EXPLORING THE POTENTIALS OF HANDS-ON INVESTIGATIVE TASKS FOR CURRICULUM EVALUATIONS

Pauline Vos & Wilmad Kuiper
University of Twente, Faculty of Educational Science & Technology, Netherlands

A performance assessment for grade 8 students was administered in the Netherlands in 1995. The test consisted of practical, investigative tasks. Dutch curriculum experts considered this test to fit well with the Dutch mathematics curriculum, which is based on Realistic Mathematics Education (RME). However, Dutch students scored lower than expected on this practical test. This result demanded follow-up research. Therefore, the performance assessment was repeated in the Netherlands, in 2000. Trend data (1995-2000) show that the achievements of Dutch students show little gain. Also, the research gave new evidence on issues pertaining to reliability and comparability in practical tests.

INNOVATING ASSESSMENT IN MATHEMATICS

When measuring student achievement in mathematics for a large population, in many cases, paper-and pencil tests have been developed and issued to students. Especially, multiple-choice items have been popular because of automated scoring and their presumed high reliability. Yet, this method for measuring student achievement has come under debate, especially with respect to multiple-choice questions. These are associated with low-valued factual knowledge, asking for limited thinking processes. Assessment methods have been altered, although it proves difficult to break with traditions and find valid and reliable alternatives. The labels used for innovated assessment are for example: performance assessment, practical assessment, alternative assessment, and authentic assessment (see e.g. Burton, 1996; Clarke, 1996; Niss, 1993; Wiggins, 1989). The descriptions show considerable overlap and the terminology applies if some of the following criteria are met:

- testing through open questions and for higher order skills,
- being open to a range of methods or approaches,
- making students disclose their own understanding,
- allowing students to undertake practical work,
- asking for performances and products,
- being as an activity worthwhile for students' learning, and
- integrating real-life situations and several subjects.

To assess student achievement, the used formats can be portfolio, observation, interview, and so forth. But in large-scale testing, these formats can be too labour-intensive. Baxter & Shavelson (1994) compared the exchangeability of different

assessment methods. These were observation, notebook reports, computer simulation, short answer questions and multiple-choice questions. They found that observation yielded most detailed information on the achievement, with notebook reports providing a reasonable "surrogate". All other tests failed to approximate the same information.

An example of a large-scale study attempting to realise an innovation in assessment is the TIMSS Performance Assessment. In 1995, this test was optional within TIMSS (Third International Mathematics and Science Study), the international, comparative study. In TIMSS, a test is developed and translated into different languages and then issued to students of different nations, in order to compare their achievements internationally. Within TIMSS, in 1995, two different tests were available. Besides a standard written test, a practical test with hands-on items was developed. The practical investigative tasks of this Performance Assessment were considered to complement the written test with a higher focus on practical skills and a lower focus on knowledge reproduction. This practical test was developed from the educational vision that seeks coherence between procedural, declarational and conditional cognition. Students are expected to investigate systematically, contrary to cookbook-demonstrations. Being provided with manipulatives and instruments, they are tested through open tasks like: designing and executing an experiment, observing and describing their observations, using calculators, looking for regularities, finding notations and interpretations of their measurements, etc. The TIMSS Performance Assessment can be associated with Gal'perin's view of *learning by doing* in which mental acts (manipulating objects in the mind) develop from material acts (manipulating tangible objects) (Van Dormolen, 1993). Though, in the assessment, manipulatives and instruments are not seen as mere demonstrators of taught concepts. They are integrated into the assessment to trigger investigative activities (Harmon et al., 1997; Garden, 1999).

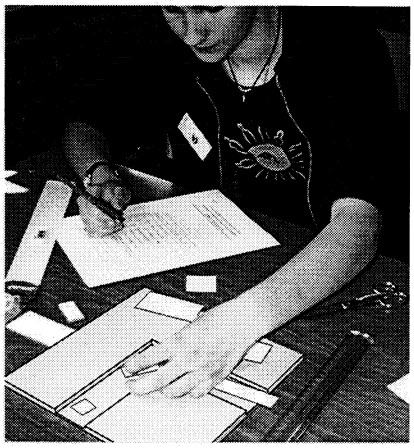

Figure 1: The task *Around the bend.*

TASKS OF THE TIMSS PERFORMANCE ASSESSMENT

The TIMSS Performance Assessment is administered in a circuit format in which students take turns in visiting stations. At each station they find a task, which guides them to carry out a small investigation. They write their answers on a worksheet. There are mathematics tasks, science tasks or combined tasks (overlapping between science and mathematics). There are five tasks with a clear mathematical focus:

- The task *Dice* is related to probability: students are given a die and a transformation rule for each throw (even: plus 2, odd: minus 1). They are asked to throw 30 times, record their findings and explain why one result (the "4") has a higher frequency.
- The task *Calculator* is related to number sense: students are given a simple calculator and are asked to discover a pattern in the multiplications of 34x34, 334x334 and 3334x3334. As the calculator holds only eight positions in the display, this is not an obvious task. The second part of the task consists of factorising 455 into two integers between 10 and 50.
- The task *Folding* is related to symmetry and spatial abilities: students have to make certain required figures by cutting, using a pair of scissors. Because only one cut is allowed for each figure, the paper has to be folded.
- The task *Around the bend* (see Figure 1) is related to scale drawing and finding rules: students are given a cardboard model of a corridor and have to cut rectangles (modelling furniture). By testing which rectangle fits through the corridor, they have to find a rule for the critical lengths.
- The task *Packaging* is related to measuring and the design of nets: students are given four table tennis balls and have to design different boxes for these.

Besides tasks with a mathematical focus, the test also contains tasks from biology, chemistry and physics. In these tasks, science investigations meet with mathematical activities, as students have to measure using instruments (using stopwatches, rulers, thermometers, and scales). But other mathematical activities are also required. For example, the task *Rubber Band* covers the topic of extrapolation. In this task, a number of washers are attached to a rubber band. Students have to measure the stretching of the band, related to the number of washers. With only ten washers given, students are asked to predict the length of the rubber band, if twelve washers were attached. Another task, named *Shadows*, is related to geometrical transformations. Students are given a torch, a card and a white screen. They have to project a shadow, which is twice as wide as the object, and find a rule for the distances between torch, card and screen. Finally, the task *Plasticine* asks for problem solving heuristics. Students are provided with a two-sided (uncalibrated) balance, two weights (20g and 50g) and a lump of plasticine. They are asked to make smart combinations in order to produce pieces of plasticine of 10g, 15g and 35g. Details of all tasks can be looked up in Harmon et al. (1997).

THE TIMSS PERFORMANCE ASSESSMENT IN THE NETHERLANDS

In 1995, the Netherlands participated in TIMSS for grade 8, both with the standard written test and the practical test. This was not without debate, as the mathematics curriculum in the Netherlands differs from the mathematics curriculum in many other countries. The Dutch mathematics curriculum is based on the principles of Realistic Mathematics Education (Freudenthal, 1975; De Lange, 1983). Dutch students learn mathematics starting from real-life contexts. Also in assessment, each test item has a

theme like transport, retail prices, or sports. Each mathematics test item has narrative texts explaining the context. The mathematical tasks are embedded into the theme.

Objections to participating in an international comparative study pertained to the validity of the comparison: if Dutch students had learnt mathematics through a different approach, they would not be able to show their particular competencies. From the RME-point of view, the written TIMSS test was too traditional and therefore, results based on it were not authoritative (Bos & Vos, 2000; Kuiper, Bos & Plomp, 1997). On the other hand, Dutch mathematics curriculum experts valued the additional test, the TIMSS Performance Assessment. It was considered to match well with the intentions of the Dutch curriculum. Instead of narrated contexts, students would now apply their mathematical skills in tangible contexts.

DUTCH STUDENTS' RESULTS IN TIMSS-1995

In 1995, 18 countries participated in both the written TIMSS test and in the TIMSS Performance Assessment for grade 8. Most countries reached a position in the international comparison on the practical Performance Assessment, which had a comparable ranking to the position in the standard written test. If a country ranked high on the league-table of one test, it would rank high on the other test. But the Netherlands was a marked exception here. Despite the fit of the TIMSS Performance Assessment with the Dutch intended mathematics curriculum, Dutch grade 8 students did not score as expected. Unlike on the written test, their achievement was at the level of the international average and not significantly above.

Therefore, an understanding of Dutch students' achievements was needed. Maybe, TIMSS in 1995 had come too early. The new RME-based curriculum was only introduced in 1993. At some schools the textbooks had not yet been replaced. And maybe teachers had not yet had time enough to adopt their instruction to the new curriculum. Thus, if the TIMSS Performance Assessment could be replicated at a later stage, trend data could establish whether the new curriculum was starting to settle. The repeat of the TIMSS Performance Assessment was planned for 2000. Unfortunately no other countries were interested in participating.

DESIGN OF THE STUDY AND RELIABILITY ISSUES

The TIMSS Performance Assessment was repeated by copying all international TIMSS protocols of 1995 (see Harmon et al., 1997; Garden, 1998). A two-stage stratified sample of 50 schools was drawn. The instruments of 1995 were re-used. The science tasks were maintained because of their mathematical aspects and to maintain the task-interaction effects during testing. For 2000, a sample of n=234 students at 27 schools was realised. This response of 54% is good according to Dutch standards. In 1995, with more funds being available, withdrawing schools had been replaced, resulting in n=437 students at 48 schools being tested.

Also for coding of students' answers, the 1995 procedures were followed. But this proved to be an ambiguous task. As already pointed out in the international 1995 TIMSS Performance Assessment report (Harmon et al., 1997), inter-scorer agreement can vary considerably. As a check on coding, two independent coders coded 10% of students' work. In the extreme case of one sub-item in the task *Shadows*, the Dutch coders only agreed in 52% of the cases on the correctness of students' work. As Zukovsky (1999) has pointed out, the coding of answers is conditional to the coders' background (e.g. coding experience, subject matter knowledge, teaching experience, etc). Some tasks showed such a low inter-scorer agreement that the results had to be doubted.

In another case, with the task *Rubber Band*, the protocol did not cover a strategy that was used by a considerable number of Dutch students. In this task the students had to measure and record the stretching of the rubber band with each washer. The result would yield an irregularly increasing graph (growing with 2-5 mm per washer) with slightly diminishing growth. But approximately 10% of Dutch students did not measure. Instead, they drew a graph that grew consistently with exactly 5 mm per washer. Their graph was a perfect straight line. There was no appropriate code for this styling strategy (a heuristics which reduces realism from the onset). Depending on the interpretation of the scheme, a coder could either give full or no credit.

Comparability of testing circumstances of some tasks in the Performance Assessment proved problematic. Although test instruments of 1995 were copied in 2000, there were minimal mutations in the laboratory equipment. These mutations were within the narrow range that the international protocol allowed. One example will illustrate the effect. In the task *Shadows* a torch is used. The torch used in 1995 gave a vaguer shadow, while the torch of 2000 gave a sharper edge to the shadow. The latter made student's measurements easier giving them more time for remaining items in the task.

To eliminate unreliable and incomparable results, two tests were carried out. First, for each task, Cronbach's alpha was calculated for 1995 and 2000 separately. Results higher than 0.6 were considered acceptable. Second, a chi-squared test was carried out, revealing that answer patterns differed because of altered testing circumstances. As a result, four science tasks (*Magnets*, *Rubber band*, *Shadows* and *Plasticine*) had to be eliminated from analysis. Fortunately, the five mathematics tasks passed these tests.

Based on these five mathematics tasks, the initial Dutch agitation about the disappointing test results dwindled. In the international reports, the results of the task *Plasticine* had been included into the Mathematics league table. This task had been pulling down the overall results of Dutch students. Omission of the unreliable results raised the Dutch position above the international average. Table 1 shows the results in TIMSS 1995 of the 19 countries that participated both in the standard written test and in the TIMSS performance Assessment. The first column shows the mathematics results of the TIMSS written test. The second column shows the results on the

TIMSS Performance Assessment as presented in the international TIMSS report. This average score is based on five mathematics tasks *Dice*, *Calculator*, *Folding*, *Around the bend* and *Packaging* plus the combined task *Plasticine* (Harmon et al., 1997). The third column shows the results based on the five mathematics tasks only. The position of the Netherlands in each league-table is indicated in grey. As can be seen in the two columns for the Performance Assessment, the average scores of each country do not change much after deletion of the *Plasticine* task. The difference between the average scores of a country is at most 2% (with the exception of the Netherlands: +3, and Iran: -6). In fact, the correlation of the two scores at country level is r=0.97 (n=19). Yet, the position in the league-table can vary considerably, because of the large number of countries with only slight differences in their scores. The positions of countries with ranking 2 up to 13 are based on scores that are very close. Therefore, it can be concluded, that the test in itself is robust (the average scores do not change much after deleting of a task). Yet, the presentation in a league-table is misleading because differences between almost equal scores are enlarged as average scores (on a continuous scale) are being transformed into a ranking (on a discrete scale).

Table 1: Ranking of 19 countries in 1995 on two TIMSS mathematics tests.

TIMSS Written Test		Performance Assessment (six mathematics tasks)		Performance Assessment (five mathematics tasks)	
Country	Score points	Country	Avg % correct	Country	Avg % correct
1 Singapore	643	1 Singapore	70	1 Singapore	70
2 Czech Rep	564	2 Switzerland	66	2 Romania	67
3 Switzerland	545	3 Australia	66	3 England	65
4 Netherlands	541	4 Romania	66	4 Netherlands	65
5 Slovenia	541	5 Sweden	65	5 Norway	65
6 Australia	530	6 Norway	65	6 Australia	65
7 Canada	527	7 England	64	7 Switzerland	64
8 Sweden	519	8 Slovenia	64	8 Slovenia	64
Intl average	*509*	9 Czech Rep	62	9 Sweden	63
9 Nw Zealand	508	10 Canada	62	10 Nw Zealand	61
10 England	506	11 Nw Zealand	62	11 Canada	61
11 Norway	503	12 Netherlands	62	12 Scotland	61
12 USA	502	13 Scotland	61	13 Czech Rep	60
13 Scotland	498	*Intl average*	*59*	*Intl average*	*59*
14 Spain	487	14 Iran	54	14 USA	54
15 Romania	482	15 USA	54	15 Spain	54
16 Cyprus	474	16 Spain	52	16 Portugal	49
17 Portugal	454	17 Portugal	48	17 Iran	48
18 Iran	428	18 Cyprus	44	18 Cyprus	42
19 Colombia	385	19 Colombia	37	19 Colombia	35

RESEARCH RESULTS FOR THE REPEAT STUDY

By repeating the study, a trend in the achievement of Dutch students could be established. The resulting scores are given in Table 2. For each task the average percentage of correct scores on the items was calculated.

Table 2: Mathematics tasks from the TIMSS Performance Assessment 1995-2000, average percentage correct score of Dutch students.

Task	1995 (n=437)	2000 (n=234)
Dice	**77**	**74**
Calculator	**62**	**59**
Folding	**73**	**77**
Around the bend	**68**	**69**
Packaging	**52**	**54**
Average score on mathematics tasks	***66***	***67***

Compared to 1995, the scores do not show any significant changes on the five mathematics tasks. The average score correct of 66 on these tasks in 1995 does not differ significantly from the average score of 67 in 2000. Also, on each task separately, the shifts were statistically insignificant.

These results show that Dutch students have not gained in practical, investigational skills. This could be caused by the classroom practice, in which they never encounter hands-on tasks like the ones in the TIMSS Performance Assessment. Students told us, during testing, that they had never done this before. Still, the tasks match well with the intended Dutch curriculum. But the testing practice has stuck with a paper-and-pencil format in which students have to read the texts in which real-life contexts are described. Tangible real-life contexts with manipulatives are habitually not utilised in Dutch testing. However, as an additional result, the TIMSS Performance Assessment proved to be an eye-opener to many Dutch mathematics teachers. During the testing sessions, they observed the tasks and how their students coped with these. Some teachers admitted that they had never thought mathematics could be tested in this way, through a mathematical practical test. As such, the TIMSS Performance Assessment could prove to be part of the exemplary material that is needed to support curriculum reform (Fullan, 1991).

CONCLUSION

The TIMSS Performance Assessment clearly has potentials in monitoring students' mathematical investigation skills. It is a valid addition to standard, paper-and-pencil tests. It also shows, that manipulatives are useful to organise time-restricted hands-on tasks in mathematics, linking mathematics to other areas. Still, much more experience is needed when it comes to reliability and comparability issues.

REFERENCES

Baxter, G.P. & Shavelson, J.R. (1994). Science performance assessments: benchmarks and surrogates. *International Journal of Educational Research, 21(3),* 279-298.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Kelly, D.L. & Smith T.A. (1996). *Mathematics achievement in the middle school years. IEA's Third International Mathematics and Science Study.* Boston, USA: Boston College.

Bos, K.Tj, & Vos, F.P. (2000). *Nederland in TIMSS-1999, exacte vakken in leerjaar 2 van het voortgezet onderwijs* [The Netherlands in TIMSS-1999, mathematics and science in grade 8]. Enschede: Twente University.

Burton, L. (1996). Assessment of mathematics: what is the agenda? In M. Birenbaum & F.J.R.C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes, and prior knowledge (pp 31-62).* Dordrecht: Kluwer.

Clarke, D. (1996). Assessment. In A.J. Bishop, et al. (Eds), *International Handbook of Mathematics Education (pp 327-370).* Dordrecht: Kluwer.

Dormolen, J. van (1993). *Wiskunde werklokaal- het gebruik van materialen en instrumenten bij het leren van wiskunde* [Mathematics laboratory – the use of materials and equipment for learning mathematics]. Utrecht: APS.

Fullan, M.G. (1991). *The New meaning of educational change* (2nd ed.). New York: TCP.

Freudenthal, H (1973). *Mathematics as an Educational task.* Dordrecht: Reidel.

Garden, R.A.. (1999). Development of TIMSS Performance Assessment Tasks. *Studies in Educational Evaluation*, 25 (1999), 217-241.

Harmon, M., Smith, T.A., Martin, M.O., Kelly, D.L., Beaton, A.E., Mullis, I.V.S., Gonzales, E.J. & Orpwood, G. (1997). *Performance assessment in IEA's Third International Mathematics and Science Study.* Boston: Boston College.

Kuiper, W.A.J.M., Bos, K.Tj. & Plomp, Tj. (1997). *Wiskunde en de natuurweten-schappelijke vakken in leerjaar 1 en 2 van het voortgezet onderwijs. Nederlands aandeel in TIMSS populatie 2* [Mathematics and the science domains in secondary 1 and 2. Dutch participation in TIMSS population 2]. Enschede: Twente University.

Lange, J. de (1983). *Mathematics, Insight and Meaning*. Utrecht: IOWO.

Niss, M. (Ed.) (1993). *Investigations into Assessment in Mathematics Education, an ICMI Study.* Dordrecht: Kluwer.

Wiggins, G. (1989). A true test: towards more authentic and equitable assessment. *Phi Delta Kappan, 76(9),* 703-713.

Zuzovsky, R. (1999). Problematic aspects of the scoring of the TIMSS practical performance assessment: some examples. *Studies in Educational Evaluation 25 (1999),* 315-323.