

Test Purpose and Item Type

Geoffrey P. Mason
university of victoria

Cet article considère le problème du type d'éléments les plus appropriés aux énoncés d'objectifs éducationnels des "programmes-tronc" (core curricula) et des "programmes-tronc élargis" (extension of the core). On suggère que puisque la réussite du "programme-tronc" sera inévitablement mesurée à partir de critères référentiels, des éléments du genre réponse construite seront normalement requis. Des éléments du type réponse sélective, tout particulièrement à choix multiple, pourraient être utiles à l'évaluation cumulative d'au moins une partie du "tronc-élargi," puisque les mesures dépendront probablement en ce cas de normes référentielles. Cependant, en matière d'évaluation formative destinée à guider l'apprentissage, les éléments d'examen doivent absolument refléter les réponses que l'on désire voir apprises, de telle sorte qu'on utilisera généralement des éléments du type réponse construite.

The recent Popham-Ebel (1978) debate on criterion versus norm-referenced measurement has tended to point up some of the less clear patches in the field of measurement and evaluation. Central to many of the current problems, it seems, lie the issues of test purpose and of item type. This paper attempts to examine the two basic types of item in relation to currently emphasized core curricula, criterion versus norm referenced measurement, and formative and summative evaluation.

In essence, Popham argued in the debate that most educators are primarily concerned with populations of knowledges, understandings, and skills which are susceptible to fairly precise description. In that the elements of a given population are known, a sample of them can be obtained in some systematic manner for estimating the level of a student's achievement with respect to that population. Further, as comparisons among students are largely irrelevant to the educational process, evaluation of achievement should be only by direct reference to an established criterion level of performance on the sample of items selected.

Ebel presented the contrasting view that in general for most educators the populations of important knowledges and skills and their interrelationships for the courses they teach are almost infinitely great, so that these populations defy complete description both qualitatively and quantitatively. As the elements of these populations cannot be completely described, systematic sampling is not possible and, this being the case, the

only appropriate basis in education for the evaluation of most student achievement is a norm referenced one, i.e., by comparison with the performance of others on the same tasks.

It was perhaps unfortunate that, for the purpose of debate, these two positions were presented as antagonistic ones, whereas, in fact, for most instructors at any level, they may well be complementary. The recent and widespread "back to the basics" movement has resulted in many teachers again being held responsible for the teaching of a rather carefully defined body of knowledge and skills frequently referred to in the past as "minimum essentials." An example of such an elaboration of basic knowledge and skills to be acquired by students has recently re-emerged in one province after a lengthy absence as a statement by the Ministry of Education (1977) of the core curricula for the public schools of British Columbia. In that the population of knowledges and skills of a core is usually well defined, and the intent of the specification of a core is the achievement of a high level of performance on it by nearly all students, the criterion-referenced approach to the measurement of student achievement of the core is clearly the appropriate one to use.

Although the specification of a core of work to be achieved is now contained in many, if not most, statements of public school curricula in Canada, the mastery of the core alone is rarely held to be a sufficient educational accomplishment for better students. Considerations of individual differences lead to different expectations in the quality and quantity of work to be completed. In those cases where the subject matter of a given course can be strictly sequenced, better students frequently are led to mastering more units than their slower colleagues. In those other cases where the subject matter is not susceptible to the strict sequencing necessary for the mastery approach, a statement expressing the need for an extension of the core or enrichment for better students generally accompanies the definition of the core. Statements of this enrichment usually reflect the approach of Ebel in viewing the number of relationships incurred by such extension as very large and consequently not susceptible to precise prescription. In that the extension of the core customarily lacks the definition required for criterion-referenced measurement, norm-referenced measurement of student achievement appears both necessary and appropriate. The writer and his colleagues, in fact, have for several years established courses with a core of behaviorally stated objectives, the mastery of which obtains a grade of C or Pass. For extension beyond this core of work, norm-referenced measures are used for awarding A and B grades on the grounds that the extension work involves a quality of performance not easily defined, and that the meaning assigned to the higher grades by the university community is essentially a normative one.

ITEM TYPE AND THE CORE CURRICULUM

There are two basic types of test question: items that require a response to be *constructed*, e.g., sentence completion, paragraph, and essay items; and items that require the correct answer to be *selected*, e.g., multiple choice, matching and true-false items. The type of item appropriate for tests at the end of a unit of instruction on both the core and the extension of the core deserves some consideration.

If the educational outcomes for the core curriculum have been adequately stated in measurable terms there should be little doubt as to the appropriate type of item to be used. The item type will be dictated by the behavior required in each outcome statement. For many courses this leads to tests containing a variety of item types or tasks which are usually of the constructed-response type requiring the student to remember information, state understandings, or “think” by applying principles. This is largely because core curricula tend to be anchored in essential “real-life” tasks, and essential tasks which require a person to identify a correct answer when supplied, rather than to supply the answer itself, are relatively few. Naturally, it makes little sense to demand a high performance on a specified core of fundamental tasks, and then attempt to measure this performance by test items not directly reflecting these tasks, although this is by no means an uncommon practice.

It has been previously indicated that the objectives for the extension of the core work are invariably far less precisely stated than the outcomes of the core. In that these objectives are not readily translatable into a form amenable to criterion-referenced measurement, and in that the purpose of the extension work may be viewed as providing a challenge for better students, a norm-referenced approach may be appropriate for the measurement of the extension work.

While it should be evident that, as in the case of the core, the test items used must be related to the objectives of the instruction, nevertheless 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. Some years ago Educational Testing Service (1963) made a strong case for the use of multiple-choice questions under a variety of circumstances. The technical superiority of multiple-choice items, together with the high reliability and ease of marking them and their ability to achieve the maximum spread of marks desirable for those wishing to rank students, makes them a particularly useful tool for the measurement of at least the “knowledge” and “comprehension” levels of the extension of the core work.

A frequently used justification for the use of multiple-choice items in the place of constructed-response items is that the correlation between test results obtained by the use of the two item types is very high, so that one type may be conveniently substituted for the other. However, it has

long been known that the level of remembering when measured by recognition tasks is far higher than when measured by recall tasks. This difference in the absolute level of performance is irrelevant to the task of simply ranking students. It is, however, an important reason why selected-response type items cannot be substituted legitimately for constructed-response type items in criterion-referenced measurement where a given standard of recall is required.

FORMATIVE AND SUMMATIVE EVALUATION

Scriven (1967) made a valuable distinction between the gathering of data during the time a program is being developed for the purpose of guiding the developmental process and the gathering of data to make an overall assessment or decision with regard to the program. These processes Scriven called respectively formative and summative evaluation. Bloom (1971) extended these concepts to apply to the evaluation of student learning. Formative measures may include class quizzes, homework assignments, scheduled tests, and classroom questioning when the intent is to guide the students' learning. Summative measures occur at the end of an instructional period and are directed at the intent of the instruction as specified by the learning objectives.

In the last section it was pointed out that it is frequently argued that the high correlation between achievement measured by selected-response and by constructed-response items, and the greater ease in marking and technical superiority of objective items, particularly multiple choice, justify their use. While there may be some validity to the argument when the purposes of a test are both to evaluate the work of students at the end of a unit of instruction and to rank the students, the argument does not hold either when the purpose of the test is the feedback of information to each student to assist learning, or the evaluation is to be criterion-referenced.

There is a much-quoted phrase in certain circles in education that "every test should be a learning experience." As with many other educational maxims there is a grain of truth in the statement. It would, in fact, be difficult to make any test "learning free" although what might be learned will differ from individual to individual taking the test (e.g., frustration, dislike of the teacher, the inappropriateness of staying out late the night before an exam, and so on). While the potential for constructive learning inherent in most testing situations should be fully utilized, it should be recognized that, in all but the mastery learning model of instruction, the "learning experience" with much material in school must eventually be terminated, the amount learned summatively evaluated, and the next part of the work begun. Nevertheless, it is certainly true that the purpose of most tests in educational institutions is *formative* and that good educational practice requires that the feedback of results to students should be appropriate and as immediate as possible.

Several years ago, Skinner (1961) addressed the problem of the nature of measurement and feedback during the learning process. In his paper he states that it is essential for the student to *compose* a response, rather than recognize it from a set of alternatives such as in a multiple choice item, on the grounds 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. Skinner explains:

Those who have written multiple choice tests know how much time, energy and ingenuity are needed to construct plausible wrong answers. They must be plausible or the test will be of little value. In a multiple choice *test* [author’s note: summative evaluation?] they may do no harm, since a student who has already learned the right answer may reject the wrong answers with ease and possibly no undesirable side-effects. The student who is *learning* [author’s note: formative evaluation?], however, can scarcely avoid trouble. Traces of erroneous responses survive in spite of the correction of errors or the confirmation of the right answer. (pp. 393-394)

Skinner goes on to point out that the selection of a wrong answer and the subsequent learning that the selection was wrong does not mean that this response will never occur again, for the student “may later find himself recalling that somewhere he has read that . . . [the faulty answer], and of course, somewhere he has” (p. 394).

It must be admitted that there are certain learning tasks which demand recognition: the rapid identification of the shapes of different aircraft in military training and the identification of works by different artists where the style is too complex to be described adequately in words are examples of these. However, for the *formative* evaluation of the vast majority of educational tasks, the technical superiority of selected-response type items is far outweighed by the learning considerations expounded by Skinner, so that constructed-response type items should inevitably be used.

CONCLUSION

Because of the apparent favoring of subjective over objective type items in this paper, it may appear to some as an advocacy of a regression to the “pre-scientific” era of educational measurement. What in fact is being argued is that recent developments in the field are now enabling much greater precision in the statements of instructional objectives by educators than was previously the case. Content validity requirements for the measurement of educational achievement and learning considerations during the instructional process both require that the test items used reflect the behavior required by the learning outcomes. In that most outcomes require a recall rather than a recognition type of behavior, constructed-response type items will generally be required in both formative evaluation and in criterion-referenced summative evaluation.

REFERENCES

- Bloom, B. S., Hastings, J. T., & Madaus, G. F. *Handbook of formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.
- Educational Testing Service. *Multiple choice questions — a close look*. Princeton, N.J.: Author, 1963.
- Ministry of Education, Province of British Columbia, *Guide to the Core Curriculum*. Victoria: Author, 1977.
- Popham, W. J., & Ebel, R. L. *The relative merits of norm-referenced measurement and criterion referenced measurement for purposes of educational evaluation and instruction*. Paper presented at annual meeting of the American Educational Research Association, Toronto, March 1978.
- Skinner, B. F. Why we need teaching machines. *Harvard Educational Review*, 1961, 31, 377-398.
- 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.
-

Geoffrey P. Mason is a Professor in the Department of Psychological Foundations of the Faculty of Education at the University of Victoria, Victoria, British Columbia V8W 2Y2.