Int. J. Engng Ed. Vol. 12, No. 6, p. 408–418, 1996
Printed in Great Britain.

0949-149X/91 $3.00+0.00
© 1996 TEMPUS Publications.

Multiple Choice Questions: The Value of Item Analysis*

V. FRITH and R. O. HECKROODT
Department of Civil Engineering, University of Cape Town, Private Bag, Rondebosch, 7700, Republic of South Africa

Multiple choice questions are used in an introductory engineering course at the University of Cape Town. Item analysis for each question provides information about facility, and about the relative plausibility of foils. Comparison of response patterns for the most able and the least able students yields information about discriminatory power, while the full variation of response pattern with ability reveals the ability level at which the question discriminates most effectively. This analysis is invaluable for the refinement and selection of questions for tests, examinations, tutorials and computer-based teaching aids. Item analysis also provides important feedback about teaching effectiveness.

INTRODUCTION

MULTIPLE choice questions are particularly useful in examinations in disciplines in which there is a requirement for a high recall of facts and where it is possible to design questions that test one piece of knowledge at a time. The advantages and disadvantages of multiple choice testing are well known [1–3], the main advantage being that the assigning of marks is objective and done automatically. The major disadvantage is that writing good questions takes considerable skill and practice, but above all, time. There are many references dealing with the principles and practice of question design [1–6], and a few concern themselves with applications to engineering education in particular [7–11].

Since examinations in engineering courses generally are required to test students' skill at more analytical types of problems, it appears that in general examinations in engineering fields use methods most appropriate to testing this kind of skill. There are however areas in the engineering curriculum where use of multiple choice questions is appropriate, especially in introductory courses of general scope. An example of such a course, where multiple choice questions have proved most useful, is an introductory course in engineering materials at the University of Cape Town (UCT). In this course questions of this kind have been found useful in class tests, examinations, and as the basis of group discussions in tutorials. Specific examples of these questions will be presented and discussed in this paper.

It must be stressed that effective use of multiple choice questions must include adequate analysis of how each question has performed. This is usually carried out by a computer system designed for the purpose. At UCT this system allows for automatic entering of the students' choice of answers and provides the lecturer with an easily interpreted printout of the analysis results. (Figures in this paper will illustrate the format of this printout.) The analysis system was written and is administered by User Support Services, a section of Information Technology Services at UCT. (It is available from the following address: Information Technology Services, University of Cape Town, Private Bag, Rondebosch, 7700, South Africa.)

One of the great advantages of multiple choice testing is that statistical analysis can readily be carried out (usually by computer) on the results obtained by students taking such a test. This provides quantitative feedback on the quality and validity of the questions, so that the best questions can be collected for future use [12], and poorer questions can be redrafted or rejected [13]. The individual multiple choice question is often referred to as an item and the statistical analysis of the results obtained for each question is known as item analysis.

It must be stressed that the value of item analysis extends beyond this testing of items, because the analysis also provides very valuable feedback for lecturers on the effectiveness of their teaching [14] as well as on the overall abilities and past learning of their students [13]. For example, if the majority of students select one particular wrong answer, their misconception may often be traced back to a misleading or inadequate presentation in lectures or notes. This pedagogical value of item analysis data has also been stressed by Lindeman [14], Leuba [8] and by Boone and DeMay [11].

Although there are numerous procedures available for carrying out item analysis, the following are the three most important types of information, which are sufficient for item selection and for providing feedback on teaching quality:

* Accepted 14 January 1997.

- facility,
- response pattern,
- discrimination.

Facility

The proportion of candidates who selected the right answer to an item provides a measure of its facility [13]. This is sometimes referred to as the 'difficulty' of the item [14], but since a higher proportion correct would indicate an easier question, the term 'facility' is preferred here. This is the most basic type of information derived from item analysis.

Response pattern

More refined information can be derived by looking at the proportions of candidates selecting different alternative answers within each question. From this kind of information one can identify totally implausible foils (incorrect answers or distractors), or foils that are too misleading, so that they can be rewritten or rejected. This kind of information is also most useful in identifying parts of the course that were not taught well enough [11].

Discrimination

While analysing the general response patterns is considerably more useful than measuring facility alone, by far the greatest value can be derived from comparing the response patterns when students are classified into groups according to their ability. This provides information about discrimination.

The discriminating power of a question is the extent to which it distinguishes the more able students from the less able and since the object of a test is very often to rank the students in order of ability (norm-referencing), a high degree of discrimination is obviously desirable. The discrimination is not independent of facility; a question that everyone can answer correctly or that no-one can answer, does not discriminate at all. A question with 50% facility allows for the greatest possible discriminatory power, although not all such questions necessarily display the maximum possible discrimination [1].

An efficient way to obtain information about an item's power to discriminate is to examine the gradients in the response patterns for that item, for example, the way in which the percentage correct responses varies with ability.

EXAMPLES AND DISCUSSION

Selecting items for use in future examinations

Some specific examples taken from the final examinations (written by 125 candidates) in an

Q 1	Whole class	Top	2	3	4	Bottom
Answer 1	50	80	62	42	38	28
Foil 2	13.7	0	4	19	25	20
3	11.3	8	8	12	8	20
4	2.4	4	4	0	4	0
5	0.8	0	4	0	0	0
Abstain	21.8	8	18	27	25	32

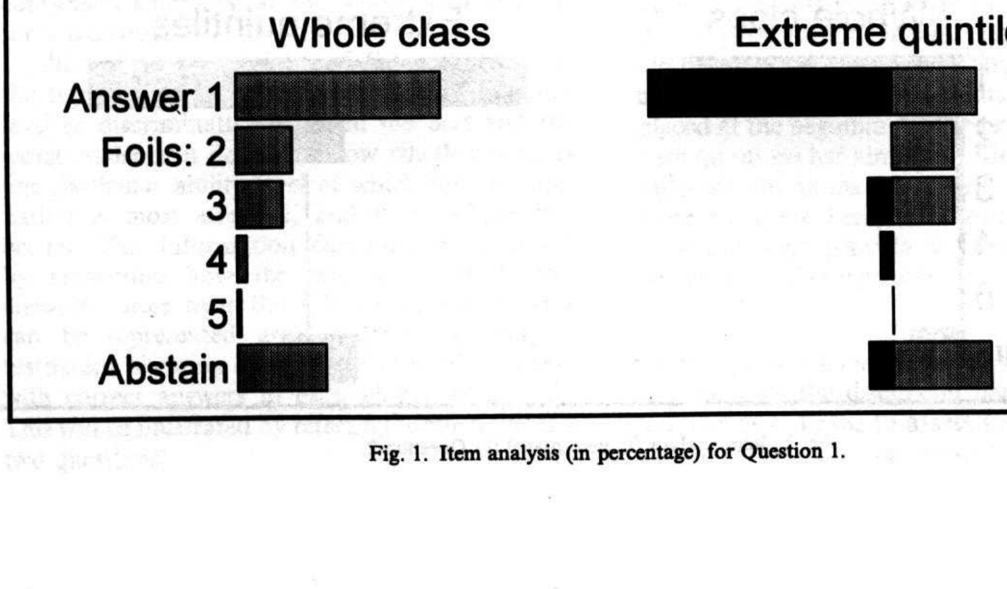

Fig. 1. Item analysis (in percentage) for Question 1.

introductory course on engineering materials will serve to demonstrate the usefulness of a comprehensive item analysis.

Question 1

Barrelling of the specimen during loading in a compression test is caused by:

Answer

1. frictional forces between platten and specimen

Foils

2. incorrect ratio between height and diameter of the specimen
3. development of bending stresses due to incorrect loading set-up
4. inhomogeneities in the texture of the specimen
5. the specimen being too large for the capacity of the testing machine

Question 2

A tensile stress on a 1 cm long specimen of a material causes a recoverable elongation of 5 cm. The material is:

Answer

1. a polymer

Foils

2. none of these
3. a composite
4. a metal
5. a ceramic

The full output from the item analysis of questions 1 and 2 is presented in Figs 1 and 2. (In all the examples and figures in this paper the foils have been re-arranged into order of decreasing number of responses. This is not necessarily the order in which they appeared in the examination.)

Before the usefulness of this data can be discussed, it is necessary to digress for a moment to explain how the table with column headings: Top, 2, 3, 4 . . . is constructed. The five columns in the table refer to the five ability groups (in this case quintiles), determined on the basis of the students' total test score. Thus the top quintile will represent the 20% of students who gained the highest marks on the test as a whole. Each row represents one of the alternative answers and the numbers in the table are the proportion (as a percentage) of students in that ability group who selected that particular alternative. The bottom row in the table represents the students in each quintile who did not answer the question (abstained).

In both Fig. 1 and Fig. 2 the information on the extreme left represents the proportion of the student body as a whole who selected each of the possible alternative answers. In Fig. 1 the facility

Q 2	Whole class	Top	2	3	4	Bottom
Answer 1	68.6	72	76	70	54	72
Foil 2	15.3	12	4	15	25	20
3	1.6	0	0	4	4	0
4	0.8	0	4	0	0	0
5	0.8	0	4	0	0	0
Abstain	12.9	16	12	11	17	8

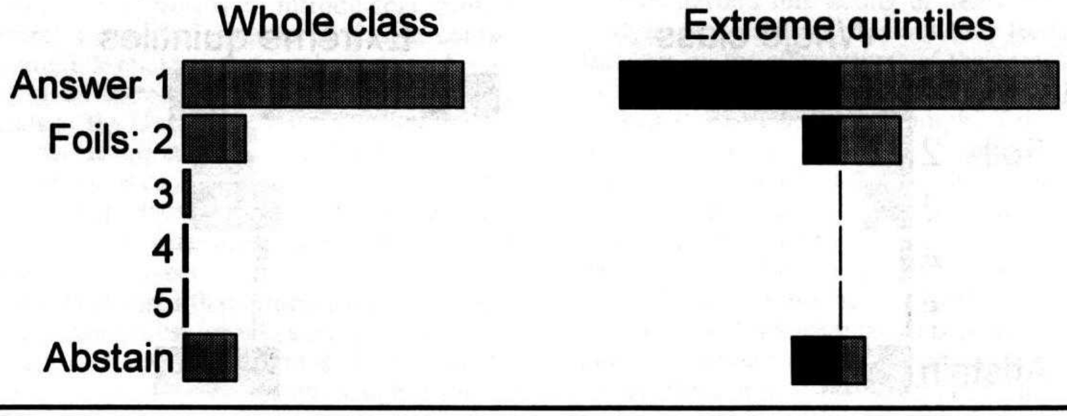

Fig. 2. Item analysis (in percentage) for Question 2.

is 50%, which is usually regarded as ideal, while in Fig. 2 it is 68.6%, which could indicate that the question is a little easy, but still quite acceptable. (This difference in facility is consistent with a subjective evaluation of the two questions: the answer to the question about the maximum elongation in polymers is more obvious.)

However, the really important difference between these questions only becomes apparent on examination of the table representing response pattern, particularly the gradient in the rows representing the variation in % correct with ability. In Fig. 1 the proportion of correct answers is greatest for the top quintile and steadily decreases as the ability decreases, while in Fig. 2 there is no such dependence on ability. Thus, while the two questions are superficially similar (from looking at the facility and the proportions selecting each alternative), question 2 is regarded as poor on the grounds that it has no discriminatory ability, while question 1 is regarded as ideal and would be retained for future use.

From the table in Fig. 1 one can also see an increase in numbers of students selecting the two most popular foils as ability decreases, so that for the poorer students the correct answer is hardly more attractive than the popular foils. The good students are less likely to be attracted by any of the foils, which indicates that on the whole this is a well-written question with a good ability to discriminate.

There is however one respect in which this question can be improved. The alternative 'The specimen is too large for the capacity of the testing machine' surprisingly attracted only one response, and is thus serving no practical function. If possible it should be replaced by a more plausible foil.

The differences in the ability of the two questions to discriminate can be seen at a glance by referring to the diagrams labelled Extreme Quintiles in Figs 1 and 2. In these diagrams the proportion of students who chose each alternative answer in the top quintile is compared to the proportions who chose each alternative in the bottom quintile (column 5). In fact this particular kind of diagram really provides a very convenient summary of the whole item analysis for a question.

Although a comparison of the responses for the top and bottom quintiles reveals the overall level of discrimination between the best and the worst students, it does not show whether there is one particular ability level at which the discrimination is most effective, and if so where this occurs. This information can only be obtained by examining how the percentage of correct answers varies over the full ability range. This can be represented graphically by plotting a histogram showing the proportions of students with correct answers in each ability group [13]. This will be illustrated by referring to the following two questions.

Question 3

A defect consisting of a displaced ion in the lattice is called . . .

Answer

1. a Frenkel defect

Foils

2. a Schottky defect
3. an intersticialcy
4. a di–vacancy
5. a stacking fault

Question 4

Which of the following will be a consequence of finer grinding of Portland cement?

Answer

1. decreased workability

Foils

2. decreased shrinkage
3. decreased rate of strength development
4. decreased rate of heat generation
5. decreased rate of hydration

Histograms of the response patterns of questions 3 and 4 are shown with the full output for those questions in Figs 3 and 4. Question 3 is regarded as an easy question (facility 78%) while the students apparently found question 4 to be moderately difficult (facility 39.2%).

The sizes of the 'steps' in the histogram give an indication of its discriminatory power. In addition the level of ability at which the item discriminates most effectively can also readily be seen. For example it can be seen that the question in Fig. 3 discriminates only between students of the lowest ability, while in Fig. 4 the biggest step occurs at the top of the ability range. This kind of information is only available from a full tabulation of response patterns, even the diagram comparing the patterns for the extreme quintiles does not reveal differences of this sort. (Notice that the 'Extreme Quintiles' diagrams in Figs 3 and 4) indicate a similar degree of overall discrimination).

It is useful to determine the ability level at which items discriminate because, while the majority of questions in a norm-referenced test (one aimed at ranking students) should discriminate between the mass of students in the middle of the ability range, it is a good idea to include some questions that almost everyone can answer. This provides encouragement to the poorer students [14], especially if placed at the beginning of the test. The presence of such questions has almost no effect on the final test outcome (in terms of student rankings) and is therefore quite benign. Likewise a few difficult questions that provide a challenge to the top students are also desirable.

Gaining feedback on effectiveness of teaching

Until now the discussion has focused on item analysis as a means to assess the quality of items, so that they can be retained for future use,

Q 3	Whole class	Top	2	3	4	Bottom
Answer 1	78.2	88	84	84	80	56
Foil 2	6.5	4	0	4	12	12
3	4.8	8	8	0	0	8
4	3.2	0	0	0	4	12
5	0.8	0	0	0	0	4
Abstain	6.5	0	8	12	4	8

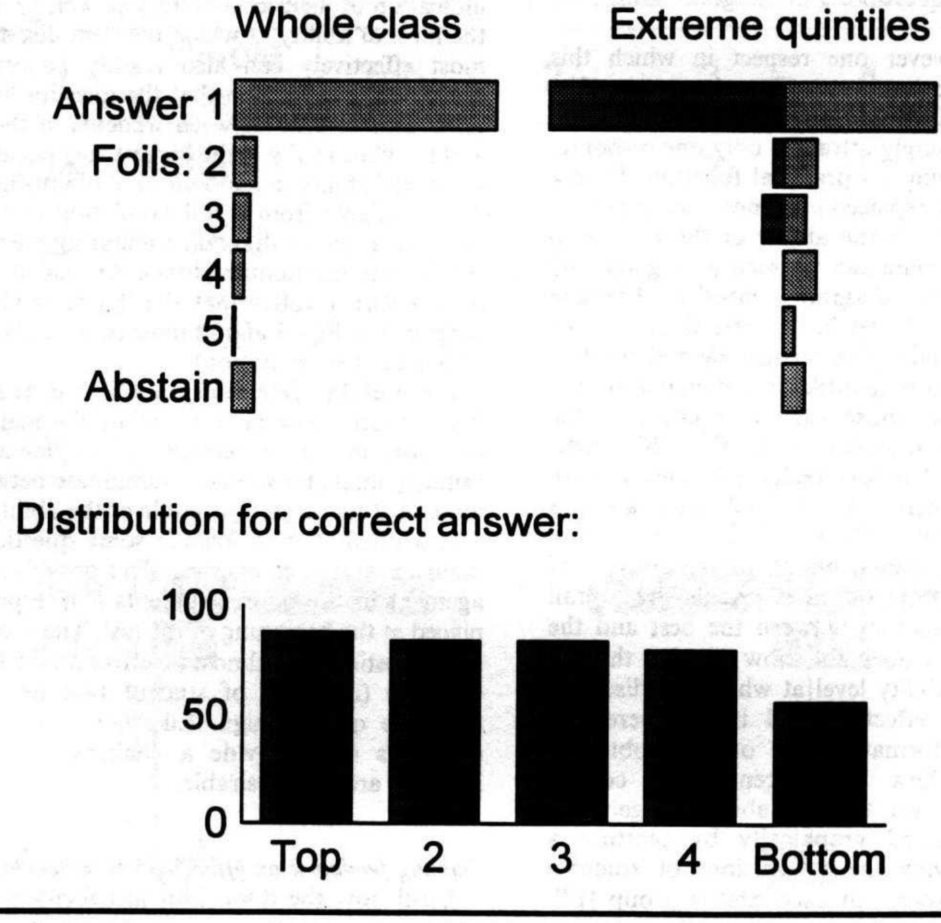

Fig. 3. Item analysis (in percentage) for Question 3.

Q 4	Whole class	Top	2	3	4	Bottom
Answer 1	39.2	80	31	34	13	37
Foil 2	25.5	0	38	22	50	18
3	5.9	0	0	11	0	18
4	5.9	10	8	0	0	9
5	3.9	0	0	0	12	9
Abstain	19.6	10	23	33	25	9

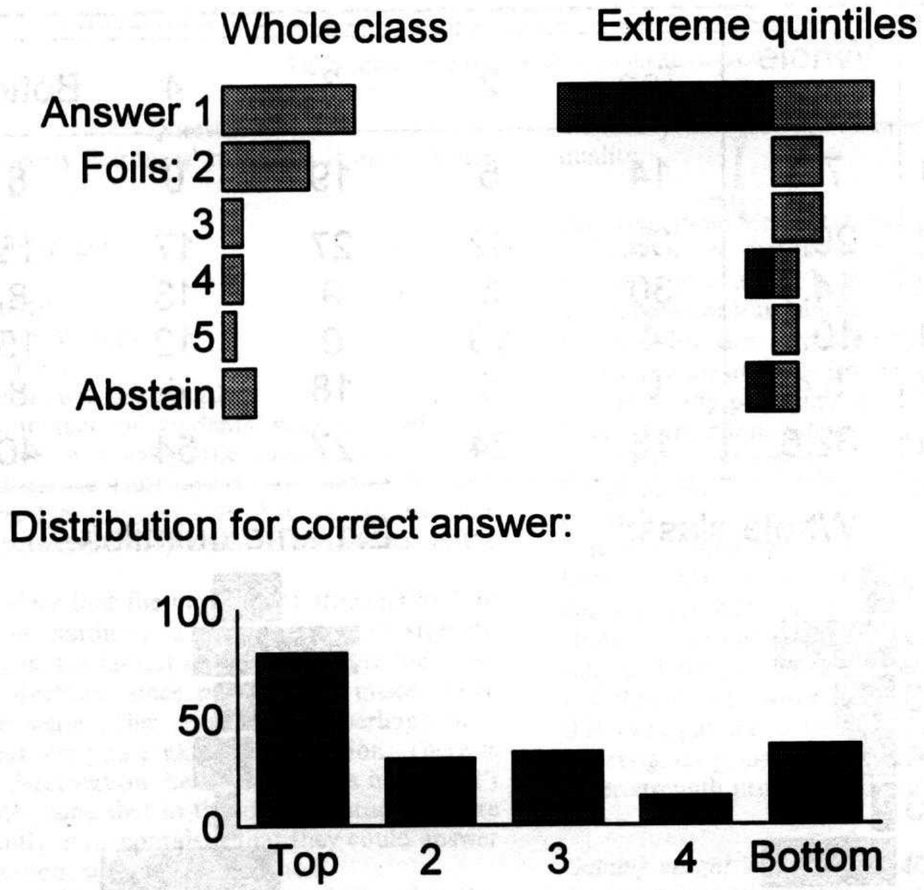

Fig. 4. Item analysis (in percentage) for Question 4.

redrafted or rejected. However, when the analysis indicates that an item has not performed well, it may be that it is not the question, but the teaching of the relevant topic that is at fault.

This point can be illustrated by referring to the following question.

Question 5

Calculate the modulus of rupture (MOR) of a rectangular bar in 4-point bending (centre-thirds loading). The length of the specimen is 40 mm, its width 10 mm and its height (thickness) 5 mm. The span of the loading system is 10 mm and the fracture load 2.4 kN.

Answer
1. 288 MPa

Foils
2. 384 MPa
3. 96 MPa
4. 1440 kPa
5. 720 kPa

The formula the students were required to recall is: MOR $= (Ps)/(wd^2)$, where P is load (N), s is span (m), w is width (m), d is depth (m) and l is length of specimen (m). Note that all the values needed to perform the calculation were given, as well as the length of the specimen, which is irrelevant to the calculation. The output of the item analysis for this question is presented in Fig. 5.

The item analysis displays an inversion; the incorrect alternative (number 2 above) attracted far more responses than the correct answer, even in the top quintile. This reveals that, even those students who remembered the formula correctly, did not realise that s in the formula represents the span, not the specimen length (l). The vast majority of the students clearly did not understand the concepts involved in testing for modulus of rupture, and at best had learned the formula blindly. It is also significant that almost a third of the students did not even attempt to answer this question.

There does not seem to be anything wrong with the way this question is written and the only reasonable conclusion that can be drawn is that the students did not understand this topic. Since this question was used in a class test there was an opportunity to attempt to rectify this situation before the final examination.

An inversion of this kind is a fairly uncommon occurrence. A more frequent problem is represented by the output in Fig. 6, which is from the following question.

Q 5	Whole class	Top	2	3	4	Bottom
Answer 1	7.7	14	5	19	0	8
Foil 2	29.3	39	42	27	17	15
3	14.1	30	5	9	13	8
4	10.9	4	19	0	12	15
5	5.4	0	5	18	4	8
Abstain	32.6	13	24	27	54	46

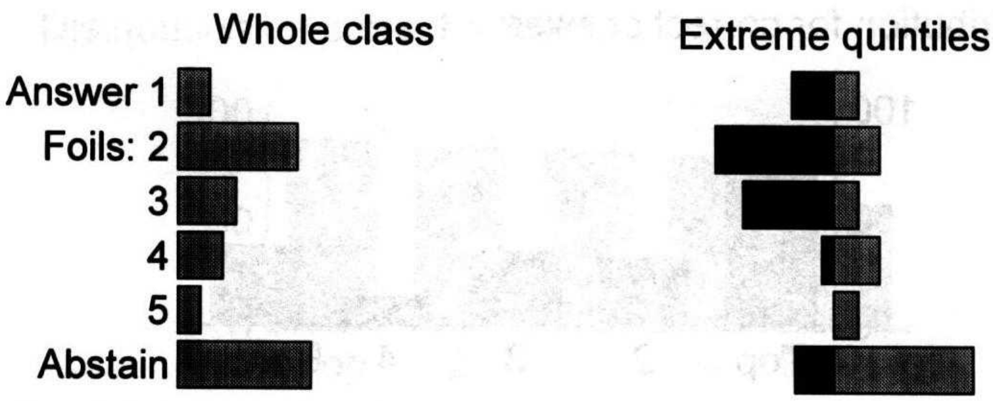

Fig. 5. Item analysis (in percentage) for Question 5.

Q 6	Whole class	Top	2	3	4	Bottom
Answer 1	20.6	44	5	19	13	23
Foil 2	26.1	13	43	45	25	8
3	20.7	17	19	0	33	23
4	16.3	13	14	18	12	31
5	16.3	13	19	18	17	15
Abstain	0	0	0	0	0	0

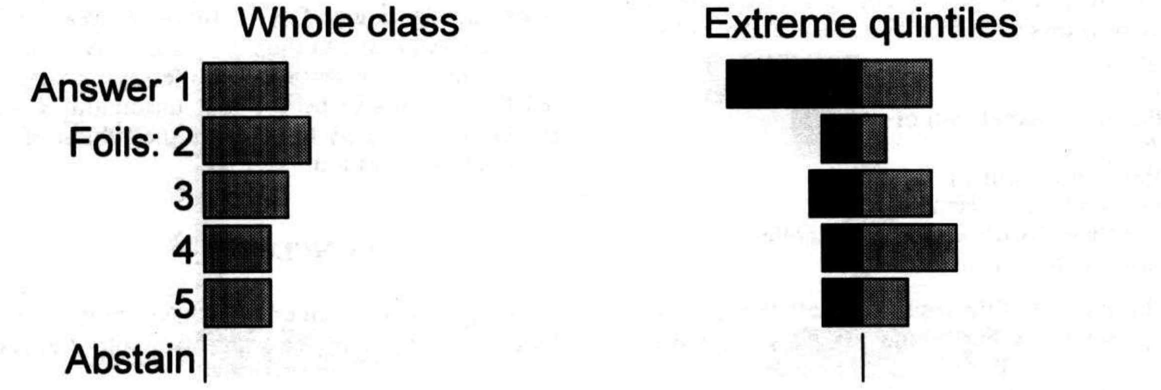

Fig. 6. Item analysis (in percentage) for Question 6.

Question 6

The property 'hardness' in metals is most closely related to . . .

Answer

1. yield strength

Foils

2. ductility
3. fracture strength
4. toughness
5. ultimate tensile strength

The number of students who selected each alternative is roughly the same; slightly more selected the incorrect answer, alternative 2 above. However the question does discriminate to some extent between the top students and the rest of the class.

It is clear that the majority of students had no idea how hardness is related to other strength properties, the largest number choosing the alternative, ductility, since perhaps this makes some intuitive sense. (They reasoned thus perhaps: 'In a hardness test you make an indentation. There is plastic deformation, hence ductility is important') It is interesting that in this case the students were apparently quite confident that they could answer the question, since none abstained.

Once again the conclusion was drawn that this section of work was not well understood. Clearly feedback of this sort is very revealing and can ultimately be most useful in improving teaching quality.

Selecting questions for tutorial discussion

Question 6 discussed above was used the following year as the basis of group discussion in a co-operative learning tutorial and was found to be ideal for this purpose, precisely because the students were not unanimous in their decisions in favour of any particular answer. It is exactly this kind of question, where responses are roughly equally divided amongst the different foils, that yields the most rewarding subjects for discussion in tutorials.

It was found that using multiple choice questions in this way provided a focus for the discussion; for instance in the above example, the students were instructed to discuss each alternative fully and clarify why that alternative was, or was not, the correct answer, as the case may be. Thus in this example the discussion helped clarify their understanding, not only of hardness, but of the other strength properties as well.

Gaining insight into student learning behaviour

A comparison of the item analyses of the following two very similar questions (which were used together in the same examination) will serve as an example of how item analysis can sometimes give insights into strategies used by students when preparing for examinations.

Question 7

In the manufacture of Portland cement, the presence of excess lime, magnesia and/or sulphates in the milled cement is likely to cause:

Answer

1. unsoundness of cement

Foils

2. false set of cement
3. insufficient workability of concrete
4. disruptive expansion of concrete
5. flash set of cement

Question 8

In the manufacture of Portland cement, the presence of excess alkalis in the milled cement is likely to cause:

Answer

1. disruptive expansion of concrete

Foils

2. flash set of cement
3. unsoundness of cement
4. insufficient workability of concrete
5. false set of cement

The analysis of the results of these two questions are presented in Figs 7 and Fig. 8. The questions appear to be equivalent in terms of the demands they make on the students and one would expect these questions to have similar facility. They have the same set of alternative answers and the wording of the stem of the questions is equivalent. However the results of the item analysis in Fig. 7 show that the students found this question very easy (facility 52% for the bottom quintile), while the question in Fig. 8 had only 25% facility for the whole class and a high level of abstaining.

Having eliminated in this case the possibility of a fault in the presentation of this topic in lectures, it was concluded that the students had 'spotted' Question 7, particularly as unsoundness in cement is a topic that had previously frequently received attention in examination papers. This kind of insight proved very instructive to the lecturer designing the course for the following year, as it was necessary to examine very carefully what it was about the presentation of the topic that had led the students to believe that unsoundness was the fault most likely to be made the subject of an examination question.

CONCLUSIONS

Multiple choice questions have been found useful and appropriate to an introductory course

Q 7	Whole class	Top	2	3	4	Bottom
Answer 1	76.6	96	88	85	62	52
Foil 2	4.1	0	0	0	9	12
3	3.2	4	0	0	13	0
4	2.4	0	0	4	4	4
5	1.6	0	0	0	0	8
Abstain	12.1	0	12	11	12	24

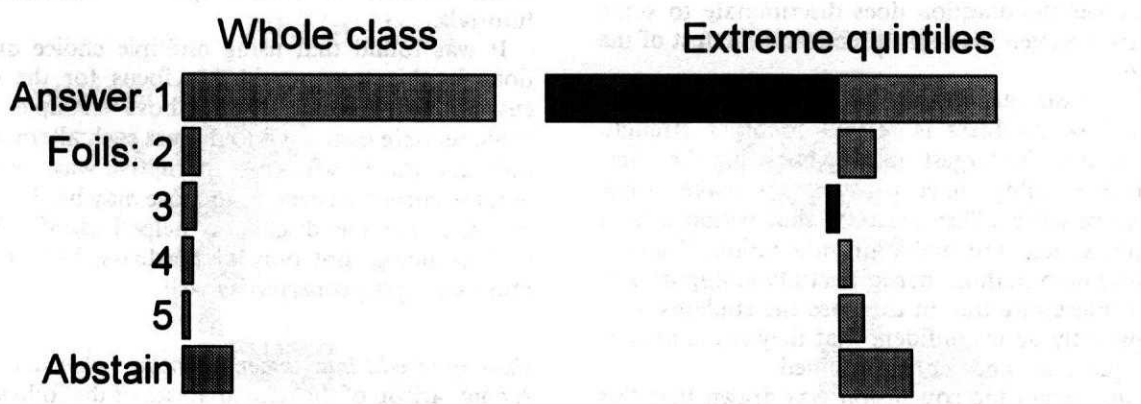

Fig. 7. Item analysis (in percentage) for Question 7.

Q 8	Whole class	Top	2	3	4	Bottom
Answer 1	24.9	48	26	34	13	4
Foil 2	11.3	20	12	8	17	0
3	7.3	0	0	4	17	16
4	6.5	4	0	4	8	16
5	5.6	0	4	8	12	4
Abstain	44.4	28	58	42	33	60

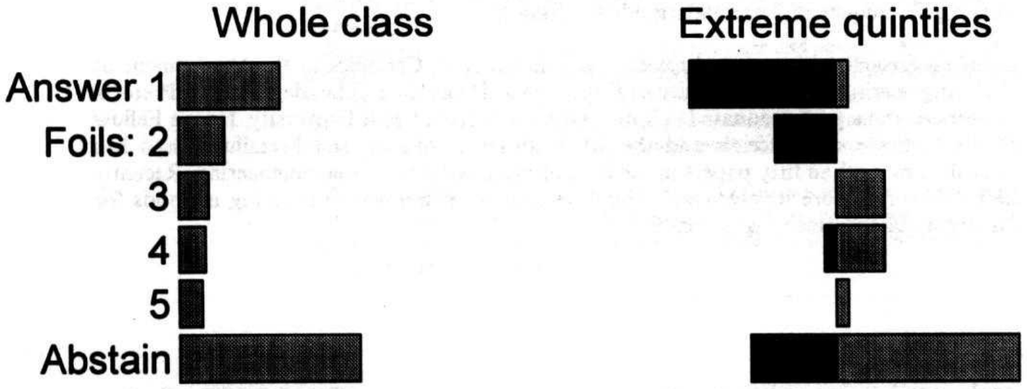

Fig. 8. Item analysis (in percentage) for Question 8.

on engineering materials at the University of Cape Town. They are used as a component of tests and examinations and as the basis for discussions in co-operative learning tutorials.

It is very important to do a full item analysis on questions used in tests and examinations. The response pattern for the class as a whole does not reveal anything about the discriminatory power of a question, but may be useful as a quick guide to facility and relative plausibility of the different alternative answers. A comparison of the response pattern for the extreme quintiles (of the distribution of student scores) provides a useful summary of the item characteristics, especially the degree of discrimination. However, it is necessary to look at the full response pattern for all ability groups to identify at which ability level a question discriminates most effectively.

Performing item analysis is invaluable for item selection for future examinations and tutorials and for use in interactive computer-based teaching aids. The ideal question for exam purposes has approximately 50% facility and a high degree of discrimination, while for tutorials the questions with 'non-ideal' statistics (poor discrimination, inversions even) have proved most effective. Item analysis also provides extremely useful feedback on the effectiveness of teaching.

REFERENCES

1. R. Wood, Multiple choice: a state of the art report, in: B. H. Choppin and T. N. Postlethwaite (eds.), *Evaluation in Education,* **1,** Pergamon, UK (1979).
2. J. Brown, *Objective Tests: Their Construction and Analysis. A Practical Handbook for Teachers,* Longmans, London, p. 5 (1966).
3. D. G. Lewis, *Assessment in Education,* University of London, p. 107 (1974).
4. G. H. Miller, R. G. Williams and T. M. Haladyna, *Beyond Facts: Objective Ways to Measure Thinking,* Englewood Cliffs, NJ (1978).
5. J. R. Hills, *Measurement and Evaluation in the Classroom,* Merrill, Ohio, p. 30, (1981).
6. J. M. Thyne, *Principles of Examining,* University of London, p. 185 (1974).
7. R. J. Leuba, Machine-scored testing, Part I: purposes, principles and practices, *Eng. Ed.* **77,** pp. 89–95 (1986).
8. R. J. Leuba, Machine-scored testing, Part II: creativity and item analysis, *Eng. Ed.* **77,** pp. 181–186 (1986).

9. P. C. Wankat and F. S. Oreovicz, *Teaching Engineering,* McGraw-Hill, USA, p. 219 (1993).
10. D. P. Kessler, Machine-scored versus grader-scored quizzes an experiment, *Eng. Ed.* **78,** pp. 705–709 (1988).
11. E. Boone and G. DeMay, Some practical aspects of multiple choice examinations, *Eur. J. Eng. Ed.* **12,** pp. 325–328 (1987).
12. C. H. Nelson, *Measurement and Evaluation in the Classroom,* MacMillan, London p. 49 (1970).
13. B. Hudson, *Assessment Techniques, an Introduction,* Methuen Educational, London, p. 150–156 (1973).
14. R. H. Lindeman and P. F. Merenda, *Educational Measurement,* 2nd Edition, Scott Foresman, USA (1979).

Vera Frith is a Senior Scientific Officer in the Ceramics Research Group in the Department of Civil Engineering at the University of Cape Town. She received her M.Sc. in Applied Science from the University of Cape Town in 1983. Her research work has been in the fields of image analysis, microstructure and body design for whitewares. In 1993 she obtained a Higher Diploma in Education at the University of Cape Town. Her interests now include mathematics education, uses of multiple choice questions and application of co-operative learning techniques in large undergraduate classes.

Oelof Heckroodt is Associate Professor and Lecturer in Ceramics in the Department of Civil Engineering at the University of Cape Town. He holds a D.Sc. degree from Pretoria University and a post-graduate Diploma in Ceramics from Leeds University. He is a Fellow of the Institute of Materials and the SA Institute of Mining and Metallurgy and has published more than fifty papers in the field of materials science and engineering. Recently he has become more involved with the development of innovative teaching methods for undergraduate engineering students.