DOCUMENT RESUME

ED 065 555 TM 001 700

AUTHOR Millman, Jason

TITLE Passing Scores and Test Lengths for Domain-Referenced Measures.

PUB DATE Apr 72

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *Academic Achievement; *Criterion Referenced Tests; *Models; Performance Criteria; Scoring; Standard Error of Measurement; *Test Construction; *Testing

ABSTRACT

Procedures for establishing standards and determining the number of items needed in criterion-referenced measures are reviewed. The discussion of setting a passing score is organized around five factors: performance of others, item content, educational consequences, psychological and financial costs, and measurement error. Classical test theory, binomial, and sequential models for determining test length are considered. An illustrative table relating test length, proficiency standard, and required accuracy is provided. (Author)

29.11

PASSING SCORES AND TEST LENGTHS FOR DOMAIN-REFERENCED MEASURES

Jason Millman

Cornell University[1]

Abstract

Procedures for establishing standards and determining the number of items needed in criterion-referenced measures are reviewed. The discussion of setting a passing score is organized around five factors: performance of others, item content, educational consequences, psychological and financial costs, and measurement error. Classical test theory, binomial, and sequential models for determining test length are considered. An illustrative table relating test length, proficiency standard, and required accuracy is provided.

In recent years there has been much attention given to criterion-referenced measures which relate test performance to absolute standards rather than to the performance of others. Popham and Husek (1969) provide a readable account of the differences between such measures and the more traditional norm-referenced tests. The purpose of this paper is to synthesize much of the literature on establishing standards and determining the number of items needed in criterion-referenced measures.

This paper is written from the following perspective. A domain (i.e., population) of dichotomously scorable test "items" is conceptualized. This population of items need not actually exist. What is important, though, is that it is described well enough so that a relatively high degree of agreement can be reached about which kinds of items are or are not members of the population. In practice, only a reasonably representative sample of items is required.[2]

The items of a domain may be heterogeneous in content, form and difficulty. In practice, however, they should be measures of a limited number of skills and knowledges so that it makes sense to establish a single proficiency standard.

Of interest is the proportion of such items a student can pass. It is assumed that some educational decision, e.g., the nature of subsequent instruction for the student, is conditional upon whether or not he exceeds a proficiency standard when administered a sample of items from the domain. Thus, attention is directed toward the individual examinee and his performance relative to the standard rather than toward producing indicators of group performance.

PASSING SCORE

"The establishment of standards of achievement...is exceedingly complex and subjective.... (It) is a task not to be attempted lightly." (Science Research Associates, 1966, p. 16) The frequent practice of employing a particular passing score (say, e.g., 65%) only on the grounds of tradition is difficult to defend, in part because it seems unreasonable to require the same level of proficiency for all domains and all individuals and in part because there are other sources beside tradition which should be considered when determining a standard. The discussion below of several such sources and practices does not conclude with a single recommended method for calculating a passing score, but rather the intention is to alert the reader to information and procedures which should be considered when a standard is being established.

None of the procedures eliminates the need for judgment. The focus of this rational thought shifts when each of the following five sources of information is utilized.

Performance of Others

One procedure which has a degree of rationality is to set the passing score such that a predetermined percent of students pass. Test construction suggestions in this situation have been provided (see, e.g., Tinkelman, 1971). Whether an individual passes under this scheme depends, in part, on the general competence of the others taking the test. This procedure is most applicable when the number of people who can or should be given some treatment or "certification" is fixed and the assessment task is to select the ablest examinees.

Item Content

One source to turn when deciding upon a passing score is the test items themselves. Each item in the test can be inspected and a judgment made about how important it is that it be answered correctly. Such a procedure is the following:

> Suppose that a standard (not to be confused with "standard scores") of satisfactory performance is to be established for the twelfth grade. The first step would be to study with care each of the individual problems...and to decide, in terms of one's own subjective notions of the adult needs of the typical high school graduate, how many of these problems the typical beginning senior _should_ be able to solve. In other words, one would have to decide subjectively what _raw_ score on this test the typical senior _ought_ to make or exceed. (Science Research Associates, 1966, p. 16)

A similar suggestion, with a probabilistic variant, has been offered by Angoff (1971).

> ...ask each judge to state the _probability_ that the "minimally acceptable person" would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (p. 515)

Alternatively, a decision might be made that to pass the test an acceptable answer must be given to all the items in one group, that some fraction of the items in a second group must be answered correctly, and that only a smaller fraction of the remaining items need be answered in an acceptable way. A plan somewhat similar to this, but which leads to a single passing score, is offered by Ebel (in press).

A different procedure for converting judgments about the acceptability of each option of each item to a passing score and letter grades has been described within a classroom context almost 20 years ago by Leo Nedelsky (1954). A procedure similar to Nedelsky's is being used in certifying candidates as eligible for the degree of Doctor of Medicine at the University of Illinois College of Medicine. (Crawford, 1970) A "Minimum Pass Level" for multiple-choice items is constructed as follows. Each option of each item is scrutinized and the options (including the keyed one) which a barely passing student might experience difficulty in discriminating are identified. Let $\underline{o}_i$ be the number of such options in item $\underline{i}$ and $\underline{O}_i$ be the total number of options in item $\underline{i}$. The required passing score is then equal to the fraction $\underline{o}_i/\underline{O}_i$ summed over all items. A student's score is merely the total number of items he answers correctly.

A weakness of this system is that it gives a premium to an examinee for knowing (not guessing) the keyed option to, say, $\underline{n}$ items. That premium is an opportunity to choose, for each of $\underline{n}$ other items, a foil which a barely passing student should not experience difficulty in eliminating and to still be above the passing level. In fields like medicine, the system permits a student who knows the correct answer to $\underline{n}$ items to select "dangerous" options to another $\underline{n}$ items and still not fail the test. One wonders why all tough-to-discriminate options are not treated as correct options and a passing score determined in a manner similar to those suggested in the opening paragraphs of this subsection.

Educational Consequences

A primary consideration in setting a required level of proficiency is its effect on future learning. If the level is set too low, students will be given instruction on new concepts and skills which they cannot handle effectively. If the level is too stringent, efficiency is reduced as students spend needless time on remedial materials. Very helpful would be any data on what happens when students with differing degrees of proficiency in a given knowledge domain are subjected to the alternative instructional sequences.

Procedures for determining passing scores when test and criteria data are both available are provided by Bormuth (1971). Although Bormuth's research involved concurrently available criteria, the regression procedures are generalizable to the situation in which performance data are acquired after the alternative educational decisions have been implemented.

In absence of data about educational consequences, the following guideline can be offered. If, on the basis of a logical analysis of the subject matter and the extant instructional system, the knowledges and skills are seen as fundamental or prerequisite to future learning, then a high proficiency level should be required. A lower passing score can be tolerated when the material is not seen as completing a necessary link in the development of some more complex concept or skill, especially if the ideas will be covered again in the curriculum. Application of solely this guideline would probably result in higher passing scores for tests on "basic" topics in mathematics than for tests covering "units" in social studies or English grammar. Tests of performances not viewed as prerequisite for future learning probably should not have passing scores and not be criterion-referenced. (Garvin, 1971)

Psychological and Financial Costs

All things being equal, a low passing score should be used when the psychological and financial costs associated with a remedial instructional program are relatively high. That is, there should be fewer failures when the costs of failing are high. These "costs" might include lower motivation and boredom, damage to self-concept, and dollar and time expenses of conducting a remedial instructional program. A higher passing score can be tolerated when the above costs are not too great or when the negative effects of moving a student too rapidly through a curriculum (i.e., confusion, inefficient learning, etc.) are seen as very important to avoid.

Emerick (1971) and Kriewall (1969) have reported procedures which utilize, at least indirectly, the ratio of the two kinds of costs in arriving at a passing score. Unfortunately, Emerick's model employs some very restrictive assumptions which makes it inapplicable to the situation being considered here and of limited value. Specifically, the domain of items are viewed as "highly homogeneous in terms of content, form, and difficulty level." (p. 322) Further, "an examinee can occupy only one status with respect to the skill being tested: mastery or nonmastery." (p. 322) Thus, when a student misses an item it is assumed to be the result of measurement error rather than partial knowledge. We agree with Ebel (1971) that "abilities, understandings, and appreciations are in the experience of almost everyone, not all-or-none adaptations. They are matters of degree. None but the simplest of them can ever be mastered completely by anyone." (p. 287)

Kriewall's model is applicable when the student has partial knowledge. It does not employ the restrictive and impractical equal item difficulty assumption, although Kriewall himself and Besel (1971) appear to argue otherwise. Kriewall's model is similar to Emerick's in that given numerical values of the degree to which certain costs or errors of classifying students will be tolerated and the length of the test, a passing score can be computed which minimizes the costs (errors). A comparison of the Emerick and Kriewall approaches is provided by Besel (1971).

Measurement Error

There is a systematic error introduced in estimating an examinee's proficiency when the test item format allows a student to answer items correctly by guessing. The passing score could be raised to take into account the expected contribution attributed to pure guessing. Alternatively, each student's score could be adjusted according to the standard correction-for-guessing formula and this adjusted score compared to the standard. Since pure, random guessing occurs rarely, adjusting either the examinee's score or the standard, as described above, will be expected to control the guessing contribution only partially. (Emerick's procedure mentioned above also takes into consideration a guessing factor; Kriewall's does not.)

An additional error in the measuring process is expected when, for reasons of difficulty of construction, inconvenience of administration, or ignorance, the variety of types of questions and content represented in the domain are not used in the test. When the test items are thus suspected to be unrepresentative, it is well to raise or lower the standard an additional amount in order to protect against the misclassification error (examinee passes when he should fail, examinee fails when he should pass) feared the more.

Even a student's corrected-for-guessing percent score on a random sample of test items will usually not equal the true proportion of all the items in the domain to which he "really" knows the answer. This expected random measurement error can be reduced by using more test items. The relation of test length and proficiency standard to the accuracy in classifying students will be dealt with in the next part of this paper.

TEST LENGTH

Recall that this paper is written from the perspective that a domain of dichotomously scorable test "items" is conceptualized and that an estimate of the proportion of such items an examinee can answer correctly is desired.

> Rather than sample problem solving behavior across a hypothetical population of pupils, it is more appropriate to measure the individual's behavior on a random sample of problems drawn from a clearly defined population of problems. The individual's relative score on this sample can then be interpreted as an estimate of his proficiency relative to that class of problems. (Kriewall, 1969, p. 37)

In this context, the test length problem is determining the size of such a sample of problems needed to acquire an estimate having a specified level of accuracy.

Classical Test Theory

The classical test theory approach to determining accuracy and test length makes use of the standard error of measurement. It follows from the assumptions of the model that the standard error of measurement is constant for all true scores, a condition that probably is not true in practice. Further, in order to convert numerical values of standard errors into probability statements dealing with score accuracies, it is necessary to know or make assumptions about the error distributions. The usual normal distribution assumption is most vulnerable in those situations often found with domain-referenced tests; namely, when the number of items used to measure a particular proficiency is small and when the performance standard approaches the ceiling of the test. Finally, the value of a standard error for a test depends upon the group of examinees on whom the test was administered, and this is in conflict with the context of the problem as described above.

Although excellent in many respects, the Livingston (1972) approach to the reliability of criterion-referenced tests does not help overcome these problems. The standard error of a test is the same regardless whether the classical or criterion-referenced reliability coefficient is used. (Harris, 1972) Remaining are the limitations of the standard error of measurement in determining the accuracy of a test score and,in turn, the required test length.

Binomial Model

If one assumes that the proficiency test represents a randomly selected set of 0-1 scored items from some domain of tasks, and if one further assumes that the experience of taking the earlier items on the test does not influence the examinees chances of passing the later items, then an exact solution to the number-of-items-needed problem is given by the binomial distribution (for infinite or very large domains) or by the hypergeometric distribution (for relatively small item domains). No assumption regarding item homogeneity (in content or difficulty) is needed. (See Lord and Novick, 1968, section 11.9, for distribution statistics associated with this model.)

No group measures or item indices computed over examinees are utilized in this binomial conceptualization. Rather, the items which an examinee can pass and those the individual fails are analogous to two colors of balls in an urn. Continuing the analogy, the test length question is, How many balls must be sampled (items administered) so that the percent of all balls in the urn of a given color (test items in the domain answered correctly) can be estimated accurately? The urns associated with other examinees are of no concern.

Tables relating test length to accuracy for a given passing score have been constructed by Millman (1972) using this model.[3] Table 1 displays the relevant data when an 80% passing score is selected.

To illustrate the use of Table 1, suppose that an educator is willing to tolerate a 25% misclassification error (i.e., classify as "pass" a student who does not know 80% of all the items or vice versa) for those students who actually know 70% or 90% of the items. Reading down the 70% and 90% columns, note that roughly 25% errors (actually 26% and 19%) will occur if a random sample of eight items from the domain are used and a passing score of seven imposed.

Table 1 about here

Sequential Models

Sequential testing procedures in which the number of items ultimately given to a student depends upon the closeness of his performance relative to a passing percentage have been suggested (Kriewall, 1969; Ferguson, 1970). These schemes are based upon earlier work by Wald (1947). Sequential testing can also be conducted within a Bayesian framework. Examples how this might be done are provided by Powers (1971).

The primary advantage of employing such models is that fewer test items, on the average, are needed to acquire a given overall level of accuracy. Such procedures are most feasible when examinees interact with computers during testing. When paper and pencil tests are used, it would appear more efficient to administer all students a somewhat more generous, but equal, number of test items.

FOOTNOTES

[1] This investigation was supported in part by the Instructional Objectives Exchange while the author was on leave from Cornell University. The reactions of Wells Hively and Paul Cieslak to an earlier draft of this paper are gratefully acknowledged.

[2] In contrast to "criterion", the term "domain" is reserved for the case when an item generation procedure is employed or a universe is postulated and the items used are considered to be a representative sample. Cronbach et al. (1963) used the notion of a universe of test items in a theory of reliability. Osburn (1968) seems to have made the first in-print statement of the precise definition of domain.

One set of procedures for generating such a domain makes use of item forms. A thorough description of the techniques and examples may be found in Maxwell et al. (1971) with further descriptions and extensions in Hussell (1969) and Rabehl (1971). Other item generating schemes have been proposed by Anderson (undated), Bormuth (1970), Guttman (see, e.g., Jordan, 1971), and to some extent by Scandura (see, e.g., Durnin and Scandura, 1971).

A less formal way of conceptualizing a domain of items is to list the specific instructional objectives included within the domain in a manner such that it becomes evident what items will be included in and excluded from the domain. This latter procedure was followed in the development of the revised collections of objectives in the basic skill areas published by the Instructional Objectives Exchange, P.O. Box 24095, Los Angeles, 90024.

[3] Also considered in the Millman reference is the mathematically parallel problem of determining the number of examinees needed to measure accurately the proportion of all students able to answer a given item correctly.

TABLE 1

Student Assessment Minimum Passing Percent 80

PERCENT OF STUDENTS EXPECTED TO BE MISCLASSIFIED

Passing Score		No. of Test Items	STUDENT'S TRUE LEVEL-OF-FUNCTIONING*							
			40	50	60	70	75	85	90	95
1	out of	1	40	50	60	70	75	15	10	5
2	out of	2	16	25	36	49	56	28	19	10
3	out of	3	6	13	22	34	42	39	27	14
4	out of	4	3	6	13	24	32	48	34	19
4	out of	5	9	19	34	53	63	16	8	2
5	out of	6	4	11	23	42	53	22	11	3
6	out of	7	2	6	16	33	44	28	15	4
7	out of	8	1	4	11	26	37	34	19	6
8	out of	9	-	2	7	20	30	40	23	7
8	out of	10	1	5	17	38	53	18	7	1
10	out of	12	-	2	8	25	39	26	11	2
12	out of	15	-	2	9	30	46	18	6	1
16	out of	20	-	1	5	24	41	17	4	-
20	out of	25	-	-	3	19	38	16	3	-
24	out of	30	-	-	2	16	35	15	3	-
32	out of	40	-	-	1	11	30	14	2	-
40	out of	50	-	-	-	8	26	12	1	-
48	out of	60	-	-	-	6	23	11	1	-
60	out of	75	-	-	-	4	19	9	-	-
80	out of	100	-	-	-	2	15	7	-	-

*The true level-of-functioning is the percent of items a student would be able to answer correctly if he were given the entire universe of items.

Students having true level-of-functioning values less than the minimum passing percent of 80 should fail a test composed of items from this universe. However, on any given test of finite length, some of these students will get over 80% of the items correct and be considered as "passers". The expected percent of such misclassifications are given in the body of the table to the left of the dotted line.

Students having true level-of-functioning values greater than the passing percent of 80 should pass such a test. The percent of these students who will be misclassified as "failures" are shown in the table to the right of the dotted line.

BIBLIOGRAPHY

Anderson, Richard C. How to Construct Achievement Tests to Assess Comprehension. Urbana: University of Illinois, undated. (mimeographed)

Angoff, William H. Scales, Norms, and Equivalent Scores. Pages 508-600 in Robert L. Thorndike (editor), Educational Measurement (second edition). Washington: American Council on Education, 1971.

Besel, Ronald. A Comparison of Emrick and Adam's Mastery-Learning Test Model with Kriewall's Criterion-Referenced Test Model. Inglewood, California: Southwest Regional Laboratory, Technical Memorandum 5-71-04, April, 1971.

Bormuth, J. R. On the Theory of Achievement Test Items. Chicago: University of Chicago Press, 1970.

Bormuth, John R. Development of Standards of Readability: Toward a Rational Criterion of Passage Performance. Final Report, U. S. Department of Health, Education and Welfare, Office of Education. Project #9-0237, June, 1971.

Crawford, William R. Assessing Performance When the Stakes are High. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March 1970.

Cronbach, Lee J., Rajaratnam, Nageswari, and Gleser, Goldine C. Theory of Generalizability: A Liberalization of Reliability Theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Durnin, John H. and Scandura, Joseph M. An Algorithmic Approach to Assessing Behavior Potential: Comparison with Item Forms and Hierarchical Technologies. Philadelphia: University of Pennsylvania, Structural Learning Series, Report No. 63, November 1971.

Ebel, Robert L. Criterion-referenced Measurements: Limitation. School Review, 1971, 79, 282-288.

Ebel, Robert L. *Measuring Educational Achievement*. (Revised Edition) Englewood Cliffs, N.J.: Prentice-Hall, (in press).

Emrick, John A. An Evaluation Model for Mastery Testing. *Journal of Educational Measurement*, 1971, 8, 321-326.

Ferguson, Richard L. Computer-Assisted Criterion-Referenced Measurement. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, Working Paper 49, March, 1970.

Garvin, Alfred D. The Applicability of Criterion-Referenced Measurement by Content Area and Level. Pages 67-75 in W. James Popham (editor), *Criterion-Referenced Measurement: An Introduction*. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Harris, Chester W. An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests. *Journal of Educational Measurement*, 1972, 9, 27-29.

Jordan, John E. Attitude-Behavior Research on Physical-Mental-Social Disability and Racial-Ethnic Differences. *Psychological Aspects of Disability*, 1971, 18, 5-26.

Kriewall, Thomas Edward. Application of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Livingston, Samuel A. Criterion-Referenced Applications of Classical Test Theory. *Journal of Educational Measurement*, 1972, 9, 13-26.

Lord, Frederick M. and Novick, Melvin R. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley, 1968.

Maxwell, Graham and others. Curriculum Evaluation in the MINNEMAST Project: A Case Study in Domain-Referenced Testing. Minneapolis: University of Minnesota, 330 Burton Hall, 1971. (mimeographed)

Millman, Jason. Tables for Determining Number of Items Needed on Domain-Referenced Tests and Number of Students to be Tested. Los Angeles: Instructional Objectives Exchange, Technical Paper No. 5, April, 1972.

Mussell, Bruce. The Behavior of Subject-matter Informants in Constrained Descriptions of Their Subject-matter. Unpublished doctoral dissertation, University of Minnesota, 1969.

Nedelsky, Leo. Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement*, 1954, 14, 3-19.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, 28, 95-104.

Popham, W. James and Husek, T. R. Implications of Criterion-Referenced Measures. *Journal of Educational Measurement*, 1969, 6, 1-9.

Powers, James E. Bayesian Statistics and Longitudinal Studies. *Association for Research in Growth Relationships Journal*, 1971, 12, 59-82.

Rabehl, George. The Experimental Analysis of Educational Objectives. Unpublished doctoral thesis, University of Minnesota, 1971.

Science Research Associates. *ITED the Iowa Tests of Educational Development: Manual for the School Administrator*. Chicago: Science Research Associates, 1966.

Tinkelman, Sherman N. Planning the Objective Test. Pages 46-80 in Robert L. Thorndike (editor), *Educational Measurement* (second edition). Washington, American Council on Education, 1971.

Wald, Abraham. *Sequential Analysis*. New York: John Wiley and Sons, 1947.