DOCUMENT RESUME

ED 208 001 TM 810 669

AUTHOR Ebel, Robert L.; Livingston, Samuel A.
TITLE Issues in Testing for Competency.
INSTITUTION National Council on Measurement in Education, Washington, D.C.
PUB DATE 81
NOTE 7p.
AVAILABLE FROM National Council on Measurement in Education, USPS 823120, 1230 17th Street N.W., Washington, DC 20036 ($2.50).
JOURNAL CIT NCME Measurement in Education; v12 n2 Sum 1981

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Competence; *Criterion Referenced Tests; Multiple Choice Tests; *Norm Referenced Tests; *Professional Occupations; *Statistical Analysis; Test Construction; *Testing Problems; Test Reliability; Test Validity; *Verbal Ability
IDENTIFIERS *Competency Based Assessment; Competency Based Certification; Noncognitive Attributes

ABSTRACT

This issue of Measurement in Education is presented in the form of a dialogue between Dr. Robert L. Ebel, Distinguished Professor of Educational Measurement at Michigan State University, and Dr. Samual A. Livingston, Program Research Scientist at the Educational Testing Service. Alternative views on some aspects of the use of tests in assessing professional competence are presented. Livingston and Ebel direct special attention to the shortcomings and virtues of verbal knowledge, multiple-choice items, norm-referenced tests, conventional test statistics, and test validation. Livingston is more convinced of the shortcomings of the first four and the virtues of the fifth than is Ebel. Despite their differences, both agree on the need for psychometric excellence. (AL)

NCME
measurement in education
A SERIES OF SPECIAL REPORTS OF THE NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

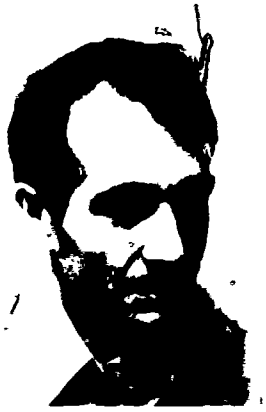
Volume 12 No. 2 Summer, 1981

Issues in Testing for Competency

Robert L. Ebel and Samuel A. Livingston

ABOUT THIS ISSUE

Robert L. Ebel



Samuel A. Livingston

This paper represents a departure from those published in the past issues of ME. The paper is presented in the form of a dialogue between Drs. Robert L. Ebel and Samuel L. Livingston. The topic, testing for competency, is one that has been addressed before by these two measurement specialists and others. What makes this paper interesting is the interchange of ideas about the topic between Ebel and Livingston.

The major thrust of this paper concerns an issue that is as old as testing itself. How can we assess the competencies needed to perform specific jobs that are not necessarily school oriented? How do we asses the skills and knowledge necessary to function effectively as a physician, as a barber, as a teacher or any other occupation? Can we indeed assess these competencies? If we can, *when* can we measure those necessary skills and knowledge? *In what mode* can we measure them, will traditional paper-and pencil tests suffice? Must we think of job performance observations as a major tool? *How can we employ the disciplines* inherent *in classical measurement* devices to observation of job performance? Can we?

While not all of these issues are examined equally, the reader will find some interesting points of view expressed by two highly respected men in our field. Dr. Robert L. Ebel is a Distinguished Professor of Educational Measurement at Michigan State University, East Lansing, Michigan. Dr. Samuel A. Livingston is a Program Research Scientist at The Educational Testing Service, Princeton, New Jersey. **HCR**

Prologue

ISSUES IN TESTING FOR COMPETENCY

This article reports alternative views on some aspects of the use of tests in assessing professional competence. It grew out of an exchange of letters between the two authors following a conference in Atlanta, Georgia, on [illegible]nber 7, 1978. At that conference, sponsored by the National Commission for Health Certifying Agencies, Ebel presented and undertook to define the following propositions:

Assessment of Competence

1. Certification of competency is essential to the maintenance of excellence in any profession.
2. Periodic re-assessment of competence, and recertification is desirable for the maintenance of professional excellence.
3. The major component of professional competence is verbal knowledge.

4. Written tests can provide effective assessments of a person's verbal knowledge

Criterion/Norm-Referencing

5. Criterion-referenced tests are intended to identify examinees who have reached a certain criterion on one or more aspects of proficiency.
6. The particular elements of knowledge to be tested are identified more specifically on a criterion-referenced than on a norm-referenced test.
7. It is seldom advisable for a test of professional competency to focus sharply on a limited number of discrete, sharply defined competencies.
8. Criteria of competency in practice of a profession tend to be norm-referenced.
9. Procedures for determining the passing score on a test of competency should be developed as rationally as possible, and then described in explicit detail.
10. Criterion-referenced tests can be evaluated using the same statistical procedures that were developed for non-referenced tests.

Job Relatedness

11. A test of competency in a profession is job-related if it reflects a rational analysis by expert practitioners of the essential functions of the professional.
12. A consensus of experts provides the only sound basis for specifying the content of a test of competency.
13. Knowledge is a necessary, but not a sufficient condition for effective performance.
14. Good test questions require the examinees to apply the knowledge they possess.
15. There is a high correlation between ability to recall and ability to apply knowledge.

Validity

16. The validity of a test of competence is determined by the tasks it includes and by the reliability of the scores it yields.
17. What a test measures is usually what it appears to measure.
18. Statistical validation of tests of competency is seldom feasible.
19. On a good test of competence, test-taking skills cannot be substituted for knowledge of the subject.
20. Other means of assessment; interviews, recommendations, biological data blanks, assessment centers, etc , are supplements not alternatives to written tests.

Reliability

21. A reliability coefficient is the correlation between the scores from two or more indpendent measurements of competence for the individuals in a particular group.
22. The reliability of a set of test scores depends on the number and quality of the test questions, and on the range of talent in the group being tested.
23. The best statistical evidence of the quality of a test of competency is its reliability coefficient.

Non-Cognitive Assessment

24. Non-cognitive characteristics include interests, attitudes, values, other traits of personality, and psychomotor skills.
25. It is practically impossible to obtain valid measures of a person's non-cognitive characteristics from a paper and pencil test.
26. It would be hard to defend the use of measures of non-cognitive characteristics as part of a process of selection for certification or licensure.

Test Construction

27 Those who prepare tests of competence should:

a) Be themselves outstandingly competent in the field
b) Be skilled in expressing ideas concisely and clearly.
c) Be guided by professional advice on how to write effective test items.
d) Be willing and able to take time to do the job well.

28. A committee of experts appointed to design and build a test of competency will work most effectively if guided and supported by specialists in test construction.

29. Standard item forms are wholly adequate for the development of excellent tests of competence.
30. Good test questions deal clearly and concisely with matters of fact.
31. The substantial cost of preparing a good test of competence is part of the price of professional excellence

Livingston agreed enthusiastically with many of the propositions. But he also reacted to several of them with skepticism or disbelief. The questions he raised, the views he expressed, and Ebel's reaction to those views are set forth in the remainder of the article.

Dialogue

Proposition 3

"The major component of professional competence is verbal knowledge.

Livingston:

As I understand the term, "verbal knowledge" means knowledge that can be expressed in words by the person who has that knowledge. Many health professionals have a great deal more professional competence than they can express in words. For example, an x-ray technologist must be able to place a patient in the correct position for the prescribed x-ray exposure. Being able to name the correct position is not sufficient. Being able to describe the correct position in words is not necessary. Being able to recognize a verbal description of the correct position is neither necessary nor sufficient. And when x-ray technologists are taught to position patients, the teaching is not primarily verbal; it is "hands-on."

I have developed performance tests in x-ray technology and in dental assisting and dental hygiene. Even the instructors in these fields find it difficult to express their practical knowledge in words (which is one of my functions in the test development process). I suspect that a similar situation (with respect to verbal and nonverbal knowledge) exists in many other health professions also.

In many professions the knowledge and skills that are most important are *not* verbal. They can sometimes be translated into verbal terms (with varying degrees of facility by different people) but as used on the job they are primarily non-verbal. Verbal knowledge is often not sufficient. A surgeon needs to know not only the name of the diseased organ and its condition; he must recognize it and its condition by sight and touch and must have the manual skills to perform the necessary correction. Blindfold him and tie his hands behind his back; his verbal knowledge will be as complete as ever, but he will be useless as a surgeon (though he may be of some use as a surgical consultant!)

An important part of my job is the development of performance tests and other behavioral measures. I work with experts in occupational fields. One of my main functions in this activity is to translate their practical knowledge into verbal terms. In the process I acquire a fair amount of verbal knowledge, but very little practical knowledge. Your proposition 13 ("knowledge is necessary, but not a sufficient condition for effective performance") seems to contradict proposition 3 to some extent. I like proposition 13 better. Some knowledge is always necessary. But in some occupations, the amount of knowledge required may be overshadowed by the skills involved. And not all knowledge is verbal knowledge. (How much of a symphony conductor's knowledge is verbal? Or a diamond cutter's?)

Ebel:

I agree that professionals, and all of us, have a great deal of knowledge that does not consist of verbal propositions, and that can be expressed only imperfectly in words. But for teaching and testing knowledge, the imperfect expression may be almost all that is available to us. Occasionally a diagram or picture, or even a live demonstration, may helpfully supplement our verbal descriptions. But, when we are attempting to impart or assess knowledge, the main burden of communication must be carried by words, I believe. With skilled writers and readers, speakers and listeners, and on many subjects, the imperfections are far outweighed by the efficiency and flexibility of verbal communication. Perceptual-motor skills, so important to the dental hygienist, the juggler and the concert pianist are another matter. To assess competence in those skills there is no adequate substitute for a performance test, obviously.

How much of the competence of a typical professional depends on the verbal knowledge he or she has, in contrast to perceptual or psychomotor skills? I know of no way in which a conclusive answer to this question, based on hard evidence, could be obtained. It might be easier to get hard evidence on a related question. How much time do students in medical schools spend acquiring knowledge as opposed to developing skills? My rough guess is that at least 75%, perhaps as much as 90% of the time is spent acquiring knowledge that can be expressed in words.

Some people do not value verbal knowledge highly, perhaps because they have had difficulty in acquiring it and do not possess much of it. These people are likely to say that words are less important than deeds, and to suggest that the relation between verbal knowledge and on-the-job performance is likely to be low.

I think they are wrong. It is hard for me to imagine a physician capable of treating a particular patient's ailment successfully who would be unable to describe in words the process of diagnosis and treatment. Granting the physician's need for certain perceptual and motor skills, I find it hard to imagine such a professional who can describe in words what should be done but still be unable to do it. The more verbal knowledge my physician has relevant to any disorders that afflict me, the safer I feel in his or her hands.

Is there really a contradiction between 3 and 13? A component can be a major component (3) without being sufficient in itself to do the whole job (13). I agree that competence in some professions (e.g. concert violin playing) is almost totally psychomotor. Such cases are a small minority, I believe.

Proposition 4

"Written tests can provide effective assessments of a person's verbal knowledge."

Livingston:

I agree that written tests can provide effective assessments of a person's verbal knowledge. The problem is that in the world of testing, "written test" too often means "multiple-choice test". The crucial difference is the prompting; it is much easier to recognize a correct

response than to supply one But in many real-world situations the options are not laid out clearly before us. Multiple-choice tests allow incomplete knowledge to masquerade as complete knowledge

Often a person in a job or situation will neglect to do something because he or she just did not think of it By presenting the correct action as one of a series of options, we remind the examinee of something he or she may not have remembered if it had not been presented. I have seen many test items in the health professions which I could answer, despite my lack of training in the relevant fields, *only* because the options were presented If the correct answer had not been presented, I would not have been able to supply it.

Incidentally, I am not at all sure that we should be as firmly committed as we seem to be to finding four alternative answers to each multiple-choice question. For the past two years I have been trying (unsuccessfully) to persuade my Educational Testing Service colleagues of the value of two-choice items for questions that are essentially dichotomous Our current practice is either to write an additional two or three distracters or to combine two or more two-choice items into a single four-choice or five-choice item. In the first case, we often end up with an item that actually tests for fine distinctions that do not reflect the original purpose of the item. In the second case, we throw away good information about the examinee by failing to score each piece of knowledge separately.

Ebel:

Research has shown over and over that the correlation between multiple-choice test scores and scores on any other means of measuring the same achievement are as high as the reliabilities of the two methods of measurement will allow. Unless one regards the absolute level of the score as dependable and important, one will not find multiple-choice test scores misleading.

Does the real world provide us with clearly laid out options? Often it does. It does to the voter, the investor, the umpire, the shopper, the home buyer, the mail sorter, the file clerk, and legislator, the judge, and a host of other decision makers. Even when it does not, the process of discovering and laying out of the options is seldom as difficult or as crucial an element in a wise decision as is the choice among them.

Multiple choice items do indeed help the examinee by offering prompts. Without doubt this increases the probability that a correct response will be given. If it were necessary to know for sure that the examinee could think of the answer to that particular question all on his own with unaided recall, then the prompting would tend to invalidate the item. But if the function of an item is to serve as one of a multitude of probes of the extent and depth of the examinee's structure of knowledge, then the prompt does no harm It does not give the answer away to the uninformed. It simply helps the informed. That help seldom, if ever, spoils the reliability of a test by making it too easy. If the prompting offered by a multiple-choice test consistently harms that test as a measure of achievement or aptitude, it ought to be possible to demonstrate the harm with empirical evidence. I know of no such evidence.

Indeed, one of the benefits of the "prompting" offered by multiple-choice test items is to define the examinee's task more fully and specifically than can be done with an open-ended question. There is less room for the happenstance of capricious recollections or chance ideration to spoil the precision of the measurement. Open-ended questions tend to yield less reliable scores than multiple-choice scores. This is due in part to uncertainties in scoring But it is also due in part to errors introduced by lack of precise definition of the test in the question, and by the examinees' good or bad luck in happening to think of the best interpretation, or procedure, or answer to give. On balance it seems to me that multiple-choice prompting is likely to do more good than harm.

You surely are on the right track in pushing for more use of two-choice items. I have been doing the same. In teaching students to write good true-false items I urge them to think of such items always in parts. For example:

1. An eclipse of the sun occurs when the moon is new (T)
2. An eclipse of the sun occurs when the moon is full (F)

I have recently been experimenting with a combined two-choice form, to compare it with the usual true-false. Here is an example.

3. An eclipse of the sun occurs when the moon is (a. full, b new). Though I have long defended true-false item form, the results of recent tests show that the two-choice form is better. The contortions item writers sometimes go through to adapt two-alternative problems to the four-alternative form are often wonderful to behold. What is worse, and potentially more harmful, is that the need to offer four alternatives leads item writers to avoid questions for which there are many, for which only two reasonable alternative answers exist. We have over-estimated the harm that guessing is likely to do on two-choice items. Your arguments in favor of such items are sound, and ought to be persuasive.

Proposition 8

"Criteria of competency in practice of a profession tend to be norm-referenced."

Livingston:

I agree that standards of competence in most professions tend to be norm-referenced. But should they be? Is it fair to deny a person the chance to practice a profession simply because enough other people are better at it? And couldn't there ever be a situation in which the public needs to be protected against the level of competence (or incompetence) represented by the average practitioner?

Professional standards need not be either purely norm-referenced or purely criterion-referenced. As the supply of persons available to do a job increases, it makes sense to increase *both* the number of persons credentialed in the job and the required level of proficiency. However, there may be an absolute minimum standard, below which it is better to leave a job undone. As an example of this last point; suppose we had a valid test for air traffic controllers, and an applicant has scored at the chance level. It would be better to close down the airport than to let him direct takeoffs and landings.

Ebel:

The public interest is best served I believe, by certifying a sufficient number of the best, not by certifying all (or only) those judged adequate on some basis or other. If good workmen are not available to do a necessary job, we must make do with some not so good.

The example you gave of the air traffic controller suggests that there are situations in which no help is preferable to incompetent help. I agree. But I suspect that such situations are not common. Much depends on the circumstances of the particular situation. Can we define

REPORTS AVAILABLE

Back issues of *Measurement in Education* are available for $2.50 each per single issue and for 75¢ each in quantities of 25 or more for a single issue

Vol 1,	No 1	*Helping Teachers Use Tests* by Robert L Thorndike
	No 2	*Interpreting Achievement Profiles — Uses and Warnings* by Eric F Gardner
	No 3	*Mastery Learning and Mastery Testing* by Samuel T Mayo
	No 4	*On Reporting Test Results to Community Groups* by Alden W Badal & Edwin P Larsen
Vol 2,	No 1	*National Assessment Says* by Frank B Womer
	No 2	*The PLAN System for Individualizing Education* by John C Flanagan
	No 3	*Measurement Aspects of Performance Contracting* by Richard E Schutz
	No 4	*The History of Grading Practices* by Louise Witmer Cureton
Vol 3,	No 1	*Using Your Achievement Test Score Reports* by Edwin Gary Joselyn & Jack C Merwin
	No 2	*An Item Analysis Service for Teachers* by Willard G Warrington
	No 3	*On the Reliability of Ratings of Essay Examinations* by William E Coffman
	No 4	*Criterion-Referenced Testing in the Classroom* by Peter W Airasian and George F Madaus
Vol 4	No 1	*Goals and Objectives in Planning and Evaluation — A Second Generation* by Victor W Doherty and Walter E Hathaway
	No 2	*Career Maturity* by John O Crites
	No 3	*Assessing Educational Achievement in the Affective Domain* by Ralph W Tyler
	No 4	*The National Test-Equating Study in Reading (The Anchor Test Study)* by Richard M Jaeger
Vol 5,	No 1	*The Tangled Web* by Fred F Harclerod
	No 2	*A Moratorium? What Kind?* by William E Coffman
	No 3	*Evaluators, Educators, and the Public — A Detente?* by William A Mehrens
	No 4	*Shall We Get Rid of Grades?* by Robert L Ebel
Vol 6,	No 1	*On Evaluating a Project — Some Practical Suggestions* by John W Wick
	No 2	*Measuring Reading Achievement — A Case for Criterion — Referenced Testing and Accountability* by S Jay Samuels and Glenace E Edwall
	No 3	*Security in a Citywide Testing Program* by Anthony J Polemeni
	No 4	*Dear Mama — Why don't they love me anymore?* by Thomas J Fitzgibbon
Vol 7	No 1	*The Discrepancy Evaluation Model (Part I)* by Andres Steinmetz
	No 2	*The Discrepancy Evaluation Model (Part II)* by Andres Steinmetz
	No 3	*What Graduate and Professional School Students Think About Admission Tests* by Leonard L Baird
	No 4	*The Paradox of Education Testing* by Robert L Ebel
Vol 8,	No 1	*FACE UP — A Framework for Assessing Career Educators Utilization Programs* by Bruce W Tuckman
	No 2	*Grade Equivalent Scores* by Gary Echternacht and *If Not Grade Equivalent Scores — Then What?* by Jeanne M Plas
	No 3	*Criticisms of Testing — How Mean is the Median?* by Henry S Dyer
	No 4	*The Superintendent and Testing — Implications for the Curriculum* by Herbert C Rudman
Vol 9,	No 1	*A Consumer's Guide to Criterion — Referenced Test Item Statistics* by Ronald C. Berk
	No 2	*Perspective on Intelligence Testing*, by Roger T Lennon
	No 3	*Bias*, by Eric F Gardner
	No 4	*The Use of Tests in Admissions to Higher Education* by Mary Fruen
Vol 10,	No 1	*Some Comments on Criterion-Referenced and Norm-Referenced Achievement Tests* by William A Mehrens and Robert L Ebel
	No 2	*Does the Rasch Model Really Work? A Discussion for Practitioners* by R Robert Rentz and Charlotte C Rentz
	No 3	*A Perspective On The Uses and Misuses of Standardized Achievement Tests* by Peter W. Airasian
	No 4	*The Practical Side of Criterion-Referenced Test Development* W James Popham and Elaine Lindheim
Vol 11,	No 1	*The Courts and Your Minimum Competency Testing Program — A Guide to Survival* by Thomas H Fisher
	No 2	*The Anatomy of a Scholastic Aptitude Test* by Roger T Lennon
Vol 12,	No 1	*Attitudes Toward Standardized Tests Students, Teachers and Measurement Specialists* by Frank P Stetz and Michael D Beck

minimum competence in such a way that an air traffic controller just below the minimum should be forbidden to try to ever help any planes to land regardless of how urgent the need?

Proposition 10

"Criterion-referenced tests can be evaluated using the same statistical procedures that were developed for norm-referenced tests."

Livingston:

I flatly disagree with the notion that the statistical procedures developed for norm-referenced tests can be used to evaluate criterion-referenced tests. Many of these statistical procedures process only the relative information contained in the test score. They are based on deviations from the group mean. If criterion-referenced testing means anything, it means that the absolute level of a test score is important information that should not be disregarded.

It often happens that a test will discriminate much better at some levels of ability than at other levels. Conventional statistics fail to take this fact into account. Suppose we have an examinee population such that most examinees are well above "minimum acceptable proficiency". Suppose we have two tests, one constructed to discriminate best in the range of ability that would be described as "minimum acceptable proficiency", the other constructed to discriminate best in the higher range where most examinees' abilities lie. Conventional test statistics would make the second test appear better than the first. But if our purpose in testing is to discriminate among examinees who have a least "minimally acceptable proficiency" from those who do not, we would do better to use the first test.

Ebel:

Except in closed and very limited universes of knowledge (e.g. the 100 basic facts of addition or the correct spelling of words on a prescribed list) it is impossible to obtain a score that has valid absolute meaning. In other cases the apparent absolute meanings are really relative to the subjective and more or less arbitrary standards of the test constructor or the test scorer. Such standards tend to be inconsistent from test to test and therefore undependable. Nor are there many

instances in the assessment of human performances where it is important to know the absolute level of a test score. More is almost always better, and we make do with the best we can get

It is true as you say that one could build Test A so that it would yield more reliable scores over all than Test B but less reliable pass-fail distinctions at a particular score level, if Test B is designed specifically to discriminate at that level. But would a sensible person use Test A to do the job that another test was designed to do specifically? And of several tests designed to do Test B's job, would not the most reliable of them result in the most dependable discrimination? Finally, is it because of the limitations of "conventional test statistics" that Test A, designed to one job, yields more reliable scores than Test B, designed to do quite another job?

Proposition 18

"Statistical validation of tests of competency is seldom feasible."

Livingston:

To argue that statistical validation of tests of competency is seldom feasible is to take a defeatist position. We could be doing more about this sort of thing than we do. In many cases the criteria would be rare events — critical incidents of various types. But statisticians in various fields have developed and are developing techniques for using those kinds of data. A few years ago you could have said, with as much justification, "Statistical determination of the causes of cancer is seldom feasible."

Ebel:

The position may be defeatist, but on the record of experience it seems to me to be clearly true. The reasons why it is true seem to me to make future success unlikely. If experience and reason teach me that I have been wrong, I will recant.

Epilogue

Accurate assessments of professional competence are essential to the effectiveness of a profession and to the welfare of a society. At the lower end of the scale such assessments are used to afford or to deny the opportunity of practicing the profession. At the upper end they grant or withhold highly valued certificates of special excellence. Concern for the quality of assessments of professional competence is surely justifiable.

Specialists in testing agree on many of the criteria of quality for tests used to measure competency. On some issues, however, their opinions differ. These differences are inevitable, given the complexity of the problems and the limits of our knowledge. Examination of different points of view on the issues is helpful in adding to our understanding of them, and ultimately to resolving them.

In this article Livingston and Ebel directed special attention to the shortcomings and virtues of verbal knowledge, multiple-choice items, norm-referenced tests, conventional test statistics, and test validation. Livingston is more convinced of the shortcomings of the first four and the virtues of the fifth than is Ebel. But despite their differences on these issues, they agree on a common objective of their efforts. For want of a better term that objective may be called psychometric excellence.

To those to whom differences of opinion are disturbing and distasteful, this thought may be reassuring. Given the task of measuring competence in a particular profession, the tests that Livingston and Ebel would help competent professionals to produce might be hard to distinguish.

URGENT NOTICE REGARDING NCME 1982 ANNUAL MEETING

IN THE EVENT OF A CANADIAN POSTAL STRIKE, SUBMIT PROPOSALS TO: Ross E. Traub c/o M. Slakter, 309 Baldy Hall SUNY at Buffalo, Buffalo, N.Y. 14260

IF IN DOUBT, CALL (416) 923-6641, ext. 460

NCME NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

USPS 823120
1230 17th Street, N.W
Washington, DC 20036

Second class postage paid at Washington, D C and additional mailing points