Assessing the Effectiveness of Pencil-and-Paper Tests for School Mathematics

M. A. (Ken) Clements and *Nerida F. Ellerton*

The University of Newcastle Edith Cowan University

An extended Newman interview technique was used to gain additional information on responses to 16 pencil-and-paper questions (8 short-answer, and 8 multiple-choice) by 65 students in three Year 8 classes in three NSW regional high schools. The data suggest that about one-quarter of students' responses could be classified as either: (a) correct answers given by students who did not have a sound understanding of the mathematical knowledge, skills, concepts and principles which the questions were intended to "cover"; or (b) incorrect answers given by students who had partial or full understanding.

Introduction

Pencil-and-paper tests are widely used to assess mathematical achievement. For example, in Australia, state-wide testing programs use pencil-and-paper tests; in the United Kingdom the national testing of students at different year levels is through pencil-and-paper tests; the largest mathematics competition in the world, the Australian Mathematics Competition, is based on results from pencil-and-paper tests, and the International Association for the Evaluation of Educational Achievement (IEA), which is supported by funds provided by governments around the world, has used pencil-and-paper tests in its three major international mathematics achievement studies.

It is not widely known, however, that recent research has generated data which suggest that students who give correct answers to pencil-and-paper mathematics items sometimes have little or no understanding of the mathematical concepts and relationships which the tests were designed to measure (Frary, 1985; Gays & Thomas, 1993; Hembree, 1987; Thongtawat, 1992).

Despite this questioning of the effectiveness of pencil-and-paper tests, education authorities continue to believe that so-called "valid" and "reliable" pencil-and-paper tests can satisfactory measure student understanding of mathematical knowledge, concepts, skills, and principles.

Aim

The aim of the research was to analyse the responses of 65 Year 8 students to 16 pencil-and-paper mathematics questions. For each student's response to a particular question, a decision needed to be made with respect to the following:

1 Did the student give a correct answer, an incorrect answer, or no answer to the item?

2 So far as the concepts and relationships involved in the question were concerned, did the student have (a) no understanding, (b) some understanding, or (c) a sound understanding?

A third variable was associated with the form of the question—was the question of the multiple choice or of the short answer variety? No other form of question was used.

Method

For this study the authors developed and used an extended form of the Newman error analysis technique (Clements, 1980; Ellerton & Clarkson, 1992; Newman, 1983) to investigate *both the errors and correct answers* given by students to items on pencil-and-paper mathematics tests.

The aim of each interview was to ascertain the level of understanding associated with each response by each student to questions on two 16-question pencil-and-paper instruments especially developed for the study. For each response the level of understanding was assessed according to a 3-point scale: "0" was allocated if a student did not recognise, or had no grasp of the necessary concepts; "1" was allocated if a student recognised which concepts might apply but had only a limited understanding of the necessary concepts; and "2" was allocated if the concepts and relationships were recognised and were well understood. Further details relating to the criteria for, and method of, locating these scores will be discussed later in this paper.

Development of the Study Instruments

A large number of multiple-choice pencil-and-paper questions used in reputable research studies (for example, the questions used in the Second IEA study—see Rosier, 1980) were shown to 10 experienced teachers in a regional city in New South Wales, and each teacher was asked to choose the 30 items most relevant to the NSW Year 8 mathematics syllabus. On the basis of responses given by these teachers, 16 multiple-choice items suitable for Year 8 students in NSW were chosen by the authors (in consultation with three experienced research assistants) for use in the study. These 16 items comprised the trial form of the multiple-choice instrument for the study.

The authors (once again with the help of the three research assistants) then constructed 16 short-answer (non multiple-choice) questions, each question being parallel to one of the 16 multiple-choice items which had been selected. Numbers and/or operations (or other components of the multiple-choice items) were changed so that parallel questions did not involve exactly the same calculations. These 16 items comprised the trial form of the short-answer instrument for the study.

For example, Question 8 for the short-answer instrument asked students to state the area (in square metres) of the yard in Figure 1(a). The corresponding multiple-choice question asked students to find the area (in square centimetres) of a brass plate whose dimensions were as shown in Figure 1(b). Students were instructed to select their response from the following 5 options: 16 cm^2; 24 cm^2; 32 cm^2; 64 cm^2; and 96 cm^2.

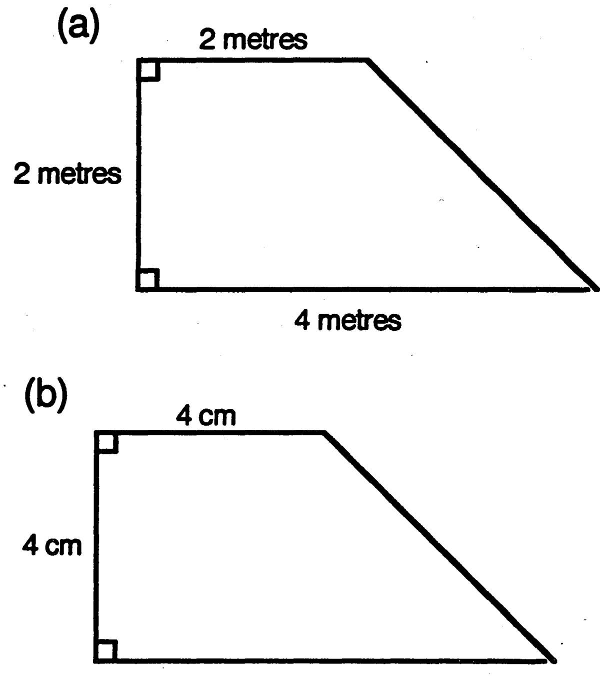

Figure 1. Diagrams for parallel questions on the area of a trapezium.

Both trial instruments were taken by students in two Year 8 classes in New South Wales, and all students were allowed ample time to complete all items. The students' responses were analysed, with particular emphasis being placed on checking whether all 16 items on the trial instruments were well worded, and whether pairs of parallel questions were of the same order of difficulty.

One of the authors (Clements) carried out an extended Newman interview with one of the students. This interview was audiotaped and videotaped for future reference, and was observed live by the three research assistants. This interview served as the basis for discussion with the research assistants on how the extended Newman method should be applied in the interviews, and how data from such interviews could help identify the different strategies used by the students when they tackled each of the questions.

On the basis of the analysis of the students' responses to the trial instruments, minor changes to the two trial instruments were made, and two final 16-item instruments were constructed.

The Main Study

The two final instruments were given to 65 students in three Year 8 classes in three schools in New South Wales. Half the students in each class took the multiple-choice instrument first, and the other half took the short-answer instrument first (with students sitting next to each other taking different instruments). Subsequent analysis indicated that the order of taking the instruments had no influence on results. Sixteen of the 32 items (8 multiple-choice and the 8 parallel short-answer questions) were chosen as "interview" questions, and it was the students' responses to these questions which were analysed in the extended Newman interviews.

After the students had answered the questions on the final instruments, each student was given an extended Newman interview on 16 interview questions. These interviews, which were conducted by trained research assistants within three days of the students answering the questions, were audiotaped, and during the interviews detailed notes were taken by the interviewers on the thinking employed by each student on each question.

After the 3 interviewers had completed a total of 24 extended Newman interviews (8 each), they met with the authors and discussed the range of responses, including answers and strategies, which students had used for each of the 16 questions. Agreement was then reached on "Processing Scores" which would be associated with particular methods or ways of approaching items—"0" (no grasp of the concepts and relationships with which the question is concerned), or "1" (some understanding), or "2" (a sound understanding).

Processing Scores

After all 65 students had been interviewed, and each response tentatively allocated a Processing Score of 0, 1 or 2 by the interviewer, the three interviewers met with the authors and discussed the Processing Scores they had allocated. Where there was disagreement, this was used as an opportunity to sharpen the classification criteria. Through this iterative process, consensus was reached for the Processing Scores of the 65 students on all 16 questions.

Results

Sixteen 3 x 3 grids—one grid for each of the 16 questions—were constructed. A cell in any one of these grids showed the number of students (out of 65) whose responses to the question (covered by that grid) were classified as belonging to that cell. Table 1 shows a 3 x 3 grid with double entries for the two parallel questions on the area of a trapezium which were described earlier in this paper. The upper entry in each cell refers to the short- answer question, and the lower entry to the parallel multiple-choice question.

Table 1 Number of Responses in the 9 Cells on 2 Parallel Questions on the Area of a Trapezium (65 responses altogether, for each question

	No Understanding	Some Understanding	Full Understanding
Short-Answer (Q8) CORRECT	0	0	7
Multiple-Choice (Q9)	9	0	15
Short-Answer (Q8) INCORRECT	39	6	10
Multiple-Choice (Q9)	31	4	6
Short-Answer (Q8) NO ANSWER	2	0	1
Multiple-Choice (Q9)	0	0	0

The overall pattern of the results for the 16 questions is presented in Table 2. One dimension of the grid in Table 2 is concerned with whether correct, incorrect, or "no response" answers were provided; the other dimension is concerned with whether students had no understanding, some understanding, or a sound understanding of the concepts and relationships associated with the questions. Each cell of Table 2 contains two percentages: the upper entry in each cell, gives the appropriate composite percentage for the 8 short-answer interview questions, and the lower entry the composite percentage for the 8 parallel multiple-choice questions. Thus, for example, Table 2 indicates that 15.2% of all responses to the 8 short-answer questions were such that students gave incorrect responses even though they had a full understanding of the concepts, skills, and relationships inherent in the questions.

Table 2 Percentage of Responses in the Nine Cells from the 65 Year 8 Students on 16 Questions

	No Understanding	Some Understanding	Full Understanding
Short-Answer CORRECT	4.4%	1.2%	33.1%
Multiple-Choice	5.6%	1.2%	36.4%
Short-Answer INCORRECT	36.3%	5.8%	15.2%
Multiple-Choice	38.5%	6.0%	11.9%
Short-Answer NO ANSWER	3.1%	0.4%	0.5%
Multiple-Choice	0.4%	0%	0%

Discussion

If a question is such that students who have no understanding, or only partial understanding, of the concepts, skills and relationships associated with the question often give a correct answer to the question, then from a measurement perspective, the question is unsatisfactory. Similarly, if a question is such that students who have full understanding of the concepts, skills and relationships associated with a question often give an incorrect answer (or no answer) to the question, then from a measurement perspective, the question is also unsatisfactory.

In Table 1 27.5% of the short-answer classifications and 24.7% of the multiple-choice classifications lay outside the cells corresponding to "correct response and full understanding" and to "incorrect response (or no response) and no understanding." Thus, altogether, about one-quarter of the responses are associated with inadequate assessments of student understanding.

Entries in Table 1 draw attention to the limitations of assessment via short-answer and multiple-choice pencil-and-paper instruments. For the multiple-choice question, 9 of the 65 responses were correct, even though the students had no understanding; 6 of the students' responses were wrong yet the students were found to have full understanding. For the parallel short-answer question, 10 of the students who gave an incorrect response had full understanding.

The data and the analyses in this paper raise fundamental questions about the reliability and validity of pencil-and-paper multiple-choice and short-answer instruments which are used for the assessment of students' mathematical understanding. At issue, of course, is the meaning of the terms "reliability" and "validity."

Acknowledgement

We wish to acknowledge the contributions made to this study by Paul Bulbert, Paul Forbes, and David Graham who conducted interviews with students and engaged in discussion with the authors.

References

Clements, M. A. (1980). Analyzing children's errors on written mathematical tasks. Educational Studies in Mathematics, 11 *(1), 1-21.*

Ellerton, N. F., & Clarkson, N. F. (1992). Language factors in mathematics learning. In B. Atweh & J. Watson (Eds.), Research in mathematics education in Australasia 1988-1991 *(pp. 152-178). Brisbane: Mathematics Education Research Group of Australasia.*

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. Journal of Education Measurement, 22, *21–31*

Gays, S., & Thomas, M. (1993). Just because they got it right, does it mean they know it? In N. Webb & A. Coxford (Eds.), Assessment in the mathematics classroom *(pp. 130–134). Reston, VA: NCTM.*

Hembree, R, (1987). Effects of noncontent variables on mathematics test performance. Journal for Research in Mathematics Education, 18, *197–214.*

Newman, M. A. (1983). Language and mathematics. *Melbourne: Harcourt Brace Jovanovich.*

Rosier, M. J. (1980). Changes in secondary school mathematics in Australia 1964-1978. *Melbourne: Australian Council for Educational Research.*

Thongtawat, N. (1992). Comparing the effectiveness of multiple-choice and short-answer paper-and-pencil tests. Penang: SEAMEO/RECSAM.

Extra references to be incorporated and pilot study data to be included

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. Journal of Education Measurement, 22, *21–31.*

Yarroch, W. L. (1991). The implications of content versus item validity on science tests. Journal of Research in Science Teaching, 28, *619–629.*

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The students' perspective. Journal of Educational Research, 80, *352–358.*

Singh, J. (1991, Dec. 18). On idle: Objective tests harmful says lecturer. The Star. *Daily newspaper, Kuala Lumpur, Malaysia.*

46 students interviewed, 8 q's responses to 368 questions 184 short answer, 184 multiple choice, etc