CHEMISTRY EDUCATION:
RESEARCH AND PRACTICE IN EUROPE
2000, Vol. 1, No. 3, pp. 323-328

SHORT REVIEW OF RESEARCH
(INVITED CONTRIBUTION*)
Assessment

Alex H. JOHNSTONE and Abdullah AMBUSAIDI
University of Glasgow, Centre for Science Education

FIXED RESPONSE: WHAT ARE WE TESTING?

Received: 26 May 2000

ABSTRACT: Objective testing is playing a larger part in higher education. This paper sets out to look critically at this form of assessment in the light of research so that examiners can take a more informed view of the technique. [*Chem. Educ. Res. Pract. Eur.*: 2000, *1*, 323-328]

KEY WORDS: *Assessment; objective testing; multiple choice questions; higher education*

* *Editor's note*: This paper was to be part of the Invited Symposium on Assessment of the 5th ECRICE, but Prof. Johnstone was unable to present it at the Conference. The initial title was 'Towards invalidity in assessment'.

INTRODUCTION

Fixed response (objective) testing has been around for a long time. Early in the 1900's tests of this kind were being developed to match the interest in measuring the Intelligence Quotient (I.Q.) of children (Black, 1998).

The commonest form of question was the familiar multiple choice in which a question or statement (the stem) was presented followed by a series of choices from which a selection had to be made. The test answer (the key) was accompanied by other options, the function of which was to distract (distracters).

All sorts of ingenuity was applied to make variants on the basic multiple-choice format. A block of data was presented and a clutch of questions was set using the common data. Others provided a set of categories and examinees were required to fit instances into these categories. Pattern seeking in the form of "odd-one-out" questions became, and still is, popular. Mensa frequently uses questions of this type to select its "chosen few" along with series questions where a sequence of numbers or pictures is offered and the testee has to add the next in the series.

There are many more types and variants of fixed response questions, but their common factor is, that the testee is confined within the question to select "the" response, which has been preordained by the setter to be "correct".

This seems to be an attractive form of testing when a teacher is faced with assessment of a large class. The prospect of marking hundreds of free response scripts by hand while consuming midnight oil, and other beverages, is not a pleasant one. How much easier to have fixed response questions which students answer with a mark on a card and then a machine marks them. National

examination boards, schools and universities use the technique. The American Chemical Society and other bodies use fixed response papers to place students in higher education and some North American universities use such tests throughout their courses.

The tests have an appearance of sharpness and precision and are capable of objective machine marking and statistical manipulation, which appeals to scientists. Although they are called objective tests, the only thing about them which is objective is the marking. The objectives, which they purport to test, have been subjectively chosen. The questions have been written subjectively to fit these objectives. When the test is over, subjective judgement is made about the score, which will be deemed to have passed. The advantages of this kind of test need to be examined more critically This paper will be devoted to such an examination, not to debunk fixed response testing, but to help practitioners to apply realistic caution to their use.

Courses and Outcomes

Before we proceed with that, we must ask some questions about our courses and their outcomes. What kind of behaviour do we expect to see in our students as a result of their having attended our courses? We certainly expect to see an increase in knowledge, an ability to use that knowledge routinely, evidence of understanding and an ability to apply that understanding in new and creative ways. We also want to see a steady movement from teacher dependence to self-confident, self-motivated learning. If we are going to recognise the achievement of these aims, we need some assessment instruments to help us. However, there is no point in using a thermometer to measure pressure or using a ruler to measure mass. In the field of assessment, absurdities of the kind mentioned above are perpetrated daily. One of the most absurd is the inappropriate use of fixed response questions.

Some concerns about Fixed Response Assessment

Before we look at research findings in this area common sense should alert us to some problems. The often-voiced complaint is that students guess from the menu offered and this is not unlikely. All kinds of statistical ingenuity have been applied to offset this. Wrong responses are sometimes penalised by the subtraction of some fraction of a mark to discourage guessing. A recent study (Burton & Millar, 2000) shows the futility of this on statistical grounds. If the purpose of the test is to place students in order of merit, there is no need for any deduction, because the rank order correlation between the raw scores and the "doctored" scores is usually in excess of 0.95!

The examiner must also decide whether an "educated" guess is preferable to no response at all. Professionally we often are in a position when an educated guess is necessary.

Another measure, which is often taken to reduce the effect of guessing, is to increase the number of distracters. Some examiners prefer to use five option questions to four option questions in the belief that guessing is reduced from 25% to 20%. However, the effort required to produce a fifth plausible distracter is so great that it hardly warrants the effort to reduce blind guessing by such a small amount. So many fifth distracters attract few, if any, students and so tend to reduce the questions to four options. Research done many years ago (Handy & Johnstone, 1973) showed that students tackle fixed response questions in two ways; by recognition of an answer or by elimination of distracters. If guessing has to take place, it is between the uneliminated options (often two) making the guessing factor 50%! Some writers see this latter technique as a multiple true-false situation.

There are other considerations besides guessing, which might make us cautious about fixed-response testing, the most important of which is the nature of the distracters. The examiner tries to offer options, which are plausible, but are they necessarily plausible to the students? By the very nature of fixed-response questions, there is no way of answering the most fundamental of questions: "Why did the student choose this option". Students may choose the "correct" option for an entirely wrong reason or may choose a "wrong" option for a very good reason; but there is no way for the student to reveal his reasoning. A recent study revealed that *30% of students get the right answer for the wrong reason* (Tamir, 1990). It is not unusual to find questions in which the "best" students, on the test as a whole, do badly, while the "weaker" students do well. This is called negative discrimination. The good students are seeing something in the distracters, which the weaker students are not seeing and so are not being distracted. Some examiners use the test results to see which "wrong" response was most popular and then assume that this indicates a learning difficulty or a misconception. However, since the students have no "say" in what the distracters are, they are not able to reveal their actual misconceptions. Some fixed response items take a form not far removed from begging the question such as, "When did you stop beating your wife?"

Fixed response questions are sometimes denigrated because they seem to test "only recall" of information; but do they test "recall" or "recognition

These are two quite different, but related, skills. The reader will appreciate the experience when, unsuccessfully trying to recall a name, she can still recognise that name from a list. Generally recognition is an easier process than recall. Students often operate on fixed response questions by recognising a key word in one option (Cassels & Johnstone, 1978; 1984) or being deflected from the "correct" answer by an unfamiliar word. It is possible to increase the success rate (facility value) of a question, by up to 20%, by using an easily recognised word.

Fixed response questions are capable of operating in ways other than recognition. It is possible to test at levels of comprehension (using information in a familiar way), application (using information in an unfamiliar way) and some other Bloom categories (Bloom, 1956). There has been a tendency to regard these categories as a hierarchy with recognition or knowledge at the bottom of the heap. This is a rather narrow view since none of the other levels can operate without recall of information or techniques. An "umbrella" diagram may be a better representation (Figure 1) of the various aspects of testing than a hierarchy implying superiority. Knowledge and basic skills underpin and support the others, which are modes of use of knowledge.

Research in fixed response assessment

Much work has been done in this field to measure the reliability of this form of assessment. The word "reliability" has a technical meaning in that a reliable test is one, which would give very similar results when applied on more than one occasion to the same (or similar) group of students. It is the same meaning of reliable, which could be applied to any measuring

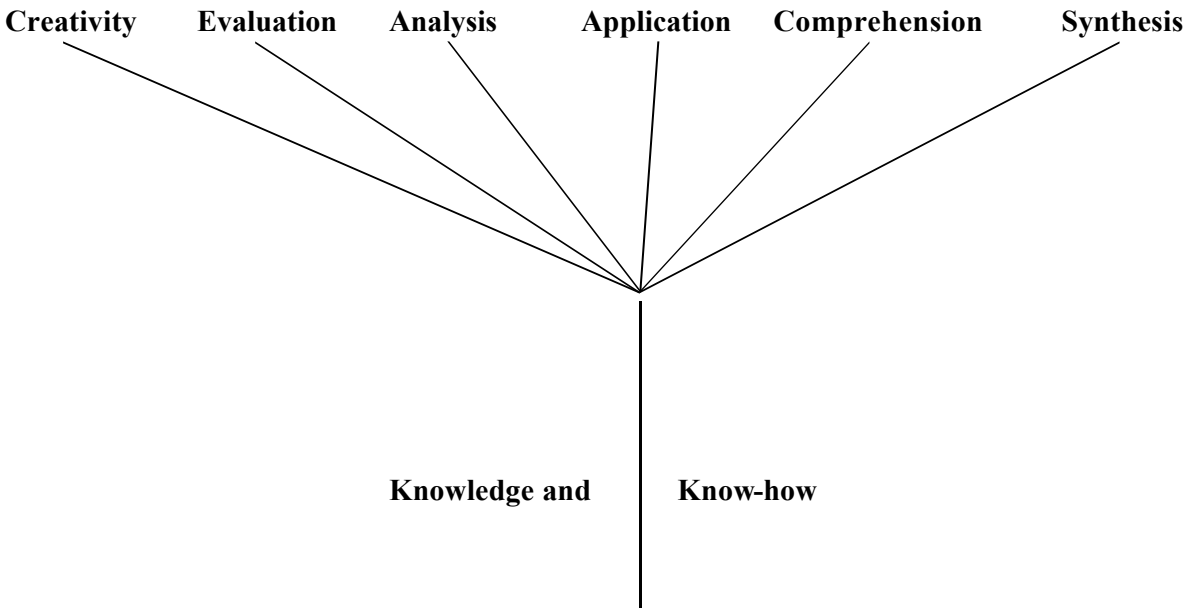

FIGURE 1. *Umbrella diagram (Johnstone after Bloom).*

instrument. A given thermometer dipped into boiling water on two occasions should give the same reading. In general, the research has shown that fixed response tests are reliable if applied UNCHANGED to two similar groups of students. Table 1 shows the results of applying a test simultaneously to two groups of students from the same class. The facility value (F.V.) is the fraction of the sample choosing the "correct" answer.

However, if any changes are made to the test, the reliability of individual questions is severely reduced. Table 2 shows the effect of moving the distracters within a question (Johnstone & Ambusaidi, 2000). Notice the effect upon the Facility Value (the proportion of the class choosing the best answer*). The facility value depends upon the chance arrangement of the options in the question. Not a single word has been changed in the questions, only the position of the distracters.

Changes in Facility Value are also reported when the position of a question (unchanged internally) is altered in the test (Ace & Dawis, 1973). When a test is being compiled from a bank of pretested questions (of known facility value), the facility values may no longer remain stable

TABLE 1. *Facility values* for the same test on similar populations.*

Question	F.V. for Group 1 (N = 100)	F.V. for Group 2 (N = 100)
1	0.86	0.88
2	0.43	0.38
3	0.87	0.83
4	0.69	0.72
5	0.66	0.70

* Facility Value (F.V.) = proportion of group correct

TABLE 2. *Response pattern changes with distracter order changes.*

QUESTION 1		QUESTION 2		QUESTION 3	
Test1	Test 2	Test 1	Test 2	Test 1	Test 2
A. 15.8	C. 22.9	A. 15.4	D. 5.8	A. 9.9	B. 8.9
B. 38.4 *	B. 48.7 *	B. 48.5 *	B. 60.4 *	B. 14.7	D. 9.9
C. 39.6	A. 25.3	C. 9.9	C. 18.6	C. 58.2 *	C. 64.7 *
D. 4.3	D. 1.2	D. 22.1	A. 13.9	D. 15.7	A. 15.2

* = correct answer

and this is particularly obvious if the questions are edited to give an even spread of A, B, C and D as the "correct" options (Ace & Dawis, 1973).

If you have a large class, which has to be split into two groups for examination on two occasions, you might want to use the same test. To avoid 'news' from the first group being passed to the second group, you may scramble the order of the questions for the second test, but keep the questions themselves unchanged. There is no guarantee that the two tests are now comparable. The averages of the two may be similar, but the performance of specific questions between the tests may well differ significantly!

This brings us to the prickly question about the validity of this type of assessment. Validity means that the test is measuring what you set out to measure: presumably chemical ability. However, if minor (and chance) alterations to questions internally or to their position in the test can cause statistically significant changes in facility value, it is clear that we are measuring something besides chemical ability.

Our recent research has been trying to discover what this "extra" is and to find ways of eliminating it. We hope that this can be the subject of a later paper.

Validity is clearly of prime importance. If the questions are measuring things other than what we want to measure, they are potentially poor sources of data. No amount of clever and even sophisticated statistical and computational treatment can make the outcomes any better. This is a classic case where "garbage in gives garbage out". An "industry" has grown out of selling banks of questions, mark-sense cards for student responses, mark-sense readers, computer programs for scoring the cards and whole computer packages for student self-assessment. It has all the attractive features of apparent efficiency, labour saving, quick return of results and objectivity. However, if the data-gathering instrument (the test) is invalid, this entire edifice sits on insecure foundations.

Many of the reservations, which have been raised in this paper about fixed response testing, could be levelled at other forms of testing also. It is time for taking a sober look at the strengths and weaknesses of all kinds of assessment and trying to match them to the different kinds of objectives of a course (Gibbs, 1995). It is certain that no one method of assessment is adequate for testing a course and that a battery of test methods is required to allow for a fair measure of our students' attainments (Balla & Boyle, 1994).

It is instructive to listen to students' views on testing. Some students, particularly those who tend to be surface learners (Entwistle, Thomson, & Tait, 1992; Entwistle & Entwistle, 1991), prefer

fixed response questions where there is a clear "right and wrong" and they regard open response questions as, "waffle, not knowing if you are answering the question". If we are trying to develop independent learners who can think for themselves, such students dislike fixed response questions because such questions give them no room to show what they can do and what their thinking is (Mackenzie, 1999).

Medical students undergoing a Problem Based Learning Course feel that fixed response testing is going counter to the spirit and purpose of the course.

Some researchers (Entwistle, Thomson, & Tait, 1992; Entwistle & Entwistle, 1991; Mackenzie, 1999) feel that fixed response testing encourages surface learning, because that is what is rewarded. If we wish to move students on, the assessment methods must be instrumental in promoting deep learning. The "sound bites" of fixed response may be reinforcing shallow, bitty, unconnected learning.

In the compass of this brief paper it is not possible to explore the huge field of assessment, but we hope that we have provided enough material to stimulate thought and debate among examiners and to inject a modicum of caution.

REFERENCES

Ace, M. & Dawis, R. (1973). Item structure as a determinant of item difficulty in verbal analogies. *Educational and Psychological Measurement, 33,* 143-149.

Balla, J. & Boyle, P. (1994). Assessment of student performance: a framework for improving practice. *Assessment and Evaluation in Higher Education., 19,* 17-28.

Black, P. (1998). *Testing: Friend or Foe? Theory and Practice of Assessment and Testing*. London: Falmer Press.

Bloom, B.S. (1956). *Taxonomy of Educational Objectives Handbook 1. Cognitive Domain*. London: Longmans.

Burton, R. & Miller, D. (2000). Why tests are not as simple as a, b or c. *The Times Higher* (4 Feb 2000)

Cassels, J.R.T. & Johnstone, A.H. (1978). What's in a word? *New Scientist* 78, 432.

Cassels, J.R.T. & Johnstone, A.H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education 61*, 613-615.

Entwistle, N. & Entwistle, A. (1991). Constructing forms of understanding for degree examinations: The student experience and its implications. *Higher Education, 22,* 205-227

Entwistle, N., Thompson, S., & Tait, H. (1992). *Guidelines for promoting effective learning in higher education*. University of Edinburgh.

Gibbs, G. (1995). *Assessing student centred courses*. The Oxford Centre for Staff Development.

Handy, J. & Johnstone, A.H. (1973). How students reason in objective tests *Education in Chemistry, 10,* 99.

Johnstone, A.H. & Ambusaidi, A. (2000). Unpublished work. University of Glasgow.

Mackenzie, A.M. (1999). *Prescription for change: Medical undergraduates perceptions of learning in traditional and problem-based courses.* Ph.D.thesis, University of Glasgow

Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education, 12,* 563-573