DOCUMENT RESUME

ED 340 768 TM 018 024

AUTHOR	Scriven, Michael
TITLE	Multiple-Rating Items.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
PUB DATE	5 Dec 91
NOTE	14p.; Commissioned paper prepared for a workshop on Assessing Higher Order Thinking & Communication Skills in College Graduates (Washington, DC, November 17-19, 1991), in support of National Education Goal V, Objective 5. For other workshop papers, see TM 018 009-023.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Educational Assessment; Essay Tests; Higher Education; Multiple Choice Tests; Occupational Tests; *Rating Scales; Scoring; *Student Evaluation; Test Construction; *Testing Problems; Test Items
IDENTIFIERS	*Multiple Rating Items

ABSTRACT

An alternative to multiple-choice testing is suggested for educational assessment. The use of what is called "multiple-rating items" is proposed. A multiple-rating item calls for the examinee to rate all of a set of things instead of picking one as with a multiple-choice item. The respondent has to provide a specific rating of each entry. Such items make it possible to retain the advantages of multiple-choice items (simplicity and reliability of scoring and comprehensive coverage), while avoiding most of their weaknesses. Multiple-rating items are faster to compose than good multiple-choice items and as fast as essay items if the time for composing a scoring sheet is included. They are easy to mark and can be machine-scored. Multiple-rating items do require that one learn how to construct and mark them; they are still a something of a novelty. Advantages include exceptional suitability for job-relevant training of teachers, researchers, and many other professionals; testing trans-curriculum skills; and moving students to the metacognitive level of learning how to evaluate their own work. It is suggested that most choice items beyond tests of rote knowledge and simple discrimination skills should be converted into or replaced by multiple-rating items. (SLD)

Multiple-Rating Items

Michael Scriven
Pacific Graduate School of Psychology[1]

OVERVIEW

The multiple-choice test provides an unfortunate example of paradigm paralysis. It is a highly developed technology within its paradigm limitations, but those limitations are extremely serious, especially in testing the key transcurriculum skills such as critical thinking, communication, and problem-solving. In the search for "more authentic" assessment, the alternatives preferred over multiple-choice tests are usually (i) structured observations or (ii) tests calling for constructed responses. The work evaluated in these approaches includes essays, projects, simulations, interviews, performances, and portfolios; on the borderline are tests calling for very brief structured responses—short-answer and sentence-completion tests. While constructed-response approaches undeniably have a place in serious assessment, they often run into well-known problems with (i) scoring—particularly problems of cost, time delays, and the validity of the scoring keys—and (ii) limited coverage of the curriculum because of the increased time per item. These problems often outweigh the validity gains from increased task relevance, absence of test anxiety, and the exclusion of success by guessing—and the advantage of time savings at the test-construction end. Moving to these alternatives at times seems like jumping out of the frying pan of a too-crude approach into the fire of a too-complex one.

In this paper, an intermediate alternative is developed—the use of what are here called 'multiple-rating items'. They make it possible to retain the advantages of multiple-choice items—simplicity and reliability of scoring, and comprehensive coverage—while avoiding most of their weaknesses. (They also have subsidiary advantages over multiple-choice items in savings of paper, printing, and mailing costs, especially for large-scale testing.) They tap the highest levels of the Bloom taxonomy, are faster to compose than good multiple-choice items and as fast as essay items if the time required for composing a scoring key is included. Typical items are about as fast to mark as multiple-choice items for a human but they can easily be machine-scored[2]. Other educational advantages of multiple-rating items include exceptional suit-

[1] Mailing address: POBox 69, Point Reyes, CA94956. Criticism, whether constructive or not, is much appreciated. Thanks to John McPeck for valuable comments on an earlier draft.

[2] We will refer to tests which can be scored without the use of judgment or subject-matter knowledge as 'template-scoreable' or 't-scoreable' rather than 'objective'. They can be scored very quickly by the instructor (who compares the testee's answers with a template—a scoring key constructed to overlay or match the spacing of the

ability for: job-relevant training of teachers, researchers, and many other professionals whose jobs frequently involves rating the work of others; for testing trans-curriculum skills; and for moving students to the metacognitive level of learning how to evaluate their own work, usually a path to improved performance in any subject matter. Multiple-rating items will sometimes serve as the best sole option for a short test, will often provide a better element in a mix with constructed-response items than multiple-choice items, and should always be considered as a third element in a mix of the other two types.

This all sounds too good to be true. What are the limitations and drawbacks of multiple-rating items? There are three candidates. (i) Methodologically, multiple-rating items are not as good as multiple-choice items for testing certain discrimination skills and no better for testing rote knowledge. (ii) Philosophically, they call for the testee and the test designer to make more complex and explicit evaluations than are involved in multiple-choice items, and for someone committed to the dogma of value-free social science that is a serious drawback. But since every teacher has to make the same kind of evaluation every time they mark a student assignment, and good teachers have long known how to do that well, this objection can hardly be treated seriously; it's simply a relic of imperial positivism. (iii) Practically, the problem is that multiple-rating items are something of a novelty for those used to constructing (or doing) multiple-choice items, and one has to learn how to construct and mark them, and cannot get prepackaged ones at the local store. It's always a little more difficult to do things oneself, especially new things, than to stay with the old and familiar, but when you get down to work with multiple-rating items, it turns out they are not so unfamiliar—indeed, they look like old friends released from Siberia and spruced up a little.

THE MULTIPLE-CHOICE TEST AS A PARADIGM

It is important to understand the nature of the issue here. For many people in testing, this is not just a matter of choices between technical devices, systematically evaluated in terms of their functionality. Whether they realize it or not, their behavior makes clear it is a fight over a paradigm. The general account provided here shows that multiple-choice items are simply a very primitive limit case of multiple-rating items. Fixation on them involved arresting the whole development of testing technology. It involved ignoring a number of obvious ways to generalize and strengthen multiple-choice items on purely ideological grounds. Some of these

questions in the test), or by a clerk, or by the most primitive kind of personal computer with a simple program and an appropriate mark-sensing input device. Fixation on standard multiple-choice items has protracted the sale of the overpriced standard scanning equipment and inhibited the development of simple, cheap, and more powerful alternatives, to which there are no technical obstacles but still too small a market to generate development of a competitive and creative technology (although scanners have been available at a high price since the mid-70s). With limited resources, the best approach to extended multiple-rating item scoring is probably through the now-cheap bar code readers and a hypercard-type program.

more general approaches have long been employed successfully by teachers—for example, differential weighting of different questions—and none of them involved a significant threat to the advantages of the multiple-choice item. They were dismissed either as involving unnecessary complications, or as involving what were described as arbitrary decisions. Since the decisions were evaluative ones, this was particularly unattractive to followers of the dominant value-free doctrine. However, the alternative approach associated with typical multiple-choice tests often involves unnecessary oversimplification and evaluative decisions that are much less plausible. Dedication to the multiple-choice item in this situation tells us that more is going on here than simple disagreements about technical properties.

The language of defenders of multiple-choice items involves referring to them as "objective tests", usually with the implication that they are the paradigm type of objective test[3]. Now it is an elementary error to call collections of multiple-choice items or multiple-rating items "objective tests" since the decision as to what is to count as the right answer, built into the scoring key, is often highly subjective and not infrequently wrong in even the best-known commercial tests. Obviously, tests fail to be objective as much or more through errors at the front end (*constructing* the scoring key) as at the back end (*applying* the scoring key). This labeling simply constitutes disregard for 'truth in labeling' by the testing profession, and is roughly the equivalent of the Society of Automotive Engineers defining the term 'safe' as a synonym for 'cars over 4000 lbs. dry weight'. Perhaps misleading labeling at this level of generality is felt to be acceptable because it is a matter of philosophy or value judgments rather than science or professional ethics—a distinction usually made by those about to commit some philosophical or evaluative atrocity. Whatever the reason, the occurrence of this kind of misrepresentation in a field which is closely governed by meticulous professional standards for test construction, advertisement, and use[4] is one of the signs that we are here dealing with paradigm worship, since one of the leading indicators of paradigm control is the covert appropriation of honorific vocabularies for narrow conceptions.

In any case, the supposed superior objectivity of the multiple-choice item and its poor relations—the true-false item, and the matching item—over the constructed-response test item is shared by all multiple-rating items. The fact that testing professionals omit any reference to

[3] In a typical leading text, the section entitled "Suggestions for Writing Objective Items" only discusses multiple-choice items, true-false items, and matching items; true-false are dismissed as unsound (quotation below) and matching items are described as a series of multiple-choice items. (The section starts on page 210 of the Fifth Edition of ***Measurement and Evaluation in Psychology and Education***, (Macmillan, 1991) in which Thorndike and Hagen, the authors of the first four editions of what is probably the most widely-used text on testing and measurement, are joined by Cunningham and another Thorndike.)

[4] ***Standards for Educational and Psychological Tests***, joint product of the American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education, published by the APA, 1985.

this huge class of items that are equally 'objective', i.e., template-scoreable, is another sign of deep prejudice. The simple fact is that there is no provable superiority of *any* of these types of item in respect of their true—as opposed to redefined—objectivity, let alone their validity, by comparison with properly designed and scored constructed-response items.

Study of the leading references on testing reveals other symptoms of the same disease. For example: (i) systematic treatment of the obvious generalizations of the multiple-choice item paradigm is rarely provided; (ii) possible modifications are rejected with the semi-religious fervor accorded to attacks on paradigms; (iii) the explicit definition is much wider than the implicit definition—a typical device for claiming ownership of a domain most of which is not discussed. The paradigm status of the multiple-choice item can only be understood fully in terms of the zeitgeist of behaviorism in psychology, which led to minimizing the use of evaluation at the expense of utility. Thus we got 'grading on the curve' and 'objective testing', both of them involving massive ideological commitments which were disguised so well that their supporters thought they had no such commitments.

These remarks constitute a fairly serious indictment of many—although by no means all—of the leading members of a professional specialty[5]. They occur at the beginning of this essay because in their absence the non-specialist reader will find it hard to believe that the paraphrases and quotations used to represent the defense of the multiple-choice test are typical. Even the best defenses in th's matter are poor defenses because the issue has transcended the realm of normal reasoned argument and become a matter of attacks upon and defenses of something transrational—a paradigm. It is, however, a paradigm which never deserved that status, and the time has come to move beyond treating it as any more than one convenient tool in the testing toolkit. We begin by describing the more powerful alternative.

MULTIPLE-RATING ITEMS vs. MULTIPLE-CHOICE ITEMS

A multiple-rating item calls for the testee to *rate all* of a set of things instead of *picking one* of them as with a multiple-choice item. The rating vocabulary to be used by the respondent to a multiple-rating item (often referred to here as a Rating item) varies between tests and possibly between items on one test. While it is limited, and prescribed for each test or item—which keeps the scoring simple—it is much richer than the single check mark used in the "pick the best" task of the multiple-choice item (here called a Choice item). Three examples of basic Rating items, and two that exhibit minor extensions follow. (i) Providing a 150-word passage

[5] Special reference should be made to the care with which some leading members of the profession address these matters. In particular, Robert Ebel in many writings, and Jason Millman/Jennifer Greene in their definitive essay "The Specification and Development of Tests of Achievement and Ability" in *Educational Measurement (3rd edition)* (Macmillan, 1989) discuss these matters with a skill that avoids virtually all of the criticisms mentioned here—although they do not explore or discuss the territory of multiple-rating items.

of prose followed by several alleged 20-word summaries of it, each of which is to be rated as Excellent, Acceptable, Poor, or Very Poor; (ii) providing an argument, followed by several proposed criticisms of it, each to be rated on the same or on a simpler or more complex scale (e.g., one which includes ratings of relevance); (iii) reproducing an advertisement or television commercial, followed by several impressions it is said to convey, each to be rated for their plausibility or, alternatively, for their probability of acceptance by the target audience; (iv) stating a problem for investigation, followed by descriptions of several processes that are to be rated as appropriate or inappropriate parts of the investigation, perhaps on the scale Essential, Desirable, Marginally Useful, Inappropriate, Unnecessarily Expensive, Highly Efficient (related examples would call for *ranking* designs for inquiry, or rating proposed *solutions* to the problem). In all these examples, any of the things rated may receive any of the available ratings, whether or not they have been used for any other; and in the last example more than one of the ratings may be applied to any one of the entries. (v) A different type of example of complex multiple-rating item involves two parts, the first part being just like one of the above—example (ii) above, for instance—and the second part providing a number of potential reasons for the judgement given in the first part, each of which is to be rated for the extent of its support of the answer to the first part, from Irrelevant, through Relevant but Weak, to Strong. Since the set of reasons given in the second part may omit some or all crucial ones, it is not a checklist substitute for thinking out the answer to the first part on its own merits.

It can be seen from these examples that the multiple-choice item and the multiple-rating item can look similar on the page although the appearance is deceptive. The following description is intended to clarify the similarities and differences and to introduce a simple terminology for the components of multiple-rating items. After item instructions (which can be provided at the beginning of the test when all items are of the same kind), both Choice and Rating items present what is conventionally called a stem, frequently a paragraph of prose, which defines a subject matter for what follows. This is followed in a Choice item by a numbered set of possible answers, the incorrect ones being called distractors (or foils). These can be described as 'options' in a Choice item because any one of them can be picked, but that language is inappropriate for Rating items, since there is no choice to be made *between* the items. We will refer to them as 'entries', sometimes using the term in a general way so that an option is a special case of an entry. As in Choice items, these are typically text clauses or sentences. (The stem and even the entries might also be pictures, music, recordings of a discussion, dances, maps or diagrams.)

In Rating items each entry requires a separate response; in Choice items, it is only required that the testee respond to one of them. In the Rating item, there is a small set of possible responses; in the Choice item, the only allowable response is selection of one entry (e.g., by checking a box or inserting the number of that entry in a box). With Choice items, writers often refer to the entries as responses, an elision for the fact that *picking one of them* constitutes a response. With Rating items, it is important to be clear about the distinction between the entry (which is part of the item on the test) and the response, which is made by the testee. The testee never selects an entry, only one of the set of responses to the entry. Only in the limit case of the Choice item does the distinction collapse, because at that point, selecting one of the entries *is* the response.

In a Choice item, the set of entries is 'guaranteed' to include one and only one correct response and several incorrect responses (distractors); in multiple-rating items, all or none of the entries may be flawless, or worthless, or at the same or different points in between. In simple cases, the Rating item might require a letter grade from A to F for each of four entries (or require that each be scored from 0 to 10, ranked from 1 to N, etc.). In such cases, we can readily arrange machine-scoreability without optical character reading capability, by designing the form with six or ten boxes listed after each entry instead of just one. In this format, the instructions for a Rating item thus have an extra dimension: instead of asking the respondent to check one box from a single *vertical column* of boxes (one box being associated with each of the entries), as in the Choice item, the instructions require checking one box from each of the *horizontal row of boxes* associated with each entry.

Thus the main difference between Rating and Choice items relates to the responses available to and required of the testee. Although both types of item require the respondent to evaluate several entries in a way that can itself be scored as right or wrong, better or worse—unlike a projective test or most of those calling for a constructed response—the respondents in a Choice item have an extremely limited evaluative vocabulary. They can only impose a *partial ranking* on the set of options, which they do by ranking one of them as better than the rest. The act of checking it rates the others as less good, but not necessarily as different in merit. The respondent to a Rating item, on the other hand, has to provide a specific *rating* of *each* entry, a rating which will usually be from the much more extensive evaluative vocabularies of *grading, scoring,* or *full ranking*—although it may also come from a listing of detailed evaluative comments from a limited set that is provided.

It will be obvious that the Rating test will usually involve more time per question and hence less paper per test—one cannot do a quick scan to spot the best (or least bad) option, make a single check mark, and move on to the next question. It should also be clear that the Choice item is just a very limited special case of the Rating approach. The key question is of course whether moving to the more general case, which clearly does not lose the fast and reliable scoreability of the Choice item, does lose something else, perhaps validity in the scoring key. We'll turn to that question in a minute.

It is important to note that the grading (or scoring etc.) involved in answering Rating items is not at all the same as having the testee *estimate the likelihood that each of their answers is correct*, a procedure which has been extensively, although—as it turned out—fruitlessly discussed in the testing literature as a possible development of multiple-choice items. This point is stressed because on several occasions testing professionals have reacted to the concept of multiple-rating items with the comment that they are equivalent to rating the probability of the response, and that research has already shown this to be unpromising. On the other hand, *estimating the likelihood of the assertions* in the entries, something quite different, can form the basis for one legitimate type of Rating item although such estimates are neither called for nor involved in most Rating items[6].

It might also be noted that the relatively simple examples given above can usefully be enhanced in two respects for certain purposes. First, the response repertoire for Rating items can easily be run up to the 'academic grade set' of 13-15 modified letter grades including plus and minus modifiers (e.g., A+ to F–), as would be appropriate in training instructors. It can also easily be expanded to scores from 0 to 100. Second, the entries for multiple-rating items are often considerably longer than the usual brief entries for multiple-choice items. In a test of the ability of future English teachers to grade essays, for example, they might consist of short English essays. In a law school, they might be short case summaries. We mention other extensions later.

MULTIPLE-CHOICE ITEMS AND TRUE-FALSE ITEMS AS LIMIT CASES

The multiple-choice item is the lower limit case of the extended sub-class of multiple-rating items where the response is from the evaluative vocabulary of *ranking*. It is the limit case because there is no feebler ranking of a set of candidates than picking one as superior to the others and saying nothing about the rest. Yet, in a profession which tends to be fixated on the multiple-choice paradigm, the very opposite is claimed: "The multiple-choice item is the most flexible of the objective test item types."[7] This claim is based on ignoring the obvious generalizations of the multiple-choice item—some of which we have already mentioned—and imagining that the only contrast is with true-false and matching items. Within the sub-class of multiple-rating items that involve ranking the entries, the most obvious extension would involve full ranking of a set of, say, two to five claims for probability, or relevance, or credibility, or critical power, or robustness.

The true-false item is the limit case of the extended sub-class of multiple-rating items where

[6] Further note. Confidence level is best thought of as an orthogonal dimension to quality or merit. It could be added to multiple-rating items as well as to multiple-choice items. Estimating confidence levels definitely has some real-world value, but the line of argument for showing that adding confidence-ratings to *items* taps the real-world skills efficiently has seemed unconvincing, especially if complicated by the forced-distribution ('Q-sort') constraints that were often included in that line of development.

[7] This is from page 223 of *MEPE*. Similar remarks are to be found in most other texts.

the rating vocabulary is restricted to *grading*. There is no weaker grading than grading from a scale with two points on it, and the classic examples of such scales are the True/False, Right/-Wrong, and Correct/Incorrect scales for the rating of actions, decisions, or claims, and the Pass/Not Pass scale for rating performances or achievement. (Note that 'weaker' is not the same as 'easier'; the Pass/Not Pass decision may be a very difficult one.) Within this 'grading sub-class' of multiple-rating items, the obvious extension is to grading using a scale with more than two points on it, for example, grading answers to a science question as Completely Correct, Partially Correct, and Completely Wrong (a good question in a test for future science teachers, and perhaps for science students). The testing literature tends to view the true-false item as inadequate[8], but it has progressed only one stage from it in treating the multiple-choice item as a paradigm and treating extensions beyond that as illicit or of minor interest.

There are three other important (sub)classes of multiple-rating items besides the grading and ranking sub-classes. One is the class of items where the required rating is quantitative—the 'scoring class'. Another is the 'apportioning class', where the response involves allocating budget dollars, points, or other valuable resources across a range of options[9]. The third, of which an example was given earlier is the 'appraisal class' where a range of specific critical and/or laudatory comments are provided, from which an appraisal is to be assembled by selecting an appropriate ensemble; diagnostic vocabularies can also be used in such items. Each of them is 'objective' in the same sense as multiple-choice items, i.e., template-scoreable.

It is worth noting that the general ranking and grading tasks just discussed, both excluded from the standard test lexicon, are exactly the ones that every teacher uses in evaluating students. These tasks are usually done responsibly and, when done properly, have considerable reliability and validity. Hence one can hardly argue that multiple-rating items where a testee is asked to do the same thing cannot be marked with acceptable objectivity. And of course, one can always increase the validity of the scoring by increasing the simplicity of the distinctions that have to be made, the same process used in multiple-choice item design.

Why were these extensions never taken seriously? Why should tasks which most teachers can do with a useful degree of objectivity be excluded from those the student is asked to do in a test? It is hard not to be reminded of the story of the Garden of Eden, where God forbade the humans from partaking of the fruit of the Tree of Knowledge because doing so would mean, as God put it, that "the man is become as one of us, to know good and evil."

[8] A typical comment: "...the true-false item is not considered to be a sound method of assessing student performance...". (*MEPE* p. 217)

[9] Apportioning is a procedure embodied in Q-sort testing methodology. Further details on the taxonomy of evaluation are provided in the author's ***Evaluation Thesaurus (4th edition)***, (Sage, 1991).

THE LITERAL vs. THE TECHNICAL DEFINITIONS OF MULTIPLE-CHOICE

It will no doubt have occurred to the intelligent reader by now that the examples of multiple-rating items give above are in some literal sense 'really' multiple-choice items. Take the simple Rating item where the testee is to award one of the grades A through F to each of several entries . In constructing the test, we set up five boxes next to each of the entries[10], one of which is to be chosen, i.e., blacked in or checked. Is that not a straightforward case of a multiple-choice item?

In classifying items as multiple-rating items rather than multiple-choice items (where we are making that distinction) it is important to distinguish between the literal and the technical meanings of the term "multiple-choice item". In the literal sense, an item is a multiple-choice item if it can be set up with all possible responses listed, from which one choice is to be made. In the literal sense, *all written and verbal tests with responses from a finite set, including all fill-in and matching tests and all multiple-rating items, are—in essence if not in appearance—multiple choice tests.* For example, a fill-in question on the date of the Versailles treaty, could have all ten digits listed below each of the four blanks for digits to make a multiple-choice 'equivalent' of the original fill-in item[11]. This is not, however, the technical sense of the term "multiple-choice item", since in the technical sense (used in texts on testing and measurement) we normally and usefully *distinguish between* fill-in and matching tests, on the one hand, and multiple-choice items on the other. And when we come to the section on multiple-choice tests in a text, they are explicitly or implicitly defined as tests with a stem and four or five possible responses to the stem, from which one is to be chosen as the best[12].

One value of the literal definition of multiple-choice item is that it identifies the theoretical limits of template-scoreability. The other is that it reminds us how ill-thought-out these categories have been, since it's clear enough that the technical definition is very much narrower than the literal one, yet measurement specialists asked to react to the idea of multiple-rating items frequently reply by saying that there is nothing new about them since they are just a special case of multiple-choice items. Whenever the literal or commonsensical meaning of a term is violated by the technical definition—as with the term "objective test" or"multiple-choice item" —there is certain to be confusion. Since the texts virtually never address these confusions although the entering student will almost certainly suffer from them, and since the comments just referred to by test specialists incorporate the confusion, it seems possible that the confusion is in the minds of many testing specialists and not just in the beginning student of the subject.

[10] Or six if we include an NA box, sometimes useful in training graders.

[11] Fill-in tests have a de facto upper limit on the number of alphanumerical symbols used in the answers

[12] *MEPE*, p. 223.

What testers usually have in mind when talking about multiple-choice items is the technical or *paradigm* multiple-choice item. The multiple-choice item paradigm is very unlike the multiple-rating item paradigm (that is, one which has not been converted to a multiple-choice item). To recapitulate: in the paradigm multiple-choice item a *small* number of alternative *statements* (usually four or five) are offered, *one* of which is to be selected as true, the others of which are *defined by the tester* as the only other possibilities to be considered. The paradigm multiple-rating item, by contrast, offers *any* number of entries (1-25 would be the usual range), which *may or may not be statements*, of which *all* are to be *rated*, but none of which are guaranteed as having any value or any different value from the others, using a set of predefined or pre-known *responses* which are in simple cases *logically* exhaustive of the possibilities and which may also, independently, run from 3 to 15 or 25.

When we reduce multiple-rating items to multiple-choice items what we actually get is so different from the *paradigm* multiple-choice item that it is sometimes referred to in texts as a completely different type of item, the 'alternative-response' item. The alternative-response item is also, in the *literal* sense, simply a multiple-choice item, since it is reducible to a multiple-choice item in much the same way as the matching and fill-in items. This literal notion of multiple-choice item is thus too all-embracing to be useful except for special purposes. The value of terms lies in the distinctions they make, not just in the content they convey. Reflection will make clear that we can even reduce all multiple-choice items to True/False items, but that is not a good reason for abandoning the distinction between the two. The fact is that we have come to accept the 'technical' or 'paradigm' multiple-choice item as the meaning of the term 'multiple-choice item', and we take it to be quite different from 'true/false item'. The same reasons make it sensible to separate multiple-choice items from multiple-rating items and from this point on we use the term "multiple-choice item" in its technical sense.

MULTIPLE-RATING ITEMS vs. MODIFIED MULTIPLE-CHOICE ITEMS

A number of ways to improve the yield from multiple-choice items have been tried, although they get very short shrift in the texts. In general, they do something to improve multiple-choice items and close the gap between them and multiple-rating items, and several are built into standard multiple-rating item practice or options. The suggestions include: (i) removing the guarantee that *only one* option is correct. This requires that each option be selected or rejected on its own merits. This essentially converts the multiple-choice item into a set of true-false items sharing a common stem. It moves the multiple-choice item one step nearer the multiple-rating item but requires the difficult construction task of creating a set of items that are definitely but not too obviously true or false. The multiple-choice item avoids that by using the evaluative term "best" in its instructions and the multiple-rating item avoids it by providing the testee with an still more extensive evaluative vocabulary.

Or (ii) one may make the set of entries open-ended—that is, remove the guarantee that it *includes one* acceptable answer; this is part of the basic multiple-rating item. This can be combined with an element of a constructed-response item by (iii) requiring fill-in of a correct answer if none of the entries is acceptable, or a justification for every rating of an asterisked entry[13]. Or one can (iv) differentially score errors so that ratings which exhibit grave lack of understanding are treated more severely than those which select an almost defensible option; this procedure is recommended as standard for multiple-rating items and is built into most essay scoring rubrics. It makes sense to everyone that since some ratings display a nearly correct evaluation of an entry, and others show a complete misunderstanding of its content or worth, the two ratings should get different marks[14]. Similar comments apply to (v) giving more points for some questions than others[15]. One may or may not elect to flag the questions where more points are available, accompanied by a risk of more negative points, for those students who need to develop some refinements in their risk-taking behavior. (vii) Finally, the 'multiple-hurdle' approach, a common feature in the sophisticated Regional Board exams in the UK, should always be kept in mind for use when appropriate. This refers to requiring a passing mark on one or more sub-sets of the questions.

The increased complications from using one or two of these options can be minimized by using them only *occasionally*. One can often afford the resources to provide one question out of twenty where no answer is correct, so the extra scoring time for the filled-in responses is slight, but they provides some measure of cheating and guessing control, as well as mentation and expression checks for remedial advice. Of course, they are hybrid items, no longer template-scoreable. The use of custom scoring profiles with zero or negative points for absurd ratings, and variable points for reasonable though less-than-best alternatives has the merit of rendering redundant any correction for guessing—and the undesirable incentive for blind guessing. (Intuitive guessing a.k.a. 'best guessing' can still pay off.) The modest extra effort involved in constructing these items and scoring keys makes especially good sense when an item-pool is

[13] This has the advantage in small-scale use that one can use a computer program running a laser printer to randomly asterisk different entries for each student's test paper, greatly reducing the pay-off from 'peek-cheating' if one makes it clear that a condition of passing is acceptable justifications of the ratings for asterisked entries.

[14] Student anxiety is reduced by using 'near-miss' scoring. One implementation of this: if two marks are given for the correct rating on an A-F scale, one mark may be given for an adjacent rating, zero for the next most proximate and a deduction for the most inappropriate rating. If one seeks to avoid negative marks, these scores are simply increased by one. But the process is probably more effective as an antidote to wild guessing if the negative marks are known to be possible. In either case, the payoff from wild guessing is reduced to near-zero. I have done some investigation of alternative scoring strategies and would be happy to send a memo about them to anyone interested enough to comment on the suggestions.

[15] The Law School Admission Test, to take an example of a highly developed test, is an assemblage of items of vastly different difficulty, all awarded one mark for the correct answer.

being developed for long-term use. For a testing service, the approach could even be justified for one-time use, depending on the customer needs profile.

THE PRODUCTION OF RATING ITEMS vs. CHOICE ITEMS

Teachers are often somewhat nervous about the prospect of setting and objectively scoring multiple-rating items. Now, the task set by a Rating item is more like the task faced by a *teacher* in dealing with student assignments than it is like the task set to a *student* by a conventional Choice item. This makes Rating items seem unusual things to be putting into a test for students; but it is in fact one of their greatest strengths, because it moves the student towards skills of self-evaluation, often the key to major progress. And once this similarity has been noticed, a source for good items is apparent; many of the mistakes made by students are important mistakes for other students to reflect on, and a multiple-rating item gives them the chance to do this, and to get feedback on their decisions in the post-test discussions.

It should also be borne in mind that the task for the professional test-constructor in producing *good* multiple-choice items is quite formidable. That is, one can delude oneself into thinking one has 'mastered' the art of constructing multiple-choice items when in fact one is producing very poor items. The talents of Rating items are a little less subtle and people often report that they feel they can more easily tell whether they have made up a good one from looking at it critically. There is also less 'wastage'. It is hard to make up plausible distractors, and most of them are 'discarded' by the respondent. But it takes a *set* of them to generate one response. In constructing a multiple-rating item, on the other hand, each entry generates a complex judgement, while—as with the multiple-choice item—one stem can still serve several entries. A discrimination similar to that forced on the respondent by the set of distractors in the multiple-choice item can be included by suitable construction of a single entry of an multiple-rating item: it is the discrimination needed to distinguish between—for example—a C and a B answer.

CONCLUSION

In deciding whether to launch forth upon the comparatively untraveled sea of Rating items, one should bear in mind the major comparative advantages over using Choice items. (i) The logic of the Choice item creates its well-known Achilles heel, the use of the 'elimination algorithm'. That is, the respondent can infer from minor flaws in 3 responses to the truth of the 4th without having enough knowledge or understanding of the 4th to recognize its truth—or its likelihood, or even its meaning—from reading it by itself. (ii) Knowing nothing at all about a topic, blind guessing still gets a substantial score on multiple-choice tests and teachers often give a D rather than an F for scores at chance expectation level. (iii) Imperfect item construction will often make it possible to use grammatical and other clues to do much better than the chance expectation, thus gaining a C on a course without having learnt anything at all from it. (iv) Opportunities in real life to agree or disagree with claims, or to select or reject options, are often not accompanied by distractors, let alone guarantees that the set offered includes the right answer, so the transfer to real situations of the skill identified by—or acquired from training aimed at—Choice items is severely restricted. (v) In the higher-order cognitive domain, for example with respect to the transcurriculum skills, the ability to *envision* alternatives is often a key part of the basis for evaluation of proposals and a testing situation in which the alternatives are not only provided but limited by fiat is extremely uncharacteristic. (vi) Use of Rating items gives practice in considerably more refined discrimination than True/False or Best Choice options, again a common real-world necessity. Given these considerations, it may be worth some effort to switch. The suggestion here is that most uses of Choice items beyond tests of rote knowledge and simple discrimination skills should be converted into or replaced by multiple-rating items which avoid all these problems.