ASSESSMENT OF MATHEMATICS ACHIEVEMENTS: NOT ONLY THE ANSWERS COUNT

Marja van den Heuvel-Panhuizen
Freudenthal Institute, Utrecht University, The Netherlands
Catherine Twomey Fosnot
City College of New York, New York, USA

Abstract

This paper addresses a study on assessing student learning in mathematics education. The study is connected to a teacher enhancement project at primary schools in New York City. The eventual goal of the project was to improve the students' mathematical achievements. Among other things, this goal includes the use of clever ways of applying mathematics for solving problems. Because standardized tests are not an adequate tool for revealing this kind of student learning a test was developed by means of which the process of mathematization was supposed to become visual. Results gained from this test in 17 classrooms with students ranging from grades 3 through 5 show that strategy-focused assessment is a valuable extension of the regular answer-focused assessment.

Introduction

The dichotomies product-process and answer-strategy play an important role in the assessment of mathematics achievements. If not earlier, this became apparent in the recent worldwide reform of mathematics education. One of the characteristics of the new approach to mathematics education is that it gives more room to processes and strategies. Nevertheless there is yet a large distance between both ends of the respective dichotomies. The answers belong to the product side and the strategies to the process side, and answers and strategies each have their own purposes. The gap is reflected above all in the difference that still exists between standardized tests and tools for classroom assessment.

Although the ways in which students solve mathematical problems are regarded as important information for educational decisions, solution strategies are scarcely covered by standardized tests. More often than not achievement scores are purely related to the number of correct answers. In tests, the answer is usually considered as the ultimate indication of the achievement level. The strategies that the students applied to find this answer are generally beyond the scope of standardized testing and belong more to the field of interviews and observations.

For a long time, only within the context of diagnostic testing, attention was paid to strategies. Data about the process of solving problems, and especially data about *errors*, was used to identify students' misunderstanding and find indications for remedial teaching. Errors in strategies were regarded as "windows on children's internal thought processes" (Baroody, 1987).

As said earlier, the reform in mathematics education has opened the door now for strategies as a more general focal point of assessment. They are no longer restricted to a remedial setting. As expressed in standards and curriculum documents of many countries, the new ideas about mathematics education emphasize that in addition to product information, process information is needed to get insight in the students' thinking. This particularly implies that special attention is paid to the *various ways* by which the students solve mathematical problems. A wide range of alternative assessment tasks and formats has been developed to reveal strategies. The main purpose of this alternative assessment is to inform the teachers about the students' way of thinking in order to provide clues for further teaching.

In other words, there is, nonetheless, a distinction between process and product. Strategies are still considered as a process variable and the answers are still the pivot of all. They are considered as the real output variable that counts for determining the achievement level. The question is, however, how tenable this is. Is this strict distinction not an unnecessary curtailment of the assessment of mathematics achievements?

The present paper cannot answer this question extensively. It only reports about a study in which was tried to use the students' strategies as an output variable, or – adapting Baroody's words – as "windows on achievements level".

The MiTC project

The study was part of the "Mathematics in the City" (MiTC) project. The MiTC Project is a large-scale project on teacher enhancement project funded by the National Science Foundation and the Exxon Educational Foundation (see Fosnot and Dolk, 2001). The City University of New York and the Freudenthal Institute of Utrecht University carried out the project jointly. Over a period of five years, the project worked on systemic reform in mathematics education from pre-kindergarten through grade 5 in five school districts in Manhattan. The project started with ten schools and in the last stage of the project about forty schools were involved.
The two key points that characterize the MiTC project are its model of improving teachers' classroom practice and its idea about mathematics education.

Approach to teacher enhancement
The heart of the project consisted of a very intensive collaboration between staff members of the project and the participating teachers. The mathematics classrooms in the project were co-taught by the project staff and the school teachers. Through joint observations and discussions between teachers and staff members the team worked continually on the development of better ways of teaching. By means of a two-week summer institute in the beginning of the school year, followed by weekly-organized institutes during the year, the teachers were offered opportunities for further deepening their professional knowledge and abilities. The structure of the project was that after a year of participation some of the teachers became staff members, which implied that they coached new teachers in the project. This mushrooming pattern meant that the initial number of forty involved teachers rose up to four hundred.

Approach to mathematics education
The project was strongly related to both the constructivist and the "realistic" view on mathematics education. Moreover, a strong basis was found in the Piagetian and Vygotskian theories on children's cognitive development. These two theories actually constitute the

socio-constructivist to mathematics education (see Cobb, 1996).

The socio-constructivist view in the MiTC project is recognizable in the fact that the students are seen as active learners within the social community of the classroom. The mathematics "congresses", in which the students, while seated on a rug, share their thinking and build up new understanding, reflect this approach to mathematics education exceedingly. In addition to this, within MiTC classrooms the Piagetian concepts of assimilation and accommodation play an important role. Teachers have to recognize the present cognitive schemes and structures of the students and must learn how to design problems and situations that can evoke new or re-ordered schemes and structures.

The key beliefs about what should be taught, and how it should be taught, are heavily inspired by Realistic Mathematics Education (RME), which is based on Freudenthal's view that mathematics must be connected to reality, stay close to children and be relevant to society, in order to be of human value (Freudenthal, 1977). Instead of seeing mathematics as subject matter that has to be transmitted, Freudenthal stressed the idea of mathematics as a human activity. Education should give students the "guided" opportunity to "re-invent" mathematics by doing it. This means that in mathematics education, the focal point should not be on mathematics as a closed system but on the activity, on the process of mathematization (Freudenthal, 1968).

Actually this mathematizing ability, including also attitude to mathematize, is the overall core goal of mathematics education. In short, this ability involves that students can use and develop mathematical tools, including models and strategies, which can help to organize and solve real life problems and pure mathematical problems.

Evaluation of student learning

In year five of the project an evaluation was planned. Because of the large-scale character of the project with its focus on change at different levels, i.e. students, teachers, and whole schools, a uni-dimensional evaluation was felt to be insufficient. Instead a multi-focus assessment was designed including macroscopic and microscopic lenses (Van den Heuvel-Panhuizen, Dolk, Fosnot, and Glick, 2000). This approach to assessment was based on the belief that the different perspectives would provide a fuller understanding of the effectiveness of the in-service intervention. The assessment contained the following three foci: student learning, teacher change, and school change. Each of them has various aspects to be examined. The present paper will only deal with the assessment of student learning and will concentrate on student competence deduced from the strategies they applied to solve the problems.

The main question to be answered about the student learning was what the project students gained from the project. To answer this question the achievements of the students from grades 3, 4 and 5 were compared to the achievements of non-project students.

Comparison by means of the standardized tests

The first analysis was done based on the results from the Corrected Terra Nova Math Test (for third-grade and fifth-grade students) and the Corrected New York State Math Test (for fourth-grade students). Although these standardized tests are not entirely in line with the project goals of mathematics education, it was decided to have the test scores from these tests as one criterion for evaluation. If students in reform project do not score well on mandated tests, the reform will fail simply because of political forces at play regarding testing. Moreover, results on sub test items can be helpful in understanding how well the reform project appears to be aligned with standardized achievement outcomes. The comparison of the tests scores revealed that in all the three grades the project students (experimental group) scored significantly better than a group of non-project students (control group). A covariance analysis controlling for entering score at the beginning of the year (which only was possible for grades 4 and 5) showed also significantly better results for the experimental group (Fosnot, Dolk, van den Heuvel-Panhuizen, Hilton, Wolf, and Bailey, under review).

Comparison by means of the MiTC test

From the very beginning of the project it was planned to have an additional evaluation of student learning more aligned with the goals of the project than the regular standardized tests mostly comprised of closed and answer-based items.

For this purpose an alternative test on number sense was developed for grades 3 through 5 using a RME assessment paradigm (van den Heuvel-Panhuizen, 1996).

Assessing mathematizing

Assessment grounded in the RME theory of mathematics education requires that the assessment tool must provide information about the process of mathematization, since that is considered as the overall core goal of mathematics education. The ability of mathematizing involves that students can solve problems by means of mathematical knowledge and concepts that can lead to models and strategies that are adequate for the problem to be solved. This goal is different from the goal of most traditional assessments, where strategies are assessed but only as process, or as a bi-product, or as an additional proof that the student had a clear understanding of what he or she was doing. In contrast, when mathematizing becomes the goal of instruction, the strategies are seen as an outcome, as the goal itself.

Within the process of mathematization several levels can be distinguished, each reflecting a certain level of the students' understanding. Regarding the content domain of number sense related to operations with numbers up to one thousand the mathematization includes concrete or mental activities ranging from carrying out a standard algorithm to making use of number relations and properties of operations in order to find shortcuts and clever and elegant procedures by means of which the answers can come across. Crucial for evaluating the mathematizing activity is to what degree the strategies show a level of maturity that includes both flexibility and effectiveness. Simply carrying out a standard procedure without taking into account the numbers involved is not judged as a high level of mathematizing.

For paper-and-pencil assessment to capture genuine mathematizing it is necessary that the students' own mathematical activity becomes visual on paper. Therefore all problems were put in a work area that the students could use as scratch paper. Another requirement for assessing mathematization is that the tasks allow several levels of mathematizing. The most important requirement, however, is that there *is* something to mathematize.

The following two problems may illustrate the kind of problems that were used in this test. The long addition problem that was in the test for grade 3 (see Figure 1) looks like an ordinary bare number problem and can trigger an algorithmic procedure. A student with number sense, however, will recognize how nicely the numbers fit together, and will adapt her or his strategy to this knowledge.

$$\mathbf{38 + 39 + 40 + 41 + 42 =}$$

Figure 1 Problem included in the MiTC test for grade 3

In the problem on the chain of beads (see Figure 2) the students have to figure out the color of the 1000th bead. In addition to this the students have to explain why they are sure about their answer. In this problem, that is from the test for grade 5, the students can apply their knowledge of multiples of three. But other, less advanced ways of working are also possible. Again, this inherent multi-level quality of the problem makes it very suitable for providing information about the students' achievements.

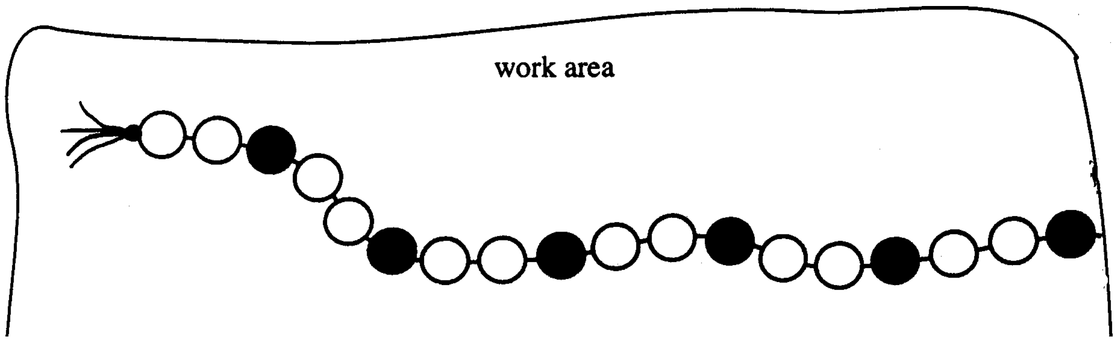

Figure 2 (Part of the) Problem included in the MiTC test for grade 5

Data collection

The MiTC test was administered in 17 classrooms in spring. The classrooms ranged from grade 3 through grade 5. Eight of the classrooms had a teacher from the control group and nine had a teacher from the experimental group. In total four schools were involved, situated in three school districts. The control classrooms were selected to match the experimental ones. In order to avoid large differences between the background of the students and the general school environment, the control teachers and the experimental teachers, in some cases, came from the same school or from the same school district. In both groups two teachers had a mixed-grade classroom. This means that in total 21 grade groups with a total of 17 teachers were involved in the data collection (see Table 1).

Table 1 Number of teachers, students, and grade groups involved in the data collection

	Control group		Experimental group		Total	
	Teachers	Students	Teachers	Students	Teachers	Students
Grade 3	3	61	6	75	9	136
Grade 4	4	72	3	43	7	115
Grade 5	3	55	2	36	5	91
	10 (8)	188	11 (9)	154	21 (17)	342

Together with the development of the test a double coding system was developed: one for the answers and one for the strategies. For the answers, the literal answer given by the student could be coded. The strategy coding was done by means of a two-digit code. The first digit referred to a general description of the strategy and the second digit specified a particular way of working within this category. It is important to explain that the categories and in particular the subcategories are not holistic but task-specific rubrics.

The coding of the students response was done blindly. The coder [1] did not know whether the students belonged to the control group or to the experimental group.

Results

A t-test analysis showed that the average percentage of correct answers in the addition problem (grade 3) did not differ significantly between the two groups (see Table 2). For the beads problem (grade 5) this is slightly different (see Table 3), but on the whole the control students and the experimental students did not diverge with respect to the total number of correct answers (see Table 4). The latter means that if this test would only have focused on the answers, the two groups might have been considered as equal in achievement.

Table 2 Answer and strategy results from the grade 3 problem (see Figure 1)

Grade 3 problem	Control group (n = 61)	Experimental group (n = 75)	
answer	48 % correct	60 % correct	n.s. (t-test)
strategy	5 % tinkering 15 % decomposing 62 % ciphering 18 % other	17 % tinkering 45 % decomposing 20 % ciphering 17 % other	p < .001 (Chi-square)

Table 3 Answer and strategy results from the grade 5 problem (see Figure 2)

Grade 5 problem	Control group (n = 55)	Experimental group (n = 36)	
answer	73 % correct	92 % correct	p < .05 (t-test)
strategy	42 % reasoning 9 % counting on/ multiplying on 4 % long division 7% wrong operation 2 % guessing 2 % other 16 % unclear strategy 18 % no strategy	75 % reasoning 14 % counting on/ multiplying on 0 % long division 3 % wrong operation 0 % guessing 0 % other 8 % unclear strategy 0 % no strategy	p < .05 (Chi-square)

However, the situation with the applied strategies is quite different. Here, a Chi-square analysis showed highly significant differences between the two groups (see also table 4).

In the addition problem (grade 3) the majority of the control group students used column arithmetic. An example of such a ciphering strategy is shown in Figure 3.

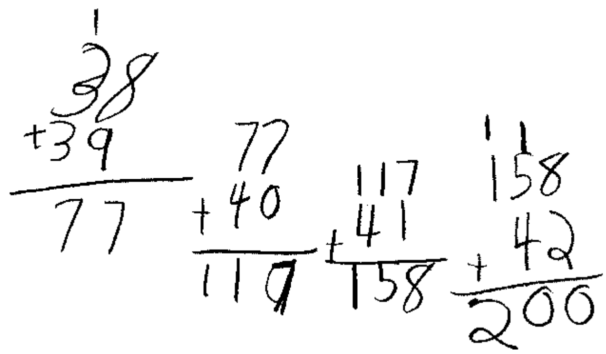

Figure 3 Example of ciphering strategy

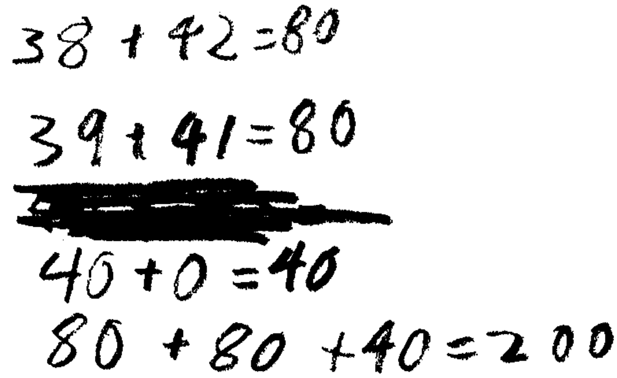

Figure 4 Example of tinkering strategy

The students from the experimental group, in contrast, brought their number and operation knowledge into play and applied a smart calculation strategy that is called “tinkering” (see Figure 4) or they used either a stringing strategy (e.g. 38 + 30; + 9; + 40; + 40; + 1; + 40; + 2) or a splitting strategy (e.g. 30 + 30 + 40 + 40 + 40 and 8 + 9 + 1 + 2). The stringing and splitting strategies are summarized by the term “decomposing”.

For the beads problem (grade 5) three quarter of the experimental students used a reasoning strategy based on knowledge of number relations. For instance, one student said:

> “I counted 10 and I know 10 go into 1000, so the last color for 10 which is white is the same color for 1000”.

Compared to the experimental group, the control group contained more students who, for instance, could not explain their strategy or students who did a long division (see Figure 5) together with the following explanation:

> “There are 2 white beads between each black bead that’s why I did 2 ÷ 1000. Then I did 1 ÷ 1000 which gave me my answer”.

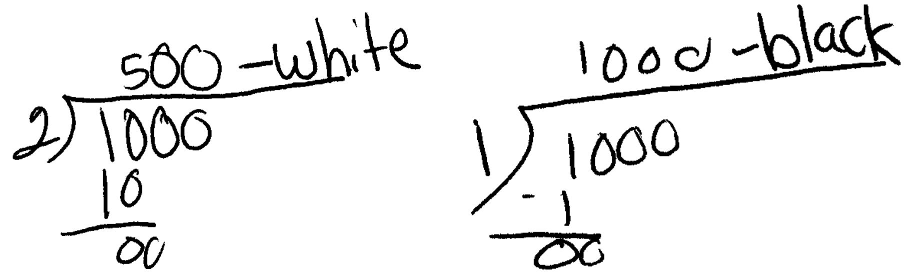

Figure 5 Example of student work that showed a long division

As is shown in Table 4 the difference in strategies was consistent in the study. Sensible tinkering strategies were applied more often by the experimental students (E+) and the ciphering strategies – in problems in which these strategies are not the most adequate solution strategies – were more often found in the work of control students (C+).

Table 4 Answer and strategy results in total for each grade

	Grade 3 11 problems	Grade 4 8 problems	Grade 5 11 problems
Difference in total number CORRECT ANSWERS (t-test)	E + ($p < .10$)	E = C	E = C
Difference in STRATEGIES (Chi-square)			
• tinkering	E + ($p < .001$)	E + ($p < .001$)	E + ($p < .001$)
• decomposing	E + ($p < .001$)	E + ($p < .10$)	E + ($p < .001$)
• ciphering	C + ($p < .001$)	C + ($p < .001$)	C + ($p < .001$)
• other	E = C	E = C	E = C

Concluding remarks

The data need further analysis (e.g. correlation between correct answer and strategy; influence of particular teachers) to conclude what the students gained from the MiTC project. The results so far give strong support to the idea that the mathematics achievements of the students in the experimental group are higher than those of the non-project students. Regarding assessment this study made clear that the measurement of mathematics achievements cannot be restricted to the answers only. This thinking fits with other developments in this area, like the work done by Suzuki (2000) who is developing a new assessment methodology in which the process of thinking in problem solving can be scored. Suzuki used the QUASAR general scoring rubric as a bedrock to develop a scale that can identify characteristics of achievement levels. In contrast with Suzuki's scale, in the MiTC study the categories were more mathematical and task-specific. The future will show what approach will be most helpful for understanding mathematics achievement.

Note

1. The coding was done by Chantal van Rooijen.

References

Baroody, A. J. (1987). Children's Mathematical Thinking. New York: Columbia Teacher College Press.

Cobb, P. (1996). The Mind or the Culture? A Coordination of Sociocultural and Cognitive Constructivism. In C. T. Fosnot (Ed.), *Constructivism: Theory, Perspectives, and Practice*. New York: Teachers College Press.

Freudenthal, H. (1968). Why to Teach Mathematics so as to Be Useful. *Educational Studies in Mathematics, 1*, 3-8.

Freudenthal, H. (1977). Antwoord door Prof. Dr. H. Freudenthal na het verlenen van het eredoctoraat [Answer by Prof. Dr. H. Freudenthal upon being granted an honorary doctorate]. Euclides, 52, 336-338.

Fosnot, C. T. and M. Dolk (2001). *Young Mathematicians at Work*. Portsmouth, NH: Heinemann Press.

Fosnot, C. T., M. Dolk, M. van den Heuvel-Panhuizen, B. Hilton, A. Wolf, and T. N. Bailey (under review). Evaluation of the Mathematics in the City Project: *Standardized Achievement Test results*.

Suzuki, K. (2000). How can the process of thinking be scored? Paper handed out at ICME-9 (WGA 10 Assessment). Tokyo, Japan.

Van den Heuvel-Panhuizen, M. (1996). *Assessment and Realistic Mathematics Education*. Utrecht: CD-ß Press / Freudenthal Institute.

Van den Heuvel-Panhuizen, M., Dolk, M., Fosnot, C.T., and Glick, J. (2000). *Macroscopic and Microscopic Lenses: A Symposium on Multi-Focus Assessment*. Symposium organized at ICME 9, Tokyo, Japan.