On Test Purpose and Item Type: A Comment on Mason

Ronald W. Marx
simon fraser university

Two major developments in the fields of measurement and evaluation took place in the sixties which have had profound effects on education in the last decade or so. The first development was the emergence of criterion-referenced measurement (Glaser, 1963), which has frequently been placed in opposition to the entrenched technology associated with norm-referenced measurement.

The concept of criterion-referenced measurement is certainly not new, as pointed out by Airasian and Madaus (1977), but since almost all educational measurement up to the sixties, and even most current usage, has been based on models developed in differential psychology, the criterion-referenced model is seen by most to be a new and rival conception of measurement. The conflict over these different models has been highlighted recently by the Popham-Ebel debate at the 1978 AERA meeting (Ebel, 1978; Popham, 1978), which Mason uses as the point of departure for his paper.

The second development upon which Mason bases much of his case is the now well-accepted distinction between formative and summative evaluation (Scriven, 1967). As Mason clearly points out, these developments are not independent, and for those of us who teach educational measurement and the many more who construct measures of learning outcomes, an integration of these two developments is welcome. The inclusion of a discussion of item type in this integration is doubly welcome, for, if successful, the synthesis of these three areas of measurement could provide a sort of conceptual roadmap for the construction and rationalization of measurement devices. Unfortunately, I think Mason's synthesis breaks down at a number of points.

ITEM TYPE AND CORE CURRICULA

Mason's position is based on the following arguments. Efficient criterion-referenced assessment of learning outcome requires well-defined task domains. In turn, well-defined task domains are possible with, for example, provincially mandated core curricula such as we have in British Columbia. But the task domains of learning objectives beyond those minimally established by the core curriculum are poorly described

by their very nature; thus it is difficult to develop high-quality criterion-referenced tests for this "extension of the core" material. As a result, it is possible to measure clearly the learning objectives of core curricula by comparing performance to an established criterion level, but the poorly defined task domains of "extension of core" objectives necessitate norm-referenced judgments since the only available standard of comparison is the peer group of the students to be judged. Furthermore, it is rare that core objectives will either be limited largely to recall-type tasks, or that such recall tasks will be representative of the kinds of transfer situations implicit in core objectives.

The result of all of this, according to Mason, is that constructed type test items are necessary for criterion-referenced assessment, presumably because of the inferences about covert behavior leading to the response (i.e., recall or recognition). I will return to this point shortly. On the other hand, "it is probably unnecessary to demand constructed-response type questions in order to demonstrate competence in all areas of the curriculum normally requiring recall type of behavior." This is true presumably because the judgments about "extension of the core" objectives are to be norm- rather than criterion-referenced.

I have two problems with this argument. First, I believe that Mason misconstrues the role of the "criterion" in criterion-referenced measurement. Second, I think that he accepts too readily the nature of the cognitive processes presumed to underlie different types of test items.

As Popham clearly points out in the debate, it is possible to attach two different meanings to the word "criterion" as it was used in Glaser's (1963) now classic paper. One meaning is that "a criterion-referenced test [is] one which relate[s] examinees' scores to a well-defined cut-off score or *level* of performance" (p. 6). The other, which Popham uses as his definition, is "that a criterion-referenced test is used to ascertain an individual's status with regard to a well-defined behavioral domain" (p. 6). Mason's use of "criterion" is closer to the first than the second meaning.

Because of this use of the term, Mason's arguments lead one to believe that judgments concerning an individual's achievement are dichotomous; learning has or has not occurred at some predetermined level of mastery. Gagné and Beard (1978) help clarify the issue by referring to *quality* (how well) and *quantity* (how much) judgments in criterion-referenced measurement. For example, as they suggest,

> answering eight out of ten questions about the Boston Massacre is often assumed to yield the same kind of assessment as answering eight out of 10 whole number multiplication problems. Actually, two different dimensions of the task domain are involved. A score of 80% on items pertaining to the Boston Massacre reveals something about *how much* is known — that is, how many relevant propositions have been stored in the learner's memory. The same percentage score applied to multiplication problems is intended to

tell us *how well* the intellectual skill of multiplying whole numbers has been learned. To interpret the latter as "how much" would be to assume that the product of each combination of whole numbers has been separately memorized by the student — an assumption that cannot seriously be entertained. (p. 267)

It might make sense to render a dichotomous judgment regarding quality of performance if the task domain entails an intellectual skill which is unitary in nature (e.g., word problems in math in which the task is to calculate required travel time between two points given distance and rates of travel, and the answer is an integer). But if a quantity judgment is to be made, then the inference the educator wants to make is, say, that the student can perform 60% of the problems in the task domain, as estimated by performance on a sample of items from the task domain included in a test. As Mason implies, quantity judgments may be converted to dichotomous judgments in the case of mastery learning, when a decision has to be made about moving to the next objective. Generally the distinction between quality and quantity judgments should help clear up the muddy water surrounding the "criterion" issue in criterion-referenced measurement by making more explicit the nature of performance implied by the task domain.

The second criticism in this area has to do with the presumed cognitive processes involved in responding to various types of test items. Mason appears to accept at face value that constructed type items require different cognitive responses than do supplied type test items. Clearly the overt behavior of the student is different, but inferring from this that the cognitive processes are necessarily different is yet another step in logic, even though it is the standard practice of authors of measurement texts to make this logical connection between overt behavior and cognitive process without clear empirical support.

There are three aspects of what Gagné and Beard (1978) call directness of measurement. First there is the *stimulus situation*, usually referred to as the context in a behavioral objective. Second is the *response provision*, which is the observed behavior component of the objective. The third aspect of directness of measurement is the *process*, or the "major verb" of the objective. While the stimulus situation and response provision relate to content validity, a determination of process is clearly a construct validity problem (see Cronbach, 1971). The answers to construct validity problems are usually slow in coming, and require both experimental and correlational studies of the type usually associated with construct validity studies of norm-referenced measurements (see Graham, 1974, for an example of such a study in the context of the measurement of learner outcome). However, until adequate empirical support is provided for the construct validity of the process component of a behavioral objective, and the corresponding process component of a test item,

judgments regarding this process ought not to be glibly made by appealling only to the topography of behavior.

It might be claimed that simple common sense would dictate that an educator would choose test items requiring behavior which would correspond with the intended behavioral outcome of instruction. While this might frequently occur, it is not necessarily true for all instructional outcomes. As for common sense, I defer to Skinner (1974):

> The disastrous results of common sense in the management of human behavior are evident in every walk of life, from international affairs to the care of a baby, and we shall continue to be inept in all these fields until a scientific analysis clarifies the advantages of a more effective technology. It will then be obvious that the results are due to more than common sense. (p. 234)

FORMATIVE AND SUMMATIVE EVALUATION

Scriven's (1967) distinction between formative and summative evaluation was made in the context of curriculum evaluation, where the unit of analysis is an instructional program. This distinction, between evaluation as a source of data for corrective feedback and evaluation as a source of data for summary judgment, has often been applied to contexts in which a person is the unit of analysis, such as in clinical supervision or in classroom learning. Mason uses this second meaning of the distinction.

Mason's position regarding these issues stems from his tacit assumption that instructional feedback plays a role as a reinforcer. This is more clearly obvious when he states "that a student needs to *produce* a response and have that response behavior reinforced, and that wrong responses in a set of alternatives are out of place in the process of 'shaping' behavior because of the tendency to strengthen unwanted forms." These are a number of problems in this position.

First, when arguing for constructed type items over selected type items based on this rationale, it is assumed that more incorrect responses will be made to selected rather than constructed type items. I do not see why this need be so. Certainly if essay type questions are used as constructed items, this assumption is obviously not tenable, as any of us who have marked essay exams can attest. Yet even if one were to grant this assumption, the role of feedback in instruction (Mason's use of formative evaluation) goes beyond that of a reinforcer.

Kulhavy's (1977) review of the role of feedback in learning from text (taking an exam?) helps clarify the distinction. In an operant analysis of behavior, a reinforcer is defined as a stimulus occurring consequent to a behavior which serves to increase the probability that the behavior will occur again. Simple-minded application of this concept has resulted in the proposition that praise or recognition following a correct response will increase the probability of that correct response occurring again. Yet

the functional quality so necessary for a stimulus to act as a reinforcer is missing here. Indeed, as most texts on applied behavior analysis in education state (e.g, Herman, 1977), it is impossible to define a priori a stimulus as a reinforcing stimulus. So while praise for correct responses *may* be functionally related to behavior, this cannot be accepted universally before the fact, making Mason's claim about the nature of feedback as reinforcing (or punishing) premature with respect to designing both effective tests and effective curricula.

Clearly, feedback operates in ways other than as a reinforcer. From a cognitive perspective, feedback should provide information regarding the qualities of a response. To this end, Kulhavy (1977) concludes by stating that feedback

> confirms correct responses, telling the student how well the content is being understood, and it identifies and corrects errors — or allows the learner to correct them. This correction function is probably the most important aspect of feedback, and if one were given the choice, feedback following wrong responses probably has the greatest positive effect. (p. 229)

It seems, then, that the formative evaluational role played by feedback from tests is not so clearly related to item type as claimed by Mason. Certainly, the reinforcing role of feedback requires empirical demonstration as opposed to a priori speculation, and the informational role is not clearly tied to item type.

In my own teaching, I have found that the technical problems associated with constructed type items are difficult to overcome. In my large undergraduate educational psychology class, it is frequently quite difficult to attain reasonable inter-scorer 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. This delay of feedback is counter-productive to the immediacy of feedback claimed to be important by Mason and others, even though a number of studies have shown that delayed feedback enhances learning in some situations (Kulhavy, 1977; Sturgis, 1972).

CONCLUSION

Mason wrestles with some important issues in his paper. As I have argued elsewhere (Marx & Winne, 1979), all too often a test item type is selected on the basis of ease of scoring, rather than sound pedagogy. This has resulted in a proliferation of the multiple choice test as the preferred technology in far too many contexts. To this end, I share with Mason a concern for sound decision making in the selection and development of evaluation devices. Just as curriculum planners and instructional designers argue for rational procedures for the creation of instructional programs, measurement specialists argue for a conceptual foundation for evalua-

tion procedures. To this end, item type should be related more specifically to task domains, including the process components of objectives, and not simply to the referencing procedure for tests or their formative or summative role.

NOTE

Thanks are due Jack Martin and Phil Winne for their comments on a draft of this paper.

REFERENCES

Airasian, P. W., & Madaus, G. F. Criterion-referenced testing in the classroom. In V. R. Martuza (Ed.), *Applying norm-referenced and criterion-referenced measurement in education.* Boston: Allyn and Bacon, Inc., 1977.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement.* Washington, D.C.: American Council on Education, 1971.

Ebel, R. L. The case for norm-referenced measurements. *Educational Researcher,* 1978, 7 (11), 3-5.

Gagné, R. M., & Beard, J. G. Assessment of learning outcome. In R. Glasser (Ed.), *Advances in instructional psychology* (Vol. 1). Hillsdale, N.J.: Lawrence Erlbaum Assoc., Publishers, 1978.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, *18,* 519-521.

Graham, D. L. *An empirical investigation of the application of criterion-referenced measurement to survey achievement testing.* Unpublished doctoral dissertation, Florida State University, 1974.

Herman, T. M. *Creating learning environments.* Boston: Allyn and Bacon, Inc., 1977.

Kulhavy, R. W. Feedback in written instruction. *Review of Educational Research,* 1977, *47,* 211-232.

Marx, R. W., & Winne, P. H. Research and evaluation in self-education. In M. Gibbons, G. Phillips, & J. W. G. Ivany (Eds.), *Self-education,* in preparation.

Popham, W. J. The case for criterion-referenced measurements. *Educational Researcher,* 1978, 7 (11), 6-10.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation.* AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Skinner, B. F. *About behaviorism.* New York: Alfred A. Knopf, 1974.

Sturgis, P. T. Information delay and retention: Effect of information in feedback and tests. *Journal of Educational Psychology,* 1972, *63,* 32-43.

Ronald W. Marx is an Assistant Professor in the Faculty of Education at Simon Fraser University, Burnaby, British Columbia V5A 1S6.