

A Reply to Marx*

Geoffrey P. Mason
university of victoria

It might be profitable if I attempt to answer briefly what I judge to be Marx' three main criticisms of my paper: (1) "misconstrues the role of the 'criterion' in criterion-referenced measurement," (2) "accepts too readily the nature of the cognitive processes presumed to underlie different types of test items," (3) "his tacit assumption that instructional feedback plays a role as a reinforcer."

In answer to the first criticism it should be recalled that the initial discussion took place within a framework of public school "core" and "extension of the core" curricula, and that there is little question that those responsible for a core curriculum, at least in British Columbia, view it in terms of mastery requirements for most students. Consequently, the use of "criterion referenced" which suggests "that judgements concerning an individual's achievement are dichotomous" would not be inconsistent with the context of the discussion. However, I believe that a rereading of the paper will show that the focus in this case was not on standards of achievement but on a "well-defined behavioral domain." Whether or not several performance standards are established with regard to the learning outcomes specified in a core curriculum is irrelevant to the thrust of the argument. I am simply suggesting that because a core curriculum usually defines rather carefully the knowledge and skills expected of the students, this logically requires a criterion-referenced interpretation of performance on test items which reflect as closely as possible those learning outcomes described in that core curriculum.

This brings me to Marx' second criticism that I "accept too readily the nature of the cognitive processes presumed to underlie different types of test items." While this is not completely true, I will state, without hesitation, that different types of test item, for example, completion versus multiple choice, frequently do require different cognitive processes. There are three ways to support this statement: by findings on item difficulty, by introspection, and by studies designed to measure this cognitive difference.

Consider first the simple case of the recall of a specific fact in a sentence-completion item versus the recognition of this same fact in a

*This reply was originally entitled "Marx' Angles on Learnin' or How to Keep Stalin'."

multiple-choice item. If the same cognitive process were involved in both instances, the two items would have precisely the same item difficulty for a given group of subjects. However, this will rarely be found to be the case except when the learning of that specific fact approaches mastery by all students. Naturally, if it can be assumed in advance that the fact in question has been mastered by all, there is no purpose for formative evaluation, and summative evaluation also becomes a somewhat empty exercise.

Let us now consider the cognitive processes which may come into play with these two types of item. Here I must report the introspective findings from a very small sample ($n = 1$). The sentence completion item measures the remembering of the specific fact with no cues other than those from the wording of the stem. The subject, presented with the item, recalls the correct answer if he can, or a wrong one, and writes it down. A similar situation also *may* obtain with a multiple-choice item. The subject may first recall the correct answer, and then identify it from among the choices given. This would be the approach of a subject confident of his knowledge of the correct answer. Note that if he follows this process but recalls a wrong answer it is quite possible that his answer will not be one of the choices given, so that he, in fact, may get a second chance at the item. However, this recall-identify sequence is not the one employed by those still in the process of learning and consequently unsure of their knowledge. Here the sequence surely approximates a search of the alternatives with a *recognition* of the correct answer; failing that, an elimination of apparently wrong answers in order to arrive at the right answer; and, failing that, as a last resort, a straight guess as to which of the presented answers is the correct one.

There are many day-to-day examples which demonstrate that recall and recognition are not necessarily the same process or even of equal difficulty. For example, it is not uncommon for those unable to recall people's names with ease to resort to all kinds of subterfuge in order to *recognize* the name of someone they are about to meet. A search of membership or guest lists, lists of names of those attending a function, notebooks, and even of place settings at banquets before the guests are seated provide useful stimulus material for recognition purposes. Recognition and recall are different processes and those for whom it is essential to remember names on social and business occasions work hard to maintain their knowledge at a recall level. A similar principle, I suggest, applies in school. If recall of information is that which is required, then recall and not recognition should be practised — and tested.

A similar analysis could be performed for the developing of a higher mental process skill. Let us consider the well-known test item using the map of the island of Serendip which shows its topography, rivers, products, cities, and railways. The test question, which is in multiple-choice

form, asks "Which of the following cities would be the best location for a steel mill?" (Educational Testing Service, 1963). With this particular item my sample, knowing little about economic geography and with the choices of answer covered up, chose a city not listed as a possibility. After the choices were uncovered the process became primarily one of eliminating the apparently most unlikely distractors. The first approach with the choices covered seems to involve a recall of principles governing the location of steel mills and then the application of those principles to a specific case. The second attempt, with the choices revealed, seemed to involve a more inductive process which attempted to derive the principles of location from the information given by the choices and the map, and then to confirm these principles by the apparent correctness of one of the choices.

The third support comes from research studies designed to investigate this specific issue. One should suffice here. A recent study from Educational Testing Service (Ward, Frederiksen, & Carlson, 1978) concludes "that there are potentially important differences between (1) assessments based on situations in which options are presented to the candidate for choice and (2) assessments based on situations in which candidates have to think of the options for themselves before reaching a solution to a problem" (p. 39). As the authors put it, in this study the free response and the multiple-choice forms cannot be considered to be parallel. It should be noted that the position taken in the original paper was not that the processes of recognition and recall are never the same, but that they are often sufficiently dissimilar to warrant a more careful matching of learning outcome and the test item than is frequently the case.

With his third criticism that "there is a tacit assumption that instructional feedback plays a role as a reinforcer," Marx attacks Skinner's position. My own view is that it is a little early to write off Skinner in the light of the evidence so far presented, but that, in any case, the interpretation of the mechanism by which increased retention as a result of feedback occurs is irrelevant to the general thrust of my argument. My position is in no way based on an acceptance of the principles of operant conditioning in this case but on the point that the behavior to be strengthened through feedback (whatever the mechanism involved) should be that behavior required by the learning outcome. The quotation by Marx from Kulhavy seems entirely consistent with this point of view.

Finally at the end of his paper Marx comes to what I believe is the real reason why recognition items are so frequently substituted for constructed type items when he states:

In my large undergraduate educational psychology class, it is frequently quite difficult to attain reasonable interscorer reliability on essay tests and papers, even when scoring criteria seem to be stated clearly. This frequently delays feedback on these assignments up to 10 days.

In other words Marx sees the marking of the responses required by the learning outcomes to be the problem and consequently attempts to justify the substitution of a more easily marked type of response for them. I have attempted to suggest throughout my paper that this is a serious error with the tail of assessment wagging the dog of instruction and learning.

The root of the above problem, I believe, lies in an incomplete acceptance or understanding of the difference between formative and summative evaluation in the classroom. Formative evaluations, being only for the purpose of advance learning, should not contribute to any final assessment. Consequently, high reliability is not essential when marking for instructional feedback. I would suggest that many test papers are marked by teachers and returned to students which could just as well have been marked by the students themselves using model answers if necessary. The only justification for a teacher's marking of all his students' papers for formative evaluation is that his students will indeed learn more than otherwise from this procedure. However, if the same time spent by a teacher in the marking of all his students' papers were spent reading a small sample for his own information, and then in a careful restructuring of lessons and materials in the light of what he had read, handsome dividends in the students' learning might well accrue.

The previous paragraph does not, of course, apply to a summative evaluation. As this is not a teaching device the problem Marx sees of obtaining immediacy of feedback is irrelevant. What is relevant to a summative evaluation is the determining, among other things, of whether or not the students have attained precisely those learning outcomes for which the curriculum was devised. That, in military parlance, is the point of the whole exercise.

REFERENCES

- Educational Testing Service. *Multiple-choice questions: A close look*. Princeton, N.J.: Author, 1963.
- Ward, W. C., Frederiksen, N., & Carlson, S. B. *Construct validity of free response and machine-scorable versions of a test of scientific thinking* (ETS RB 78-15). Princeton, N.J.: Educational Testing Service, 1978.