教育研究學報 1992，第七卷， 33-40
Educational Research Journal 1992, Vol.7, pp. 33-40

The Assessment of Practical Ability in Science Using the Partial Credit Model

Malcolm J ROSIER
Survery Design and Analysis Services Pty Ltd.

A science practical test was given to a sample of 220 students in 74 schools in Years 5 and 9 in Victoria (Australia). There were three problems, and four tasks per problem. The tasks were designed to minimise the supervision at the classroom level. A total test score was calculated using the partial credit version of the item response theory model. Each category of each of the 12 tasks was assigned a score (in logits) on the underlying scale that measured practical ability. The score for each student was measured as the location of the student on the scale. In order to simplify the description, these scores were then divided into five levels of science practical ability.

一科學實驗的測驗實施於 220 個學生中，他們來自澳洲維多利亞省 74 間學校的五年級及九年級。這測驗由三條大題目組成，每條大題目又可分為四條小題目。而每條小題目均要求學生做一項實際的工作(在每項工作中，學生可獲 0， 1，或 2 分，而全卷的滿分為 24 分)。這些題目的設計均盡量減少教師在場監督的需要。

本測驗的計分應用「測驗題回應理論模型」中的「部分績分模型」來處理。該十二條小題目的每一個分數均可在科學實驗能力績分表上取得一數值。而每個學生的成績亦能在該績分表上取得一個位置。在報告學生成績的時候，為了簡便起見，該等績分共分為五個等級，而只報告學生所屬的等級。

All modern science curricula recognise the importance of practical work, and generally try to integrate it into the curriculum package. As noted by Hofstein (1988, p. 189):

> 'The laboratory has long been given a central and distinctive role in science education. It has been used to involve students with concrete experiences and with concepts and objects.'

However, the assement of practical work has received less attention than the assessment of science content or attitudes (Doran, 1980; Tamir, 1990).

Attempts have been made to measure practical skills by paper-and-pencil tests. In the First International Science Study (FISS) there were some items which displayed pictures of scientific apparactus and procedures about which the students were asked questions (Comber & Keeves, 1972). It was assumed that students would only be able to answer correctly if they had handled such apparatus. The scores on these paper-and-pencil 'practical' items were moderately correlated with the science achievement test scores, with a median value across all the countries of 0.58. The two types of tests were related but were probably measuring different aspects of students' understanding of science. This finding was validated by more extensive analyses of the FISS data by Kojima (1974). It was concluded that the assessment of practical skills required actual hands-on testing instead of paper-and-pencil tests.

This finding was behind the decision to include the assessment of practical skills in the Victorian Science Achievement Study (Adams, Doig & Rosier, 1991). The Australian Council for Educational Research was commissioned by the Victorian Ministry of Education to conduct this study in 1990. The main focus of the study was on the assessment of science content and understanding. However, it was obvious that efforts should also be made to see how the curriculum emphasis on process skills was reflected in students' ability to handle practical tasks. Although the study was set in Australia, it has relevance to Hong Kong and other countries where the role of practical work is given a high priority.

There are two main problems in the assessment of science practical skills or other similar performance outcomes. The first problem is organisational. In whatever form the practical work is assessed, it necessarily involves the setting up of equipment. This problem is exacerbated by the desirability of devising novel procedures that the students have not seen before, which may require the purchase or construction of special equipment. Practical testing also requires more supervision than for pencil-and-paper tests.

An associated problem is the scoring. Content tests generally have only one correct answer. At the extreme, the correct answers can be included among a set of distractors in a multiple-choice format, and the scoring can be done very quickly. Indeed, with optical mark reading (OMR) facilities, the computer can do all the work and the teacher may never handle the tests. The assessment of performance, as in science practical work, should allow for variety in the responses, and scoring procedures should take account of different levels of performance.

This paper describes the Science Practical Test devised for the Victorian Science Achievement Study, and the procedures adopted to deal with problems of organisation and scoring. In particular, it describes the use of the partial credit model for estimating a total Science Practical Test score which enables more realistic scoring procedures to be employed.

Method

Classification System

Before commencing work on the practical test, it was necessary to define a classification system. Several authors have proposed classification systems for the assessment of practical work. The system adopted for this study (shown in Table 1) was based on categories defined by Klopfer (1971), incorporating some elements of the work of the United Kingdom Assessment of Performance Unit (Driver et al., 1980).

TABLE 1

A Classification System for Science Practical Tests[a]

	Classification	Cross-reference
1	Knowledge	Klopfer A1-A9
2	Following instructions	APU 2 gamma
3	Manual skills	Klopfer G1-G2
4	Observation/description	Klopfer B1-B2, APU 3 beta
5	Measurement	Klopfer B3-B5, APU 2 alpha
6	Estimation	APU 2 beta
7	Problem solving	Klopfer C1-C4
8	Interpretation of data	Klopfer D1-D5, APU 3 gamma
9	Application	Klopfer F1-F3
10	Attitudes	Klopfer H1-H6

a The table refers to the schemes defined by Klopfer (1971, pp. 591-626) and the Assessment of Performance Unit (Driver et al., 1980, pp. 180-185).

Category 1 refers to knowledge, which is strictly not part of a classification of practical tasks, but which provides a necessary foundation for higher order problem solving and application skills. Categories 2-3 focus on basic skills needed for tackling the practical tasks defined by subsequent categories. Categories 3-5 deal with skills needed for gathering and processing information. Categories 7-9 cover the range of skills involved in using the information to solve problems or to generate hypotheses. Category 10, while not directly linked to practical tasks, refers to attitudes about practical tasks and problem solving.

Sample and Organisation

Problems with organisation were initially dealt with by selecting only a small subsample of the students taking part in the main study. Three students were selected at random from each school to take part in the practical testing program lasting 60 minutes. The removal of three students from the class used for the sample in each school had a minimal disruptive impact on the normal work of the class. Results were obtained from 123 students from 41 schools at the Year 5 level and from 97 students from 33 schools at the Year 9 level. The subsample for the practical test was similar to the sample for the main study on characteristics such as age, sex, home background and science achievement.

The Science Practical Test consisted of three problems, for which the equipment was set up at three separate locations in a science room or other room with access to a sink and water. Each of the three students spent about 20 minutes on each problem. All the equipment needed for the practical tasks

was supplied by the Australian Council for Educational Research as part of the study, and schools were invited to retain this equipment at the end of the testing session.

Practical tests are labour intensive, and those used in previous studies generally required administration and assessment by persons trained for the task. This approach was not feasible for the study on financial and organisational grounds. The practical testing was therefore designed to minimise the involvement of school staff. Their role was limited to setting up the equipment and exercising general supervision of the testing program, which could in fact be done by a laboratory assistant or parent aide. The tests were also designed so that the scoring could be done at the Australian Council for Educational Research.

Scoring

The range of skills covered by practical testing cannot readily be scored as simply right or wrong, especially where students write their answers instead of selecting from multiple choice alternatives. In all performance measures there are gradations. The separate tasks within each problem were therefore scored on a 2, 1, 0 basis, where:

2 represented full competence or understanding;
1 represented partial competence or understanding; and
0 represented a scientifically meaningless or unclear response.

Tests and Results

The following section describes the major tasks in the practical test. For each of the three problems, the content and results for four tasks are shown.

TABLE 2
Scoring and Results: Worn Out Batteries

		Year 5	Year 9
B1	Identification of good and worn out batteries		
2	Correct. Identifies good batteries as A and C and worn out batteries as B and D	56%	62%
1	Partly correct. Places one battery in the wrong category.	9%	10%
0	Wrong answer. Places more than one battery in the wrong category.	34%	25%
	Missing data.	1%	3%
B2	Explanation of procedure used		
2	Systematic statement of the student's approach to solving the problem, indicating that the batteries were placed in the torch in sequence and the results compared.	11%	34%
1	Less systematic or vague statement.	38%	37%
0	Statement unscientific or unable to be interpreted.	41%	23%
	Missing data.	10%	6%
B3	Time taken to solve the problem		
2	1-4 minutes.	41%	46%
1	5-7 minutes.	21%	32%
0	8 or more minutes.	23%	19%
	Missing data.	15%	3%
B4	How does the torch work?		
2	Correct description: complete circuit through the batteries and globe.	6%	12%
1	Partially correct, with no indication of the need for a circuit: energy/power goes to the globe/torch/light.	28%	40%
0	Incorrect or guesswork.	57%	39%
	Missing data.	9%	9%

Problem 1: Worn Out Batteries

The focus for Problem 1 was on problem solving. The problem was defined for the students as follows:

You are given a torch and four batteries. Some of the batteries may be worn out. Your task is to work out which batteries are OK.

The batteries were labelled Battery A, Battery B, Battery C and Battery D. Batteries A and C were good batteries, and Batteries B and D were worn out. The questions for the four tasks were:

B1 Which batteries are OK, and which are worn out?
B2 How did you decide which were batteries were worn out?
B3 How long did you take to solve this problem?
B4 How does the torch work? Why does the light shine when the torch is switched on?

The results for these four tasks are shown in Table 2. For each task, there is a description of the kind of answer needed to score 2, 1 or 0, together with the percentage of Year 5 and Year 9 students scored at each category of each item. The percentage of students who did not give any response to the item is also given since there was no basis for allocating missing data across the valid responses.

At both the Year 5 and Year 9 levels, students tended to give good response to questions B1 and B3. However, more of the Year 9 students were able to supply good explanations for their activities (questions B2 and B4).

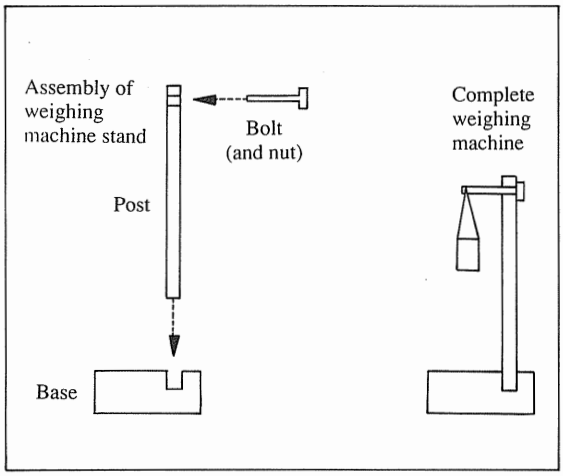
Problem 2: Weighing Machine

The main focus of Problem 2 was on measurement, and the interpretation of the data thus obtained. Students were given the following task:

In this problem you will construct a simple machine for weighing. The 'weighing machine' consists of two parts: a stand and a plastic container for holding the things to be weighed. Your task is to use the machine to weigh some coins.

The following instructions were given for the construction of the weighing machine:

The stand for the weighing machine has three parts: a base, a post and a bolt. Take these parts for the stand out of their plastic bag and put them together as shown in the diagram.



In order to finish making the weighing machine, the plastic container is hung from the bolt by means of a loop of elastic. Attach the loop of elastic to the plastic container. Take the lid off the small plastic container. Hold one strand of the loop of elastic across the top of the container. Hold the elastic so that it is firm but not stretched. Place the lid back on the plastic container so that it traps the loop of elastic. Hang the loop of elastic with the container from the bolt on the stand.

For question W1, the students were asked to make a series of measurements, initially measuring the distance from the bolt to the top of the empty container, and then measuring the distance when the container held one, two, three and four coins.

The students were then asked or instructed:

W2 How much of the stretching of the elastic is due to the weight of each coin?
W3 Prepare a graph of the distance from the bolt to the top of the plastic container (on the vertical axis) and the number of coins (on the horizontal axis).
W4 Use the graph to estimate the distance from the bolt to the top of the plastic container when there are *five* coins in the container.

The results shown in Table 3 indicate that the Year 5 students tended to have much more difficulty with these four tasks, in making the basic measurements, in drawing the graph, and in estimating the results for five coins based on the measurements for the first four coins.

TABLE 3
Scoring and Results: Weighing Machine

		Year 5	Year 9
W1	Distance from bolt to container		
2	Measurements appear to be correct, and given to one decimal point.	19%	66%
1	Measurements are suspect or irregular; given as fractions	68%	28%
0	Measurements appear to be wrong; e.g. measured from the bottom of the case.	10%	6%
	Missing data.	3%	0%
W2	Distance of stretching due to coins		
2	Values correct in terms of responses to Item W1, and given to one decimal point.	17%	40%
1	Values suspect and/or given as fractions.	12%	23%
0	Values wrong or unrelated to earlier responses.	49%	31%
	Missing data.	22%	6%
W3	Graph		
2	Good graph – points or tops of bars clearly marked corresponding to data.	16%	50%
1	Poor graph – e.g. poorly specified bar graph, measurements missing, measurements in wrong columns.	14%	20%
0	Meaningless graph unrelated to data.	5%	9%
	Missing data.	65%	21%
W4	Estimate for Coin 5		
2	Reasonable estimate based on use of data from the graph.	15%	35%
1	Reasonable estimate based other than on graph data.	16%	28%
0	Meaningless estimate unrelated to data.	5%	4%
	Missing data.	64%	33%

Problem 3: Sorting Seeds

Problem 3, dealing mainly with classification skills, was expressed as follows:

In this problem you are supplied with several kinds of seeds. Your task is to describe and test them in several ways.

The six kinds of seeds were: buckwheat, coffee beans, peanuts, poppy seeds, rice and sunflower. The tasks were:

S1 Sort the seeds by size.
S2 Sort the seeds by oiliness.

Take a piece of the paper provided (actually thin typing paper 40 gsm). Fold it in half. Place some of the seeds you are testing inside the folded paper. Crush the paper (with the seeds inside) between two coins. The oily seeds leave a translucent mark on the paper. Hold the paper up to a window. More light comes through the oily mark than through the other part of the paper.

S3 State which of the following substances contain oil: cooking oil, margarine, flour, tea leaves, peanut paste.
S4 Sort the seeds into two groups or families in terms of shape or color or hardness or some other reason. Explain why you placed the seeds in each group.

Table 4 sets out the scoring description and results for some of the tasks. These tasks were easier than for the other two problems. However, few of the Year 5 students could suggest a classification for the seeds even though this kind of process skill features in many primary school curricula.

Science Practical Test Score

The three separate problems within the Science Practical test cover different aspects of practical work. Taken as a whole, the three problems and their 12 constituent tasks covered most of the categories set out in the classification system. The aim of the

TABLE 4
Scoring and Results: Sorting Seeds

		Year 5	Year 9
S1	Sorting seeds by size		
2	Correct order: peanuts, coffee beans, sunflower, rice, buckwheat, poppy seed.	68%	69%
1	Partly correct (one out of order) or reverse order or one omitted.	20%	24%
0	Mostly incorrect (more than one out of order).	4%	3%
	Missing data.	8%	4%
S2	Sorting seeds by oiliness		
2	Correct		
	Oily: peanut, poppy seed, sunflower.		
	Not oily: buckwheat, coffee beans, rice.	37%	52%
1	Partly correct (one misplaced).	39%	34%
0	Mostly incorrect (more than one misplaced)	19%	14%
	Missing data.	5%	0%
S3	Identification of oily substances		
2	Correct		
	Oily: cooking oil, margarine, peanut paste.		
	Not oily: flour, tea.	45%	80%
1	Partly correct (one misplaced).	40%	18%
0	Mostly incorrect.	5%	1%
	Missing data.	10%	1%
S4	Classification of seeds		
2	Logical/valid classification system, correctly classificed within this system.	27%	64%
1	Logical/valid classification system, but incorrectly classified within this system.	20%	11%
0	Not a logical/valid classification system. No obvious basis for the classification.	10%	12%
	Missing data.	43%	13%

final stage of the scoring procedure was to produce a total test score to measure science practical ability.

One possible method was to sum the scores (2, 1 or 0) on the separate tasks, with missing responses for a task being assigned a value of 0. In scoring the Science Practical Test for the Victorian Science Achievement Study it was possible to use more advanced techniques based on item response theory procedures to define a linear scale calibrated in logits. In the item response theory, the ability level for each person is not derived by adding up test scores, but by assigning each student to a location on the scale.

In the basic model, the responses to each item are scored 1 or 0 (right or wrong). In the partial credit model (Masters, 1990), there can be several responses for each item, ordered in terms of levels of performance. A value (in logits) is obtained for each of the categories for each item, mapped on to the scale for the trait being measured.

The Science Practical Test was identical for both the Year 5 and Year 9 students, and was scored the same for both groups. The difficulty levels (logit values) for the items and the ability levels for the students (also measured in logits) were estimated by taking the Year 5 and Year 9 students as a single dataset. This meant that direct comparisons could be made across the two year levels[2].

Table 5 presents the threshold difficulty values (in logits) for each category of each of the 12 items, arranged in five levels. The highest item difficulty value was for category 2 of Item 4 in the Worn Out Batteries problem (shown as B4.2), requiring a good explanation of how a torch worked. Category 1 of Item 4, indicating a poorer response to the question, is at Level 3. In contrast, the lowest item difficulty was for S3.1 (at Level 1), indicating a partly correct identification of common oily substances.

TABLE 5
Threshold Difficulty Values (Logits) for Each Category of Each Item

Level 5		Level 4		Level 3		Level 2		Level 1	
Item	Logit	Item	Logit	Item	Logit	Item	Logit	Item	Logit
B4.2	2.42	W1.2	0.83	B4.1	0.63	B1.2	0.24	S1.2	-0.41
B2.2	1.60	W4.2	0.80	B3.2	0.61	S4.2	0.24	B3.1	-0.45
W2.2	1.12	S2.2	0.75	W3.2	0.50	B2.1	0.00	S4.1	-0.70
				W2.1	0.45	B1.1	-0.08	S2.1	-0.74
						S3.2	-0.08	W3.1	-0.76
								W4.1	-1.21
								W1.1	-1.80
								S1.1	-1.88
								S3.1	-2.11

In order to simplify the presentation of results, the student ability levels (in logit values) were converted to a 5-point scale, with each level containing about 20 per cent of the total sample of students. Table 6 summarises the location of each of the responses on the scale of practical ability. The highest level indicates that the students demonstrate good practical skills, are able to explain practical procedures and observations, and have a good scientific understanding of how things work. The lowest level indicates that the students gave poor responses to most of the tasks.

Table 6 also shows the distribution across levels for Year 5 and Year 9 students. There was a group of about 29 per cent of the Year 5 students (Level 1) who were barely coping with the tasks in the practical test. On the other hand, the comparison across year levels shows that some of the better Year 5 student are scoring at higher levels than the weaker Year 9 students.

Discussion

Several implications for teaching follow from these results. Continual efforts must be made to enhance the opportunities of students to develop practical work skills, especially in terms of giving clear explanations for the activities carried out. Teachers at the lower secondary level should be aware of the wide range of practical skills possessed by the primary students entering secondary school. Obviously some students, especially at the primary level, require support in developing practical skills to raise them from a level where they can barely cope with even the simplest instructions or tasks.

TABLE 6
Description of Science Practical Scale Levels with Percentage Frequency Distribution by Year Levels

Level	Range (Logits)	Description	Year 5 (n = 123)	Year 6 (n = 97)
(high)				
5	more than 1.08	Obtains good scores on the tasks and gives good explanations of their observations.	7%	37%
4	0.69 to 1.08	Obtains good scores on most of the tasks.	19%	23%
3	0.34 to 0.69	Obtains good scores on some of the tasks.	19%	19%
2	-0.15 to 0.4	Obtains good scores on a few of the tasks.	26%	13%
1	less than -0.15	Obtains poor scores on most of the tasks.	29%	8%
(low)				

The scores on the Science Practical Test were linked to the students' scores on the multiple-choice science achievement test. The correlation was 0.53 ($p<0.001$) at both Year 5 and Year 9. These values are consistent with the value of 0.58 for the First International Science Study, cited at the start of this article. The current study has demonstrated procedures for the organisation and scoring of practical work. The thorough assessment of science practical work is important and feasible, and should be attempted in order to assess important aspects of the learning of science that cannot be properly assessed by paper-and-pencil tests.

Notes

[1] This paper deals with a short version of the test. Adams, Doig & Rosier (1991) discusses the complete version.

[2] The partial credit analysis was carried out using Quest, a comprehensive item analysis program that calculates traditional and item response theory item and test statistics.

References

Adams, R.J., Doig, B., & Rosier, M.J. (1991). *The Victorian Science Achievement Study 1990.* Hawthorn: Australian Council for Educational Research.

Comber, L.C., & Keeves, J.P. (1973). *Science achievement in nineteen countries.* New York: John Wiley.

Doran, R.L. (1980). *Measurement and evaluation of science instruction.* Washington, DC: National Science Teachers Association.

Driver, R., Gott, R., Johnson, S., Worsley, C., & Wylie, F. (1982). *Science in schools. Age 15: Report No. 1. Assessment of Performance Unit.* London: Her Majesty's Stationery Office.

Hofstein, A. (1989). Practical work and science education II. In Fensham, P. (Ed.), *Developments and dilemmas in science education* (pp. 189-217). London: Falmer Press.

Kojima, S. (1974). IEA Science Study in Japan with special reference to the Practical Test. *Comparative Education Review, 18*(2), 262-267.

Klopfer, L.E. (1971). The evaluation of learning in science. In Bloom, B.S., Hastings, J.T., & Madaus, G.F., *Handbook on formative and summative evaluation of student learning* (pp. 559-641). New York: McGraw Hill.

Masters, G.N. (1990). Partial credit model. In Walberg, H.J., & Haertel, G.D. (Eds.), *The international encyclopedia of educational evaluation* (pp. 388-393). Oxford: Pergamon Press.

Tamir, P. (1990). Practical examinations. In Walberg, H.J., & Haertel, G.D. (Eds.), *The international encyclopedia of educational evaluation* (pp. 476-481). Oxford: Pergamon Press.

Author

Malcolm J ROSIER, Survey Design and Analysis Services Pty. Ltd., 249 Eramosa Rd West, Moorooduc, Victoria 3933, Australia.