part of a restricted time written test was the following:

- Choose a graphical way to show clearly the problem of the ageing population in The Netherlands.

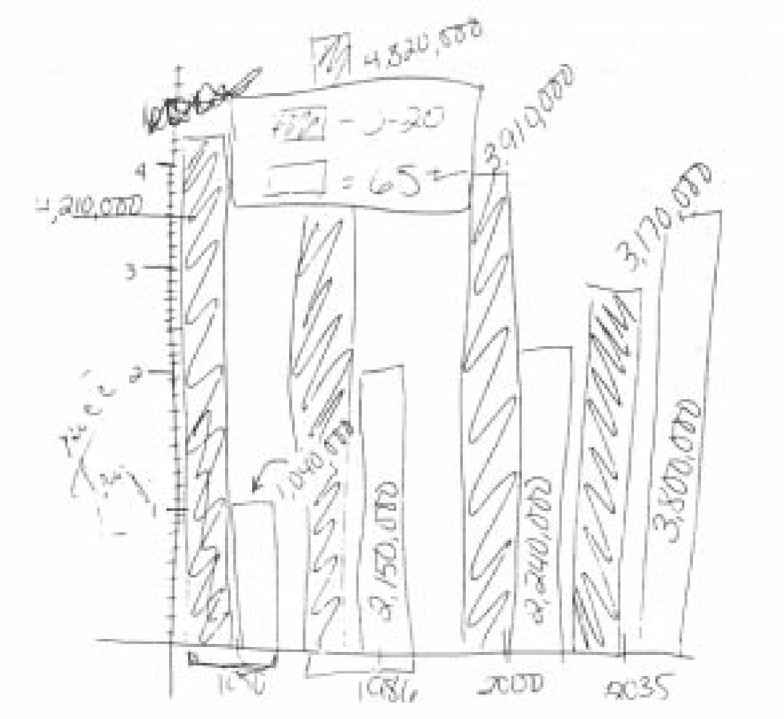
In this case the students were given only about twenty minutes to ‘solve’ this item. So time was a problem for the students. And there were others: obviously the students suffered badly from the fact that they were used to answer all questions on the test sheet. Trying to make a good graph without graph paper and on a small place (two to five inches) on your test sheet are not what you might call optimal conditions. Usually, in The Netherlands the students are given as many sheets, of any kind, as they need to complete the test.

Besides this organisational point there were other points of concern: the question was perceived as quite complex by the students. They had to:

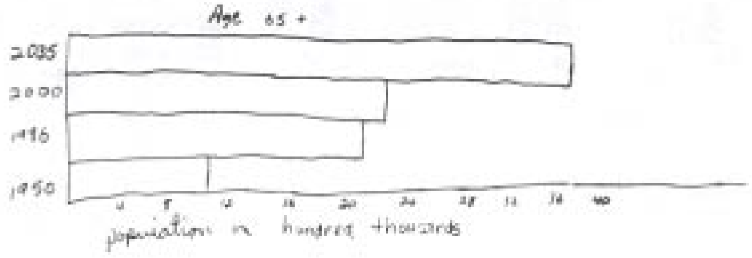
- understand the problem;
- compare the four or eight graphs;
- decide how to visualise (to construct or design);
- decide how to divide men/women;
- scale the graph;
- draw the graph.

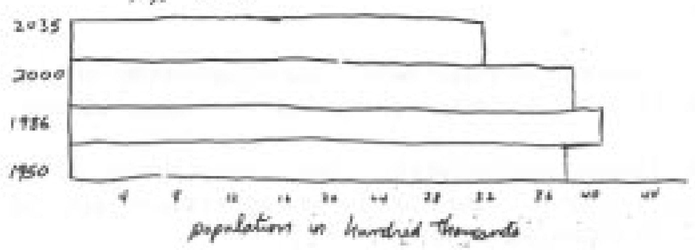
Given this complexity of the problem, the poor boundary conditions (graph paper lacking) and complete newness of this kind of question the results were not bad at all.

It came as no surprise that most students did choose some form of a bargraph. Some of the solutions were rather surprising. First a solution that shows the struggle that was caused by the lack of both graph paper and ruler:

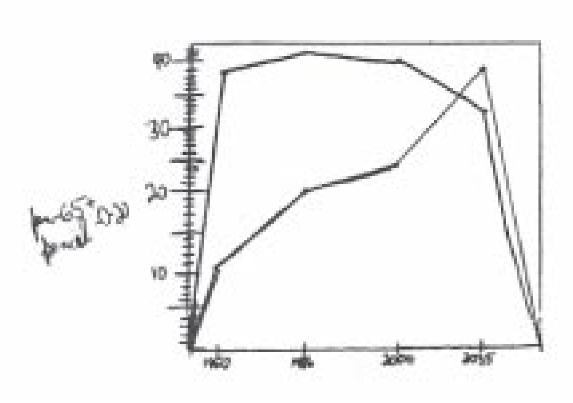


Others make two graphs: one for the 65+ group and one for the -20 group:

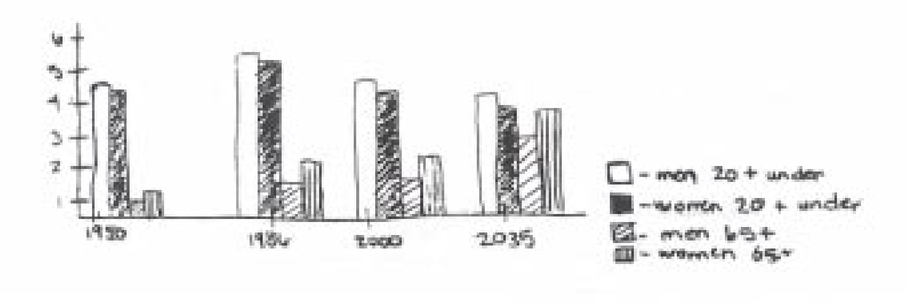




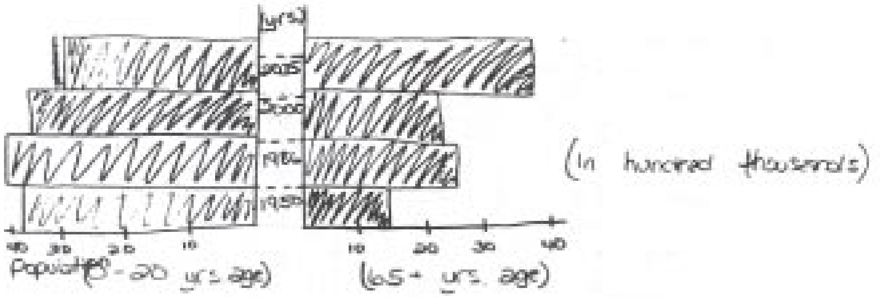
Interesting (for student and teacher) mistakes were abundant as well. The following linegraph is not bad except for the fact that the students had some fixation on the origin.



A nice graph is the following bar graph which does not only show the ageing population but also gives insight in the typical male/female distribution.



The next graph came as a surprise:



lt is an original way to draw the graph because the student designed his graph with the time axis vertical - in contrast with all we are used to. Breaking this convention the graph shows very nicely the ageing population problem of The Netherlands - in a very clear and ‘fair’ way.

Reflection on this test showed us quite a bit:
– the test was perceived as very unusual;
– the students were not used to reading a text;
– the students were not used to writing a text;
– the skills on percentages were not transferable to real problems;
– the student needs to have proper materials with him at all times;
– there was not enough time.

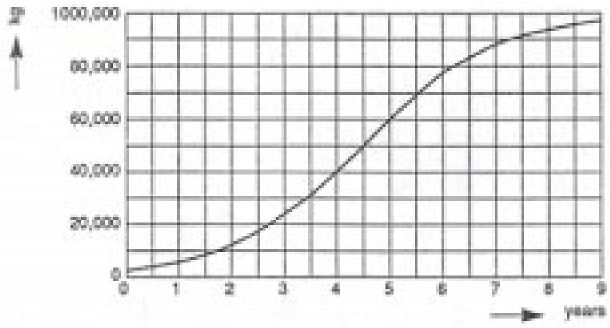
From the students’ perspective there were the following unusual aspects:

- the answer is not a number or a simple graph;
- you have to read carefully;
- you have to look at different graphs;
- you have to make decisions;
- you have to judge;
- you have to write;
- you have to think: there is no similar exercise in your textbook, so there is also no trigger mechanism;
- you have to produce (a story or a graph or a visualisation); there is not one right answer;
- you have to organise your work.

All together we might conclude that both the students, teacher and test designers learned a lot, most notably that assessment of higher order thinking skills requires a lot of higher order thinking skills.

A rather ‘simple’ item - as in contrast with ‘complex’ in the previous example - is the following:

- If no fish are caught the number of fish will increase in the coming years. The graph shows a model of the growth of the number of fish.



- Draw an increase diagram with intervals of a year, to start with the interval 1-2.

The fish farmer will wait some years before he’s going to catch fish. He wants to catch every year the same amount of fish as the first year. After every catch the number of fish increases again according to the graph.

- What would you like to advise the fish farmer about the number of years he has to wait after planting the fish and the amount of fish that he will catch every year?
  Give convincing arguments.

One should know, when looking at this exercise, that the students are not familiar with the subject ‘differentiation of functions’ but they do know the changes and rate of change context of real phenomena in a discrete way. Instead of the graph of the derivative of a function the students are accustomed to the discrete apparatus called ‘increase diagram’. So the first question is very straightforward and operationalises only the lowest level.

The other question is a different story. It was new both to the students and new in its form on the national standardised test in The Netherlands. Communicating mathematics, drawing conclusions, and finding convincing arguments are activities that all too often are not very visible in mathematics tests and examinations. Many teachers were surprised and didn't know what to think of this development, although they were prepared in some way, the experiments gave proper indications of the new approach.

Students seemed less surprised if we have to judge by the results, although their answers showed a wide variety:

> 'I would wait for four years and then catch 20 000 kilograms per year. You can't lose that way, man.'
>
> 'If you wait till the end of the fifth year then you have a big harvest every year: 20 000 kilograms of fish, that's certainly not peanuts. If you can't wait that long, and start to catch one year earlier, you can catch only 17 000 kilograms, and if you wait too long (one year) you can only catch 18 000 kilograms of fish. So you have the best results after waiting for five years. Be patient, wait those years. You won't regret it.' (Van der Kooy,1989)

The role of the context

Problem oriented mathematics education places mathematics in a context. Situated cognition is a well known, or at least well discussed subject nowadays. In realistic mathematics education the real world is used as a starting point to develop mathematical concepts and ideas. According to Treffers and Goffree (1985) context problems in 'realistic' curricula fulfil a number of functions, to wit:

- Concept forming: in the first phase of a course they allow the students a natural and motivating access to mathematics.
- Model forming: they supply a firm hold for learning the formal operations, procedures, notations, rules, and they do so together with other models which have an important function as supports for thinking.
- Applicability: they uncover reality as a source and domain of applications.
- Exercise of specific abilities in applied situations.

In an earlier article (de Lange, 1979) we tried to discriminate between the uses of context in a way that more or less fits with the four functions mentioned above. One of the functions, and for realistic mathematics education the most characteristic one, is the use of context for concept forming: the conceptual mathematisation process. This use of context has its own problems which are somewhat different from the problems in the other three classes of classification: usually we will not introduce new concepts during a test, but we apply the mathematical concepts in some way. So the three remaining classes of functionality are important when dealing with assessment:

- no function at all: no context
- context used to 'camouflage' the mathematical problem, or as Niss (1992) calls those problems: 'dressed up' contexts.
- Context is an essential and relevant part of the problem.

But a few other points need attention too. We will discuss the 'degree of reality' of a context, and how important it is or seems to be.

No context

This category hardly needs further elaboration. However we cannot pass this category without showing a recent example from a standardised test from Poland (1989) that shows a complex task without any context. The question to consider is at which level we are working here. Is this higher order because it is so complex or is it lower order but then repeatedly? Our impression is that it is the last one but we must admit that the example got us a bit confused.

Which number is 75% of:

$$\frac{\sin^2 30^0 - \left(\frac{1}{2}\right)^2 \cdot (0.8)^{-1} + \sqrt{2.25}}{\frac{11}{22} + \left(\frac{2}{3}\right)^2 \cdot (\cos 60^0 + \tan 45^0)^2}$$

Camouflage context

The context in this situation is only used to ‘camouflage’ or ‘dress up’ the mathematical problem. Most of the so called ‘word problems’ and so called ‘multistep’ problems from the NAEP are of this form. We refer for instance to the problem of Christine and her Friendly Finance Company. But there are many others.

Some problems similar to Christine’s:

- The growth factor of a bacterium type is 6 (per time unit).
  At the moment there are 4 bacteria.
  Calculate the point in time when there will be 100 bacteria.
- The interest percentage for a year is eight per cent.
  $4000 is deposited at 0 time.
  At what point in time will this amount have increased to $5000.

In this category there are also the well known items like:

- Bill weighed 107 pounds last summer. He lost 4 pounds, and then he gained 11 pounds.
  How much does he weigh now?

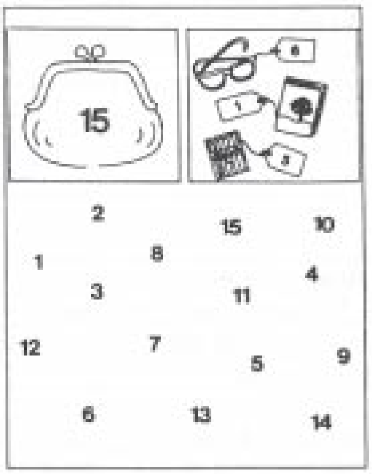
The goal that should be operationalised with the last item is:

- Identify, analyse and solve problems using algebra in equations, inequalities, functions and their graphs (Illinois State Board of Education, 1989). Although the problem certainly does not qualify higher than the lower level in this form it is interesting to notice that the item in its original form looks quite different and certainly does not operationalise the desired goal:
- Which one of the number sentences below could be used to solve the following problem?
  Bill weighed 107 pounds last summer. He lost 4 pounds and then gained 11 pounds.
  How much does he weigh now?
  a. 107 – (4 + 1) = A
  b. (107 – 4) + 11 = A
  c. (107 + 11) + 4 = A
  d. –4 + 111 = 107 + A
  e. (107) – 11) + 4 = A

The problem is not that we have to solve the problem, but that we have to analyses some kind of notation that no sane, let alone intelligent, citizen ever would use to solve the problem. An almost perfect example of a dressed up problem item that does not even reach its desired goals.

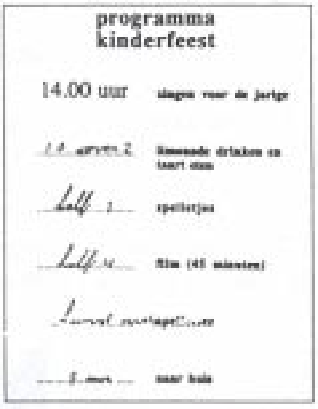
Relevant and essential context

Let us return once more to the tests developed during the MORE Project. We start with a simple test for Grade 1.



The children are asked to buy 'something' and to put a circle around the number that shows the money left in their purse. There are several degrees of difficulty here, the choice creates indications of what children are capable of. Of course preferences for a certain object play a part (relevant and essential context).

Experiments have shown however that quite a few children make numerically similar choices on tests of this kind which follow one another. Other types of problems show again the 'own production' aspect of tests coupled with relevant context use:



programma
kinderfeest

14.00 uur zingen voor de jarige

14 uur en 2 limonade drinken en taart eten

half 3 spelletjes

half 4 film (45 minuten)

kwart [illegible]

5 uur naar huis

The children are asked to make a program for a birthday party, or rather, complete it, as the starting time and the activities are already given. lt is left to the children to determine how long each activity will take. The only thing that is predetermined is 45 minutes for the movie. Like most open items this one allows a great many observation points: there must be progression of time and duration must be in tune with the activities and finally digital notation of time must be at least understood (Van den Heuvel, 1990).

But also in simple problem is, even in the multiple choice format the context can be essential like in the following example:

– Which of these would be a fairly good estimate for the width of a classroom?
  4 feet

10 feet
25 feet
300 feet

The size of the classroom is considered to be a good context for Grade 3, the length of the teacher’s desk for Grade 6. The question is of course if this is a proper way of discrimination by context. lt may be argued that the length of the teachers desk is easier to estimate than the width of the classroom:

- Which of these is the best estimate for the length of a teacher’s desk? (17)

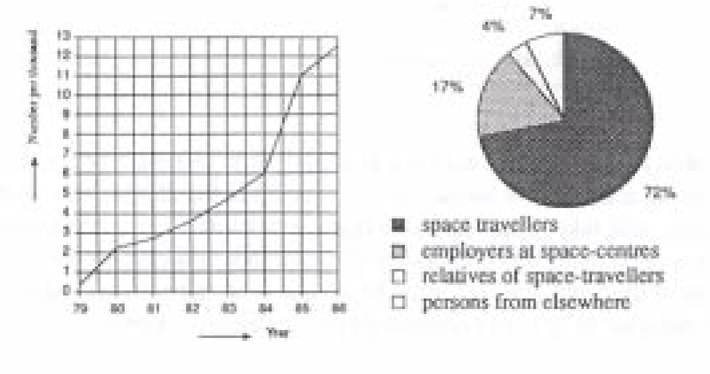
4 feet
10 inches
2 feet
15 feet
20 feet

Of course the inches might have caused some problems at Grade 3, but it would be interesting to give both items to all grades at primary level to see if the percentages of proper answers are about 64 per cent as was the case with both items in their respective grades. Or do we expect to see an increase in scores with the increase of grades? There are research results that paint a completely different picture.

Real versus artificial

It is clear from the discussion how the last problem was designed that the context was a real one and that the authors discussed the relevance and the reality of this problem for the group of students this examination was designed for (in general the problem was well received both by students and their teachers). But what to think of the following context:

When in the 21st century space travel increased considerably, a new disease came from outer space to earth. The graph shows the number of sufferers from this disease over the whole planet earth for the years 2079 till 2086.



- Draw a new graph on logarithmic paper for the number of patients.
- During which period is the increase in the number of patients nearly exponential?
- Compute up to one decimal point the yearly growing factor during this period.

The patients suffering from the disease were merely space travellers and employees at space centres. The pie chart shows the distribution of patients in 2086 over the planet Earth. The number of infected people in The Netherlands in 2086 was: Space travellers 60; employees at the space centre 5; relatives of space travellers 3 and other persons: 2.

- Make a pie chart for the situation in The Netherlands.
- Investigate if in 2086, among patients in The Netherlands there were significantly more space travellers than the 72 per cent for the whole Earth, with a significancy of one per cent.

In a hospital the disease is treated with the medicine R and R2. Every patient gets 600 milligrams Rl and 190 milligrams R2. Both medicines can be made from raw materials A and B. Every kilogram A yields 60 milligrams Rl and 10 milligrams R2 and every kilogram B gives 30 milligrams RI and 15 milligrams R2.

- Compute the minimal number of kilograms raw material (A and B together) needed for one patient.

The cost for A is $15 per kilogram and the costs for B is variable. The hospital tries to get the raw material for minimal costs per patient.

- Compute by which costs for B it is cheaper to make the medicines RI and R2 from A only.

The context is rather clear from the first sentences but we have given the complete problem in order to represent an honest picture of the problem. A first reaction might be: very, very artificial to talk about one century from here, to talk about a space related disease, it might even look like a dressed up problem because of the very anonymous information (like RI and R2 and A and B). So it is definitely not real for the students - apart from the fact that the mathematics is all too real for them - and the relevance leaves something to be desired too.

Nevertheless the information from the problem is very scientific and very relevant and real in the original source article. The years concerning the problem are 1979 till 1986 and the space disease was in reality AIDS. But the designers felt that it was not a very good idea to confront students in an examination condition with this highly emotional context. And here we are at the heart of the matter.

It seems clear that when we put much emphasis on mathematics education for preparing our citizens to be intelligent and informed citizens we have to deal with all kinds of real contexts. We have to deal with pollution with a very political component. We cannot avoid politics with a very subjective component. Traffic safety is quite an important matter with a very emotional component because many students know of causalities in their families. Health is maybe one of the most important issues at this moment for many people: the fitness trend is still very strong. But to discuss cancer, Alzheimers, heart diseases and for instance the effectiveness of certain treatments - whether or not in relation to the costs of health care as a political issue is a difficult matter.

Formats of tests

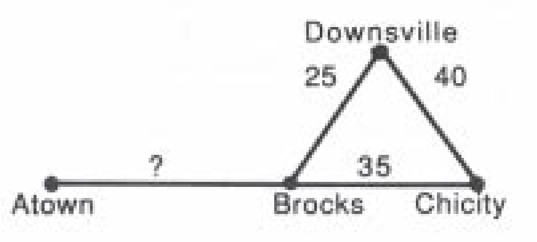
Multiple Choice

In constructing an achievement test to fit a desired goal the test maker has a variety of item types from which to choose. lt will come as no surprise that the multiple choice format seems to be the ‘best’ format if we just judge by its popularity.

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

The rules to construct a multiple choice item are simple. A multiple-choice item will present students with a task that is both important and clearly understood, and one that can be answered correctly only by those who have achieved the desired learning (Gronlund, 1968).

This is not as simple as it seems to be, as we all know, especially if we include that the item should operationalise a specific goal. To show how difficult the latter seems to be we give an example that was included (in slightly different wording) in the second IEA study - one might expect just the best items in that study:



John and Mary make a trip by car. They go from Atown to Brocks, then to Chicity to Downsville and back to Atown. The total trip is 190 miles.

– What is the distance from Atown to Brocks?
  a. 35
  b. 40
  c. 45
  d. 55
  e. 70

Of course the item would win if it was not the multiple choice format. But what is far more serious is that this item is meant to operationalise the goal: linear equations.

So the test-item designer is not only the test designer but also the solution designer. Many educators would like their students to solve this problem in their own way which in this case would seldomly include linear equations. Comparative studies have very limited value and the money spent on it is just for political reasons. Many people who are really interested in improving mathematics education and assessment that is appropriate could think of more efficient ways of throwing away money. But apart from that: the items are sometimes really not without flaws like the one above. And these are serious flaws. Because of the fact that in this way we don’t know any more what is being measured (which is not always bad) but we pretend that we do know (which is very bad). ‘American students are very poor in linear equations’ could be a meaningless statement if it was based on items like this one.

(Closed) Open Questions

Multiple choice items are often characterised as closed questions. This suggests that there are open questions as well. However we have to be careful. Sometimes the question is open by format but closed by nature. Let us illustrate this point with some examples.

– A How many diagonals has a rectangle?
  B How many diagonals has a square?

- The point (2, 3) lies in the first quadrant (1).
  Write down for each of the following points in which quadrant they are.
  A (–2, 2) b. (2, –4) c. (3, 4) d. (05, –6)

- Given is the equation $5x + 2y = 25$
  A Is (3,5) a solution?
  B Is (4,2) a solution?

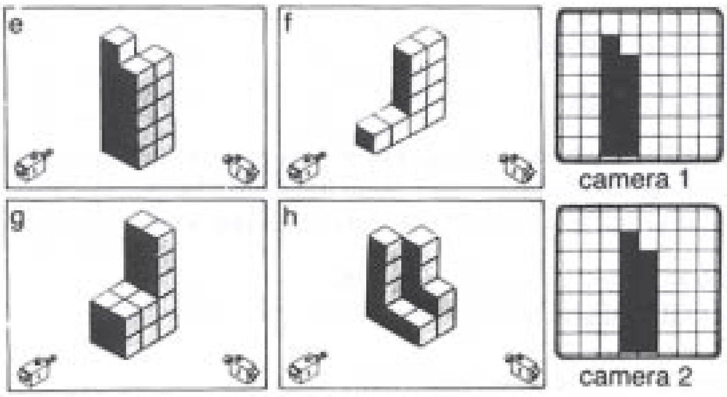
These examples are rather extreme examples of so called short answer questions, a subdivision of open questions. Although they are, technically speaking open questions, they are very closed by nature. The respondent has to answer by a number, a yes or no, a definition and maybe a simple graph or a formula. There is hardly any thinking or reflection involved. This category is mostly in close competition with the multiple choice format. The distinction between (closed) open questions and (open) open questions is rather arbitrary. This does not mean that we should not pay attention to this aspect when designing tests.

(Open) Open Questions

In our perception an (open) open question differs from the (closed) open question in respect to the activities involved to get a proper answer. This proper answer can still he just a number, or formula but the process to get there is slightly more complicated or involves higher order activities. This category differs from the next one - extended response open questions - in that respect that in the latter category we expect the students to explain their reasoning process as part of their answer. From both categories we have seen numerous examples (in this article), reason why we just give a couple in order to try to show the differences.

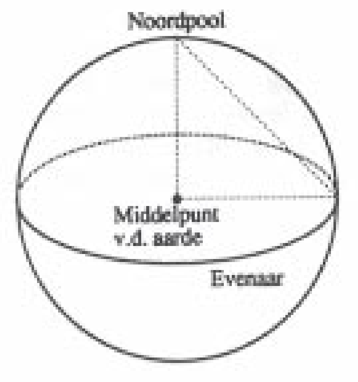
Examples for (open) open questions:

- A fruiterer buys a box of six pineapples for 60 cents. X pineapples are defective and therefore unsaleable, but the rest he sells at 18 cents each. Write down a formula for his profit P cents.
- A 400 metres running track is to have two parallel straights of 80 metres each and two semicircular ends. What should be the radius of the semi-circles?
- Which of the four buildings do you see on the TV?



- The length of the equator is about 40 000 kilometres.
  A How many kilometres is the distance from the north-pole to the equator, measured over the globe?
  B Imagine you can travel right through the earth. How far is it in that case from the north-pole

to the equator?



The following example is on the border of the (open) open questions category and the extended response open questions category:

– Explain which strawberry jam is sweeter:
  A jam made of 400 grams strawberries and 300 grams of sugar.
  B jam made of 1 kilograms strawberries and 700 grams of sugar.

The differences between the examples given in this category and the previous (closed) category are clear, if we base our judgement on the examples given. The closed questions referred to basis facts and knowledge: a definition, a simple drawing, a substitution not involving much of a process or thinking. The open questions, although also requiring a short answer, were just triggering questions, but questions that needed some thought and understanding and offered some possibilities for the student to solve the problem in his or her own way.

The last question - from a national examination - involves not only some reasoning but the students have to explain their reasoning. That 'writing down' part can be regarded as a separate goal in mathematics education, and a very valid one. If we stick to short answer questions we are not able to operationalise the communication goals. Extended response open questions do offer that possibility.

Extended response open questions

We will give one more example from a national examination, that is certainly of the extended response category but with even less freedom than the fish farm one. lt shows clearly how many different test formats we actually have at our disposal if we just focus a little bit more in detail.

– In fall the grapes that are ripe can be harvested. The taste of the grapes depends considerably on the moment that they are harvested. If they can be left out in the sun a little longer the taste will improve. But if one waits too long with the grape-harvest, there is a chance of damage caused by heavy and lengthy rainfall.

A grape farmer has the following choices for harvesting:

i Immediate harvest
The quality is 'reasonable'. Half of the harvest can he sold for direct consumption;
the proceeds in this case are $2.00 per kilogram.
The other half of the harvest can only he used for grape-juice; the proceeds for this part are $1.30 per kilogram.
In this way there is no risk involved in the harvesting.

ii Wait two weeks before harvesting

> The quality of the grapes in this case ‘is good’. The complete harvest can be sold for $2.30 per kilogram. But to wait two weeks involves some risks. If it rains more than two days during these 14 days, the grapes are that damaged that they can only be used for grape-juice; in this case the proceeds are only $1.30 per kilogram.

The grape farmer can he sure of a harvest of 12 000 kilograms. He chooses the risky second way of harvesting.

– Compute the advantages and the disadvantages that he has compared with the first strategy.

Meteorologists computed that for every day in this two week period, the chance for rain is 15 per cent.

– Compute the chance (probability) that it will rain in this two week period on more than two days.

The farmer chooses that way of harvesting that has the largest expected proceed.

– Which strategy will he choose? Illustrate your answer with a computation.

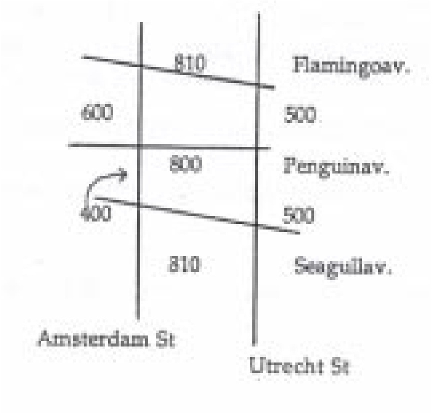
Most experts agree that we cannot operationalise all goals in mathematics education with tests of the multiple choice or open questions format. However they offer in our opinion more possibilities than are usually exploited. Properly constructed open questions, with a variety of short, long and extended responses do offer some possibilities for assessment at a higher than the lowest level - whatever name we give to the lower levels. They may be called knowledge outcomes and a variety of intellectual skills and abilities, or computation and comprehension, or basic skills and facts. Whatever the words, it is generally agreed that we need other instruments like essay tests that provide a freedom of response which is needed for measuring complex or higher order learning outcomes.

Essays

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

The most notable characteristic of the essay test is the freedom of response it provides. The student is asked a question which requires him to produce his own answer. The essay question places a premium on the ability to produce, integrate and express ideas. The shortcomings of the essay task are well known. lt offers only a limited sampling of achievement, the writing ability tends to influence the quality of the answer and the essays are hard to score in an objective way. Essays can come very close to extended response questions, especially in mathematics education. The example of the warning calling tree could be considered as an example of an essay just like the snow plow problem. This brings us immediately to an often mentioned aspect of the essay: should it be made at school or at home.

Usually the essay task is seen as a take-home task. However this is not necessary. One can easily think of smaller essay problems that could be made at school but require a day (or so). An example that was carried out at some 50 schools in The Netherlands is:



- In the map the Amsterdam street and Utrecht street are the main streets with a maximum speed of 30 miles per hour. At the intersections with the Flamingo, Penguin and Seagull Avenues are traffic lights. To get a smooth as possible traffic flow and to minimise irritation with traffic participators it is considered important that the waiting time for red lights is minimised. If a driver has a green light we should catch a green light at the next intersection just as he arrives and so on for the next light. In this case we talk of a 'green wave'. To get as close as possible to a green wave situation one has to consider for instance:
  - the duration of the green, yellow and red light periods
  - the relation between the different intersections
  - the indication of advised travel speeds.

  lt is known that a traffic light cannot be red for more than 90 seconds at a time and that each green period should be minimal 5 seconds. On the average the traffic on the main streets is four times as intensified as on the avenues.

Give advice as to how to adjust the traffic lights for each of the following subsequently more complex situations.

1. Bicyclists on the Utrecht street from north to south should have a green wave.
2. Both bicyclists and car drivers should have a green wave from north to south on the Utrecht street.
3. Both north and south traffic (cars and bicycles) on the Utrecht street should have a green wave as good as possible.
4. Both north and south traffic (cars and bicycles) on the Utrecht and Amsterdam streets should have a green wave as good as possible. Take also into consideration the traffic on the Flamingo, Penguin and Seagull avenues: this traffic should have a reasonable flow too.
   Final question:
   Which general principles and considerations should you use in a general and more complex situation.

This example was meant for students at upper secondary level - non mathematical majors. The students worked in groups of four from 9 am to 4 pm and were able to complete the task in a reasonable way. Especially the final question is an essay-like question - the other four are more or less extended response open questions to make sure that the students get started in the first place. Tasks like this one can also be made by individual students - at home or at school depending on the goal that has to be measured.

Production tests

If one of our principles is that the testing should be positive, which means that we should offer the students their abilities, and that the test is part of the learning-teaching process the 'own productions' offer nice possibilities. The idea of 'own productions' is not really new. Reports about experiences go back a long time. Treffers (1987) has introduced the distinction between construction and production, which according to himself is no matter of principle. Free production is rather the most pregnant way in which constructions express themselves. By constructions we mean:

- solving relatively open problems which elicit divergent production, due to great variety of solutions they admit, often at various levels of mathematisation, and:
- solving incomplete problems, which before being solved require self-supplying of data and references.

An example of the first: 'How to divide two bars of chocolate among four children?'
An example of the second: 'A radio message on a 5 kilometres queue at Bottleneck Bridge: How many cars are involved in the queue?'

The construction space for free productions might be even wider:

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

An example: Think out as many sums as you can with the result three (Grade l). We will describe in some detail the experiences that took place in an American ninth grade on the subject data-visualisation. The students had been working (suffering, interacting, thinking, discussion etc.) for two weeks on a text designed in The Netherlands with the philosophy of realistic mathematics education in mind. A philosophy, we think, fits reasonably well to the philosophy of the Standards as we tried to explain in the introduction. The booklet (*Data-Visualisation*, de Lange & Verhage, 1991) was intended for about five weeks of class activities. Some examples were shown earlier in this paper.

After two weeks the students got their 'final' test:

- This task is a very simple one. At this moment you worked yourself through the first two chapters of the book and made a relatively ordinary test (the test among other problems had the problem of the ageing population of The Netherlands). This task is quite different:
  Design a test for your fellow students that covers the whole booklet.
  You can start your preparations from now on: look at magazines, books, etc. for data, charts, and graphs that you want to use. Write down ideas that come up during school time. After finishing the last lesson of this booklet, you will have another three weeks to design your test.
  Keep in mind:
  - the test should be taken in one hour
  - you should know all the answers.

lt is tempting to show many exciting examples of the students' productions. To be honest, there were disappointments as well. One student just took a reasonably well chosen collection of exercises from the booklet, avoiding any risk taking or creativity. The next 'higher' design strategy is the one that mathematics teachers often use: take examples from the textbook make small changes (in exponents, coefficients or maybe context) and your test is ready. This worked for some students as well although the answers made sometimes painfully clear that these were not the brightest students if we looked at their proposed answers.

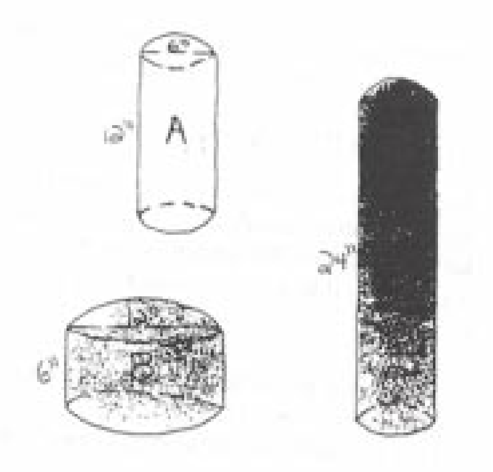
Our first example shows an exercise that operationalises the lower level:

- The following are the top 20 point leaders of the Edmonton Oiler team. Draw Stem & Leaf.

WG	208	DL	32
JK	135	KL	26
PC	121	PH	25
MK	88	KMC	23
GA	81	RG	23
MN	63	DJ	20
MM	54	DS	18
CH	15	LF	18
DH	36	BC	17
WL	32	JP	12

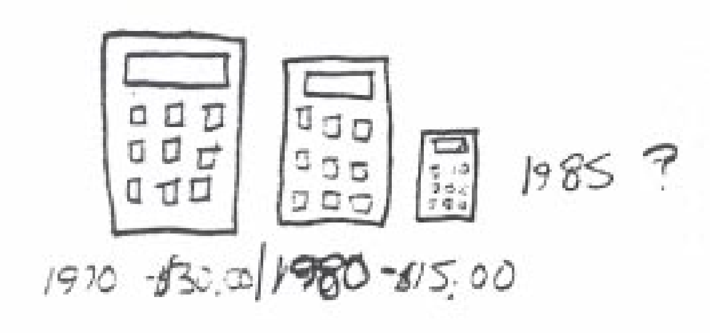
- What is the average for the 20 pieces of data?
  (4.13)
- What is the median?
  (2)
- What is the difference between the average and the median?
  (22,15)
- Why is the average higher than the median?
  (Some people had extremely high points).

The next example is interesting because it tackles the problem of misrepresentation that often occurs in pictographs. Two or three dimensional objects are used to represent one dimensional fact. So the subject ‘fair’ or ‘honest’ graphs got some attention in the booklet. One of the students made the following exercise:



- Does B or C show the volume of A doubled?
  Which cylinder, B or C, shows the volume of A more clearly?
  Why?

Another exercise on the same subject:
- How much would a calculator cost in 1985?



Some More Points of Attention

Designing tests and carrying them out in ordinary classroom practice is not a simple task as we have seen. When designing one has to be very clear which goals are being operationalised, which context to choose, the formats to consider and the practicality of carrying it our in the classroom. But there are other points that need serious consideration when choosing a balanced package of assessment tools. And they are rather obvious.

- Is the test to be taken within a fixed time interval: restricted time test?
- Is the test to be taken individually or in groups?
- Is the test to be taken at home or at school?
- Is the test a single strand test, an integrated test or even an inter-disciplinary test?
- Is the test part of a continuous assessment practice or is it a part of a more discrete scenario?
- For some people the most important question: how objectively scoreable is the test and which tools do we have in making the scoring as good and fair as possible?

Assessment: no change without problems

Let us return to the title of our article and see what conclusions we can draw. The first meaning of the title will he obvious. Assessment should have real problems which often will mean real world problems and applications. Having said that, we are immediately confronted with a number of other problems that we have to tackle and solve in the first place. To name a few:

- Teachers, test designers, parents, administrators, public officials and citizens need a new attitude towards assessment. This point is often underestimated. We cannot just rely on Assessment Summits and publications as ‘For Good Measure’ and all kind of interesting developments in assessment. Society has a certain image of assessment (SAT, CAT, etc.) that will take years to change and improve. The damage done by the assessment practices especially in the United States will take at least a decade to repair and no quick fixes are available. Nor are cheap solutions.
- Different levels of mathematical activities need different assessment tools which are hard to design and need a lot of research and testing.
- To design a balanced package of assessment will be difficult.
- To interpret the different strategies and processes that the students will show in more open assessment will be hard for teachers. Teacher training with special emphasis on assessment is not only needed but will really make teachers understand the problems we are dealing with.
- Different problems need different contexts with all kinds of variables to take into account (as discussed earlier). A special problem is to find the balance between a good context and a good mathematical problem.
- Scoring and judging the quality of all forms of assessment will be more complex and varied and will be considered to be more difficult than at present.

All of these problems may come as a surprise to a certain degree but if that is the case just shows how bad the situation is at the present moment. Assessment practices have lost their major participants: the students and the mathematics curriculum. Indeed: no change without problems.

This article has been published in Max Stephens & John Izard (Eds.). Reshaping Assessment Practices: Assessment in the Mathematical Sciences under Challenge. Proceedings from the First National Conference on Assessment in the Mathematical Sciences. Geelong, Victoria, 20-24 Nov. 1991, 46-76. Reproduced by permission of The Australian Council for Educational Research Ltd.. Copyright © 1992 ACER..

REFERENCES

Brink, J. van den (1987). Children as arithmetic book authors. *For learning of mathematics, 7* (2).

Burkhardt, H. & Resnick, L.B. (199l). *A balanced assessment package*. Nottingham: Shell Centre.

Cockcroft, W.H. (1982). *Mathematics counts: Report of the Commission of Inquiry into the Teaching of Mathematics in Schools.* London: Her Majesty's Stationery Office.

Lange, J. de (1979). Contextuele problemen. *Euclides 55.*

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

Lange, J. de, Reeuwijk, M. van & Burrill, G. (1991). *Learning and testing mathematics in context - the case: Data visualization.* Madison, WI: National Center for Research in Matheniatical Sciences Education.

Lange, J. de & Verhage, H.B. (1991). *Data-visualization.* Scotts Valley: Wings for Learning/Sunburst.

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht: Reidel.

Freudenthal, H. (1991). *Revisiting mathematics education*. Dordrecht: Kluwer.

Galbraith, P.L. (1991). Paradigms, problems and assessment: Some ideological implications. *ICMI study on assessment.*

Gravemeijer, K., Heuvel-Panhuizen, M. van den & Streefland, L. (1990). *Contexts, free productions, tests and geometry in realistic mathematics education*. Utrecht: OW&OC.

Gravemeijer, K., Heuvel-Panhuizen, M. van den & Streefland, L. (1992). *Methoden in het Reken-Wiskundeonderwijs, een rijke context voor vergelijkend onderzoek*. Utrecht: Center for Science and Mathematics Education.

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

Grossman, R. (1975). Open-ended lessons bring unexpected surprises. *Mathematics Teaching 71.*

Lapointe, A.E., Mead, N.Z.A. & Phillips, G.W. (1989). *A world of difference: An international assessment of science and mathematics.* Princeton NJ: Educational Testing Service.

McKnight, C.C., Crosswhite, F.J., Dossey, J.A., Kifer, E., Swafford, J.O., Travers, K.J. & Cooney, T.J. (1987). *The underachieving curriculum: Assessing United States school mathematics from an international perspective.* Champaign, IL: Stipes Publishing.

Mathematics Sciences Education Board (MSEB) (1990). *Reshaping school mathematics: A philosophy and framework of curriculum.* Washington D.C: National Academy Press.

Mathematics Sciences Education Board (MSEB) (1991). *Four good measure. Principles and goals for mathematics assessment.* Washington D.C.

Niss, M. (1992). Assessment of mathematical applications and modelling in mathematics. Paper presented at ICTMA 5 conference.

Phillips, D.C. (1987). *Philosophy, science and social inquiry.* New York: Pergamon Press.

Popper, K. (1968). *Conjectures and refutations.* New York: Harper.

Stevenson, H.W., Lee, S.Y. & Stigler, J.W. (1986). Mathematics achievement of Chinese, Japanese and American children. *Science 231:* 693-699

Stigler, J.W. & Perry, M. (1989). Cross-cultural studies of mathematics teaching and learning: Recent findings and new directions. In D.A. Grouws, T.J. Cooney & D. Jones (Eds.). *Perspective on Research on Effective Mathematics Teaching*. Reston, VA: National Council of Teachers of Mathematics pp. 194-223.

Streefland, L. (1990). Free productions in teaching and learning mathematics. In K. Gravemeijer, M. van den Heuvel-Panhuizen & L. Streefland. *Contexts, free productions, tests and geometry in realistic mathematics education*. Utrecht: OW&OC.

Streefland, L. (1991). *Fractions in realistic mathematics education*. Dordrecht: Kluwer.

Treffers, A. & Goffree, F. (1985). Rational analysis of realistic mathematics education. In L. Streefland (Ed.). *Proceedings of PME-9.* Utrecht: OW&OC.

Treffers, A. (1987). *Three dimensions. A model of goal and theory description in mathematics education*. Dordrecht: Reidel.

Heuvel-Panhuizen, M. van den (1990). Realistic arithmetic/mathematics instruction and tests. In K. Gravemeijer, M.

van den Heuvel-Panhuizen & L. Streefland. *Contexts, free productions, tests and geometry in realistic mathematics education.* Utrecht: OW&OC.

Kooij, H. van der (1989). Het eerste hawex-examen. *Nieuwe Wiskrant. Tijdschrift voor Nederlands WiskundeOnderwijs 9* (1). Utrecht: OW&OC.