1984

8. Achievement Test Items: Current issues

Robert L. Ebel
Michigan State University

8 Achievement Test Items: Current issues

Robert L. Ebel
Michigan State University

The writer of achievement test items is confronted with two major problems, as Lindquist pointed out nearly half a century ago (Lindquist, 1936, p. 17). The first of these is the problem of what to measure. The second is how to measure it. The solution proposed for the first problem is to focus primarily on testing for knowledge and only secondarily on testing for abilities. Cognitive abilities, it is reasonable to believe, depend entirely on knowledge. Although the term *knowledge*, as commonly used, includes both information and understanding, the most useful kind of knowledge, the kind that will occupy our attention almost exclusively, is that which involves understanding. Understood knowledge is a structure of relations among concepts. To understand is to be aware of relationships. Each of these relationships can be expressed in words as a proposition.

The solution proposed for the second problem is to present the examinee with a series of incomplete propositions, accompanied by two or more alternative completions, only one of which makes the proposition true. Many of the current issues in the writing of achievement test items are related to these two proposed solutions.

A CONCEPTION OF KNOWLEDGE

Knowledge originates in information that can be received directly from observations or indirectly from reports of observations. These observations may be external (objects or events) or internal (thoughts and feelings) (Scheffler, 1965, p. 137). Information feeds the mind and, like food for the body, it must be digested and assimilated. Thinking is the process by which these things can be accomplished (Newman, 1852, p. 134). Information that is simply stored in memory remains only information, the lowest, least useful form of knowledge.

But if the information becomes the subject of reflective thought, if those who received it ask themselves, "What does it mean?" "How do we know?" "Why is it so?", the information may come to be understood. It may be integrated into a system of relations among concepts and ideas that constitutes a structure of knowledge. This has been referred to as "semantic encoding" (Anderson, 1972, p. 146). Information that is understood, that is incorporated into a structure of knowledge, tends to be more powerful, more useful, and more satisfying. It is likely to be a more permanent possession than information that is simply remembered (Boulding, 1967, pp. 7–8).

The basis for verbal knowledge exists in the mind in a form that Polanyi (1958) has called "tacit knowledge." In that form, it is a purely private possession. But if concepts can be abstracted from these images and expressed in words, and if the relations among the concepts can be expressed in sentences, then tacit knowledge is converted into verbal knowledge. This can be communicated and thus made public. It can also be recorded and stored for future reference. It can be manipulated in the processes of reflective thinking. It is thus a very powerful form of knowledge. The peculiar excellence of human beings among all other creatures on earth is their ability to produce and to use verbal knowledge. Thinkers produce it. Teachers and students, planners, and managers use it. Classrooms and libraries and study rooms are full of it. So are conference rooms, memoranda, and reports. It would be difficult to overstate the importance of structures of verbal knowledge in human affairs (Hayakawa, 1941, pp. 15–25; Langer, 1957, pp. 200–204).

If a structure of verbal knowledge consists entirely of a system of articulated relations among concepts and ideas, can it be described *completely* by listing the elements (propositions) that compose it? Might not a complex structure involve relations or dimensions that are not expressed by the constituent elements of the structure? Indeed it is possible that a listing of the elements of a structure might omit some that *have not been* perceived or expressed in words. But it is unreasonable to believe that there might be important elements of the structure that *could not be* perceived and expressed; to cite an example of such an unperceived and unexpressed element, one would have to perceive and express it. Once it had been expressed, it could be added to the list. The conclusion that a structure of verbal knowledge can be described completely by listing the elements that compose it appears to be logically necessary. Where structures of knowledge are concerned, the whole seems to be precisely equal to the sum of *all* the parts.

THE RELATION OF KNOWLEDGE TO ABILITY

The contribution of knowledge to effective human behavior is sometimes questioned. Knowledge alone is not enough, says the businessman. It does not guarantee financial success. Knowledge alone is not enough, says the college

president. It does not guarantee scholarly achievement. Knowledge alone is not enough, says the religious leader. It does not guarantee virtue. Knowledge alone is not enough, says the philosopher. It does not guarantee wisdom.

They are all right, of course. Knowledge alone is not enough. But in this complex world of chance and change, no one thing nor any combination of things ever will be enough to guarantee financial success or scholarly achievement or virtue or wisdom. Although this is true, few would deny that the command of knowledge does contribute greatly to the attainment of these other, more ultimate goals.

The term *knowledge,* as it is used in this chapter, means considerably more than the same term means in the Bloom Taxonomy (Bloom, 1956, pp. 201–297). There, knowing something means simply being able to recall it. Having knowledge is nothing more than having information. Here, the term *knowledge* refers not only to information but also and, far more importantly, to understanding, which requires a structure of relations among concepts. In addition, the emphasis here is on useful knowledge. If knowledge is not available to be used, it is not fully possessed. Thus the possession of knowledge, as the term is used here, should enable a person to demonstrate all the other abilities and skills identified in the other categories of Bloom's Taxonomy: comprehension, application, analysis, synthesis, and evaluation. If one *knows* how to do these things, one ought to be *able* to do them.

THE MEASURABILITY OF HUMAN CHARACTERISTICS

Any important human characteristic is necessarily measurable. To be important, a personal characteristic must make an observable difference, that is, at some time, under some circumstances, a person who has more of it must behave differently from a person who has less of it. If different degrees or amounts of a personal characteristic never make any observable difference, what evidence can be found to show that it is, in fact, important?

But if such differences can be observed, then the characteristic is measurable, for all measurement requires is verifiable observation of a more–less relationship. Can integrity be measured? It can if verifiable differences in integrity can be observed among men. Can mother love be measured? If observers can agree that a hen shows more mother love than a female trout or that Mrs. A. shows more love for her children than Mrs. B, then mother love can be measured. The gist of the argument is this. To be important, a personal characteristic must make a difference. If it makes a difference, the basis for measurement exists.

In principle, then, any important human characteristic is measurable. In practice, however, many characteristics said to be important seem to be very difficult to measure. Where can one find a reliable test of ability to see relations, to

formulate hypotheses, to interpret data, to organize ideas, to draw conclusions, to solve problems, or to think?

Perhaps the difficulty may lie in the characteristics themselves. Perhaps they simply do not exist as separate, unified, measurable abilities. Perhaps what we call abilities are simply categories of tasks that have some superficial characteristics in common but which cannot be dealt with effectively by the application of a single general task-related ability. Perhaps what they may require mainly is knowledge of the special context in which the tasks arise. Take problem solving for example. The problems a physician must solve are likely to be quite different from those a chess player or a football coach or a highway engineer or a theoretical physicist must solve. No test of general ability to solve problems is likely to predict very accurately how successful a practitioner of each of these arts or crafts is likely to be. Too little of what makes a physician successful in problem solving is also likely to make the chess player, the coach, the engineer, or the physicist successful.

Many of the alleged abilities that are said to be important human characteristics have never been defined operationally, which must be the first step in developing valid measures of them. If an operational definition of one of these very general abilities could be developed, it might lead to a test composed of such a heterogeneity of tasks, with very low intertask correlations, that the test scores would be very low in reliability. When this is the case, differences among individuals in the amount of this general ability are likely to be difficult to discern. It will probably be equally difficult to show that such differences matter very much. If they make little difference on a test designed to measure them, they are unlikely to make much difference in other contexts. If this is the case, they cannot be of great importance.

It may be a waste of time and energy to try to measure "hard to measure" human characteristics. Their measurability is directly related to their importance. For the same reason it may be a waste of time and energy to try to develop these "hard to measure" characteristics through instruction. A teacher who claims to be doing so without being able to produce evidence of success in doing it (because, you see, they are "hard to measure") may be simply throwing dust in our eyes. Those who argue that "what can be easily assessed should not dictate what is taught" are mistaken. If it cannot be easily assessed it cannot be surely taught. It is not likely to be worth trying to teach.

An instructor who wishes to develop in pupils some important characteristic must first devise a method for measuring reliably how much of that characteristic each pupil has acquired. Then the instructor must devise a method for developing that characteristic. Finally the instructor ought to measure the effectiveness of his efforts. Most teachers can find a sufficient challenge to their abilities and commitments in teaching things that are not "hard to measure." They should not add unnecessarily to the difficulties and frustrations of their work by undertaking to teach and to test "hard to measure" achievements. Teachers would teach more

effectively and talk more sensibly if they would ban references to "hard to measure" qualities from their discourses.

THE RELATIVE MERITS OF ESSAY AND OBJECTIVE TESTS

Specialists in testing tend to recommend the use of objective tests in general and multiple-choice items in particular. They claim not only that objective tests are more objective and convenient but that they provide more extensive samples of the ability to be tested and yield scores of higher reliability. Critics of multiple-choice tests claim that essay tests, despite their limitations and the difficulties of using them, provide more valid measures of ability and encourage more wholesome educational practices.

In considering the relative merits of essay and objective tests, it is important to make this point at the outset. If the purpose of the test is, as it is usually, to determine how much useful knowledge a person has on some subject, then that purpose can be achieved by using either an essay or an objective test. The point is important because some believe that essay tests call for a different, and higher, level of mental ability than is required by an objective test. The fact, however, is otherwise. There is no empirical evidence to support belief in such a difference and no rational basis for expecting it.

It is reasonable to believe that any cognitive ability consists entirely of knowledge of how to do something. That knowledge is made up of a structure of elements of knowledge, a structure of relations between concepts and ideas. By testing examinees for possession of a sample of those elements, one can determine the extent and strength of their structures of knowledge relevant to the ability and, thus, the degree to which they possess the ability.

If person A knows more about a subject than person B, then A is likely to write a better answer than B to an essay question on the subject. A is also likely to give more correct answers than B to an objective test on the subject. The correctness of the answers either person gives to either type of test question depends largely on the extent and firmness of that person's structure of knowledge.

It is true that essay tests present tasks to the examinees that are distinctly different from the tasks presented in objective tests. The difference, however, is more one of form than of substance. In both cases the information used in giving the answer comes from the examinee's structure of knowledge. In both cases an examinee must choose information relevant to the question being asked. Then, with an essay test answer, the examinee must choose how to express in words the relevant items of information and the conclusions to which they lead. With an objective test, the examinee must choose how to relate the relevant items of information to the questions posed by the item and then choose which of the

answer options is best supported by the relevant information. In both cases the foundation for an answer is the examinee's structure of knowledge. In both cases the process of arriving at an answer involves making repeated choices. In both cases the examinee must apply the knowledge possessed, must relate and infer as well as remember.

The advantage that essay tests have in not suggesting the correct answer to a question, or providing clues to it, is more apparent than real. Those who are most successful in selecting correct answers to a multiple-choice question tend to be also more successful in producing good answers to essay test questions (Cook, 1955). The cues to the correct answer that multiple-choice items provide seldom give away the correct answer to one who lacks knowledge of it or ability to infer it. Multiple-choice items often prove to be too difficult to discriminate well despite the cues they may provide. If the multiple-choice items are well written, cues to the correct answer will be offset to some degree by other cues that suggest incorrect answers to poorly informed examinees. The items in which cues are likely to be most helpful are the less desirable kinds in which a previously learned answer simply must be recognized. If the item requires application of what has been learned to answer a question or solve a problem that has never been encountered before, cues will be less helpful. Presenting a good test question in multiple-choice form seldom if ever makes the question too easy to do its job well. Seldom if ever does presentation of correct answers keep objective tests from clearly distinguishing those who know more from those who know less about a subject.

Whatever theoretical advantages there might be to having the examinee produce an answer are likely to be offset by the tedious, subjective process of evaluating the answer and the unreliable scores that often result. Errors in scoring objective tests are quite rare and usually very small. Differences of opinion in judging the quality of essay test answers are often substantial. This is not to say that there are no occasions on which an essay test should be used in preference to an objective test. It is to say that a general preference for essay tests is unwarranted. The ability tested by an item is determined mainly by the content of the question, not by the form of the response.

THE MERITS OF ITEMS BASED ON REALISTIC PROBLEM SITUATIONS

For over 40 years some test specialists have recommended the use of test items based on verbal descriptions of realistic problem situations. Items of this kind are suitable for inclusion in paper and pencil tests. They are more realistic than items that test directly for possession of knowledge or for understanding of principles and procedures. They are less realistic than performance tests presented in simu-

lations of "real-life" situations. A discussion of the possibilities and problems of applied performance testing can be found in Fitzpatrick and Morrison (1971).

The inclusion in paper and pencil tests of items that present verbal descriptions of realistic problem situations has several attractions to test constructors. It demonstrates that objective tests are not limited to testing for recall of isolated, trivial factual details. Situation-based items cannot be answered correctly by simple recognition of the right answer. They force the examinee to think. They obviously require the application of knowledge to real-life problems. Realism in the test encourages faith in the validity of the test scores. These are valuable assets. But situation-based items also have disadvantages. They tend to be complex and wordy. Complexity may obscure the crucial element in the situation, complicate the task of the examinee, and thus lower the discriminating power of the items. It is true that the real problems we face in living are complex. Unfortunately, complex, real problems seldom have single demonstrably correct right answers. Giving a person a complex problem to solve may not be the best way to estimate that person's capability of solving such problems.

Ordinarily a complex test question contributes only a single unit to the total test score. It is answered correctly (1) or incorrectly (0). But to arrive at the final answer to a complex question, the only answer that counts, the examinee must provide himself with a multitude of intermediate or contributory answers that do not count. To reach a correct answer, each of a number of contributory steps must be taken correctly. A single error in any one of them may lead to a final answer that is wrong. The value of nine correct decisions can be offset by the penalty for one that is incorrect. Should not right and wrong decisions carry more nearly equal weight in judging an examinee's capabilities? Would it not be more reasonable, would it not be more informative, would it not lead to more accurate measurement of the mental ability being tested to assess the correctness of each step independently?

Some would say not, arguing that the whole is more than and more significant than the sum of its component parts; that ability to avoid even a single error during a complex process is the essence of competence. The argument is not without merit. Surely it is true that in the ordinary affairs of living, single errors can be very costly. One thing done wrong can cancel the rewards for doing many things right. But is our purpose in measuring mental abilities to imitate life? Or is it mainly to assess a person's cognitive resources, that is, the person's knowledge and mental abilities? For that purpose it may be appropriate and advantageous to take each decision into account and to assess them independently. It may be inappropriate and disadvantageous to consider only a single outcome from a sequence or cluster of related, contributory decisions.

Wordiness should make the items more time-consuming so that fewer could be included in a test of given duration. Obviously a test composed of simple items will yield more independent scorable responses per hour of testing time and

hence will tend to yield more reliable scores than a test composed of complex items. Simple test items should also be easier to comprehend and present fewer ambiguities or occasions for misinterpretation by the examinees. Because of these differences one would expect scores of higher reliability from simple than from complex items in tests of similar duration. Experimental studies by Howard (1943) and by Ebel (1953) have confirmed these expectations. It seems difficult to obtain scores of reasonable reliability in tests of reasonable duration if the test items are situation based. This has been true of patient–management problems in medicine (Skakun, 1979), of air crew problems derived from critical incidents in military aviation, and of simulations in legal education (Alderman, Evans, & Wilder, 1981). There seems to be an inverse relation between the realism of the problem situations in the test and the reliability of the scores yielded by the test.

Recognizing these disadvantages, the test constructor may still favor the use of situation-based items. For they do test examinee understanding, abilities to apply knowledge, and ability to think. Is there any better alternative? There may be. Whereas items involving complex, realistic problem situations are often inefficient, ambiguous, and indeterminate, items testing elements of knowledge tend to be efficient and can be less ambiguous and more determinate. There are reasons for believing that most cognitive abilities that can be measured by situation-based test items can also be measured, perhaps with greater efficiency and reliability, by proposition-based items. In many situations, tests composed of simple items may provide more efficient and accurate measures of mental abilities than can be provided by complex test items. In item writing as in many other arts, simplicity can be a virtue.

THE MERITS OF ALTERNATE-CHOICE ITEMS

A simple approach to assessing knowledge is available to those who can accept the idea that knowledge is a structure of relations among concepts. Each of the relations that makes up the structure can be expressed as a proposition. A proposition is simply a sentence that can be said to be true or false (Cohen & Nagle, 1934, pp. 27–30). Propositions similar in appearance to those that are part of the structure but expressing relations that are not part of the structure can also be written. The person whose knowledge is being assessed is asked to distinguish between the correct and the incorrect propositions.

This sounds suspiciously like a true–false test, as indeed it is. True–false tests, however, have been condemned by many specialists in testing, often with considerable vehemence (Adkins, 1947, p. 41; Travers, 1950, p. 42). Other authorities have suggested a different view, which I share. The faults found in true–false items are not inherent in the form but sometimes result from careless or incompetent use of it (Bergmann, 1981, p. 92; Popham, 1981, p. 243).

Both the amount of guessing pupils do in taking true–false tests (Ebel, 1968) and the amount of error that the guessing contributes to their scores (Hills & Gladney, 1968) tend to be exaggerated. Classroom true–false tests of approximately 100 items have yielded coefficients of reliability in the .80s and .90s. These results would be most unlikely if the scores were distorted seriously by guessing.

Each true–false item tests only one element in a structure of knowledge, but there can be many such items in a test. No single essential element in an important structure can be regarded as trivial. If the item is seriously ambiguous, or if it encourages rote learning, much of the fault must be with the one who wrote it. Elements in a structure of knowledge *can* be expressed clearly. They do not need to reward rote learning by being expressed in the exact words or sentences of the textbook or lecturer. The fear that incorrect propositions in a true–false test will lead to wrong learning has proved to be unjustified (Ross, 1947, p. 349; Ruch, 1929, p. 368).

Despite their intrinsic relevance to the assessment of achievement in learning, true–false test items can be ambiguous. They call for absolute judgments of truth or falsity. They do not offer different answers among which the examinee can choose. Because few statements are complete and accurate enough to be perfectly true, the examinee must decide how far the statement can deviate from perfect truth and still be called true. This is one source of ambiguity. Another is lack of clarity in the focus of the item. The element in the statement that is crucial to its truth or falsity is not identified clearly to the examinee.

An alternative to the true–false item, designed to remove some of the ambiguity, is the alternate-choice item. It consists of an incomplete statement of a proposition along with two or more alternative completions, only one of which makes the statement true. For example:

> An eclipse of the sun can only occur when the moon is:
> (1)full (2) new.

Items of this kind do not call for absolute judgments of truth or falsity. The critical element in the statement they make should be quite clear. Their indices of discrimination should be higher on the average than the indices of comparable true–false items given to the same examinees. The test scores therefore should be more reliable. A recent study has verified this expectation. Students ($N = 28$) enrolled in a class in educational measurement took parallel 25-item true–false and alternate-choice tests on each of eight units of instruction in the course. The Kuder–Richardson 20 reliability coefficients for the true–false tests ranged from .13 to .71, with a mean of .47. Those for the alternate-choice tests ranged from .56 to .76, with a mean of .66 (Ebel, 1982).

Alternate-choice items are distinctively different from the familiar four-alternative multiple-choice items in ways other than the number of response options

offered. Because they tend to be simpler and use fewer words, they take less time per item (Ebel, 1953). This could lead to higher reliability for tests of a given duration. The response options tend to be shorter, often one or two words, which focuses the attention of the examinee more clearly on the element of knowledge being tested.

One objection likely to be raised to the use of the alternate-choice items is that they deal with isolated factual details. Their brevity and specificity may be taken as indications of triviality (Highet, 1950, p. 120). But if the conception of knowledge presented in this chapter is correct, if verbal knowledge can be expressed completely as a structure of relations, if each of these relations (the elements of the structure) can be expressed as a proposition, and if each proposition is used as the basis for an alternate-choice item, then one can assess the extent and firmness of the whole structure by examining the parts that compose it (Thorndike, 1935; Wood & Beers, 1936, p. 162). The choice of a response to an alternate-choice item is simple to indicate, but the process of making it rationally may be quite complex. If a problem like the following has not been encountered before, it is likely to test understanding and application as well as recall.

> The buoyant force on a ping-pong ball immersed in water is:
> (1) greater than (2) the same as (3) less than that
> on an iron ball of the same size. (Answer 2)

Even if the problem has been encountered before, it is reasonable to suppose that the person who understands the basis for the answer is more likely than the one who does not to give the correct answer.

When using the alternate-choice item form, the item writer is free to pose questions that admit only two good alternative responses. Here are some examples:

1. The density of ice is (1) greater (2) less than that of water.
2. A point on the surface of the Earth moves toward the (1) east (2) west as the Earth turns.
3. The average size of farms in the United States has (1) increased (2) decreased during this century.

Often, as in these examples, there is only one plausible alternative to the key word or phrase in the proposition.

When item writers are obliged to produce four-alternative multiple-choice items, they sometimes do so by combining several alternate-choice items. They may present four propositions and ask which one is true or not true. They may ask if a statement is true or false, and why. The responses might be: (1) true, because A; (2) true, because B; (3) false, because C; (4) false, because D. They may ask if something is true of both X and Y. The responses might be: (1) yes, both; (2) no, only X; (3) no, only Y; (4) no, neither. They may ask the speed and direction of a change, so that the responses might be: (1) rapid increase; (2) slow

increase; (3) slow decrease; (4) rapid decrease. Presented separately the two or more alternate-choice items would yield two or more independent indications of achievement. Combined, they yield only one. The result is likely to be a loss of reliability (Ebel, 1978).

Three other characteristics of alternate-choice items give them some advantage over conventional multiple-choice items. When the response options are brief, as they usually are, they can be included as parts of a continuous sentence and need not be listed below an item stem. This makes the typing simpler and the resulting pages more compact. When it is awkward to arrange the wording of the sentence so that the response options come at the end, they can be put in the middle or at the beginning. This sometimes simplifies the wording of the item. Finally, because alternate-response items are simple in structure, they are easier to write. There are fewer opportunities for errors in item writing that might spoil the effectiveness of the item.

One other point ought to be mentioned before concluding this case for alternate-choice items. There are items like the following in which more than two good response options are readily available. For example:

1. The gas given off in photosynthesis is (1) carbon dioxide (2) hydrogen (3) oxygen (4) nitrogen. (Answer 3)
2. Most of the territorial possessions of the United States were gained as a result of the (1) War of 1812 (2) Civil War (3) Spanish-American War (4) World War I. (Answer 3)

When more than two good response options are available, the item writer should probably offer more than two.

PROSPECTS FOR A TECHNOLOGY OF ITEM WRITING

Cronbach (1970) expressed the opinion that "The design and construction of achievement test items has been given almost no scholarly attention. The leading works of the generation—even the Lindquist *Educational Measurement* and the Bloom *Taxonomy*—are distillations of experience more than scholarly analysis [p. 509]." The contrast implied here between "distillation of experience" and "scholarly analysis" is interesting. Did not Lindquist and Bloom rely on scholars to aid in the distillations? Did not these scholars analyze the experiences of which they were aware? Is it obvious that a theory of item writing has much to add to the "distillation of experience" in the development of a technology of item writing?

Roid and Haladyna (1980) have reviewed recent research on item writing, with special attention to the more or less mechanistic or semiautomatic methods of item generation. Their article contains descriptions and discussions of six classes of methods for producing test items:

1. Those in which the item writer is guided by statements of the objectives of instruction.
2. Those whose items must meet specifications of the domain of content to be covered and the forms of items to be used.
3. Those in which items are produced by linguistic transformations of segments of prose instruction.
4. Those in which mapping sentences derived from facet theory are used to define a content domain.
5. Those whose items are designed to test understandings of concepts.
6. Those in which items are stored in or actually produced by computers.

The limitations of these methods is acknowledged clearly in the review. Each method appears to have a particular application. They cannot be applied to any content level and at any cognitive level. They require ingenuity and the exercise of judgment. At present they are in the infancy of their development. Cronbach believes that they will mature into useful tools for the test constructor. Others, including this writer, are more skeptical. Roid and Haladyna endorsed Berk's (1978) observation that the rigor and precision of item-writing specifications are inversely related to their practicability.

In a sense, the item development procedures outlined in earlier sections of this chapter constitute a technology for item writing. The form and derivation of the items is specified quite precisely. The content of the items depends on the item writers' knowledge and skills. Propositions that are important and defensible must be selected. They must be expressed clearly, accurately, and concisely. Incorrect answer options that have commonsense plausibility must be provided. The judgment involved in these choices is crucial, and no algorithm or computer program is likely to provide it.

CONCLUDING STATEMENT

This chapter has attempted to make 15 points.

1. Information is the source but not the substance of knowledge.
2. Useful knowledge is a structure of relations among concepts and principles.
3. The peculiar excellence of human beings is their ability to produce and to use verbal knowledge.
4. Cognitive abilities are entirely dependent on the possession of relevant knowledge.
5. The assumption that each kind of cognitive task requires a separate special cognitive ability is unnecessary and probably unwarranted.
6. Special tasks are more likely to require special knowledge than special abilities.

7. Any important human characteristic is necessarily measurable.
8. Human characteristics that are hard to measure are likely to be of limited importance.
9. Either an essay test or an objective test can be used to measure any important cognitive achievement.
10. Multiple-choice items that *present* correct answers among the response options can indicate quite accurately an examinee's ability to *produce* correct answers.
11. Items based on realistic problem situations tend to yield unreliable test scores.
12. Items that consist of incomplete propositions each of which is accompanied by one correct and one or more incorrect completions can yield valid measures of achievement.
13. Items that provide only two response options can measure achievement satisfactorily.
14. Technologies for the mechanical or semiautomatic generation of test items are likely to be of limited value.
15. Simplicity in the conception of what to test and in the means used to test it is commendable.

Paraphrasing Plato's assessment of the ideas he attempted to illustrate in the *Allegory of the Cave,* "Heaven knows if these things are true, but this, at any rate, is how they appear to me."

REFERENCES

Adkins, D. C. *Construction and analysis of achievement tests.* Washington, D.C.: U.S. Government Printing Office, 1947.

Alderman, D. L., Evans, F. R., & Wilder, G. The validity of written simulation exercises for assessing clinical skills in legal education, *Educational and Psychological Measurement* 1981, *41,* 1115–1126.

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research,* 1972, *42,* 145–170.

Bergmann, J. *Understanding educational measurement and evaluation.* Boston: Houghton Mifflin Co., 1981.

Berk, R. A. The application of structural facet theory to achievement test construction. *Educational Research Quarterly,* 1978, *3,* 62–72.

Bloom, B. S. *Taxonomy of educational objectives: Cognitive domain.* New York: Longmans, Green, 1956.

Boulding, K. E. The uncertain future of knowledge and technology. *The Education Digest,* November 1967, *33,* No. 3, pp. 7–11.

Cohen, M. R. & Nagle, E. *An introduction to logic and the scientific method.* New York: Harcourt Brace, 1934.

Cook, D. L. An investigation of three aspects of free-response and choice-type tests at the college level. *Dissertation Abstracts* 1955, *15,* 1351.

Cronbach, L. J. Review of *On the theory of achievement test items. Psychometrika,* 1970, *35,* 509–511.

Ebel, R. L. The use of item response time measurements in the construction of educational achievement tests. *Educational and Psychological Measurement,* 1953, *13,* 391–401.

Ebel, R. L. Blind guessing on objective achievement tests. *Journal of Educational Measurement,* 1968, *5,* 321–325.

Ebel, R. L. Proposed solutions to two problems of test construction. *Journal of Educational Measurement,* 1982, *194,* 267–278.

Ebel, R. L. The ineffectiveness of multiple true-false items. *Educational and Psychological Measurement,* 1978, *38,* 37–44.

Fitzpatrick, R. S., & Morrison, E. J. Performance and product evaluation. In R. F. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hayakawa, S. I. The importance of language. *Language in Action.* New York: Harcourt Brace, 1941.

Highet, G. *The art of teaching.* New York: Vintage Books, 1950.

Hills, J. R., & Gladney, M. B. Predicting grades from below chance test scores. *Journal of Educational Measurement,* 1968, *5,* 45–53.

Howard, F. T. *Complexity of mental processes in science testing.* Contributions to Education No. 879, New York: Teachers College, Columbia University, 1943.

Langer, S. K. Language and thought. In L. G. Locke, W. M. Gibson, and G. W. Arms (Eds.), *Toward liberal education.* New York: Rinehart, 1957.

Lindquist, E. F. The theory of test construction. In H. E. Hawkes, E. F. Lindquist, & C. Mann (Eds.), *The construction and use of achievement examinations.* Boston: Houghton Mifflin Co., 1936.

Newman, J. H. C. *The idea of a university.* London: Longmans, Green, 1925. (originally published, 1852.)

Polanyi, M. *Personal knowledge.* Chicago: University of Chicago Press, 1958.

Popham, W. V. *Modern educational measurement.* Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Roid, G., & Haladyna, T. The emergence of an item-writing technology. *Review of Educational Research,* 1980, *50,* 293–314.

Ross, C. C. *Measurement in today's schools* (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1947.

Ruch, G. M. *The objectives or new-type examination.* Chicago: Scott, Foresman, 1929.

Scheffler, I. Philosophical models of teaching, *Harvard Educational Review,* 1965, *35,* 131–143.

Skakun, E. N., Taylor, W. C., Wilson, D. R., Taylor, T. R., Grace, M., & Fincham, S. M. A preliminary investigation of computerized patient management problems in relation to other examinations. *Educational and Psychological Measurement,* 1979, *39,* 303–310.

Thorndike, E. L. In defense of facts. *Journal of Adult Education,* 1935, *7,* 381–388.

Travers, R. M. W. *How to make achievement tests.* New York: Odssey Press, 1950.

Wood, B. D., & Beers, F. S. Knowledge versus thinking. *Teachers College Record,* 1936, *37,* 487–499.