DOCUMENT RESUME

ED 106 347 TM 004 463

AUTHOR Graham, Darol L.; Bergquist, Constance

TITLE An Examination of Criterion-Referenced Test Characteristics in Relation to Assumptions about the Nature of Achievement Variables.

PUB DATE [Apr 75]

NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE

DESCRIPTORS *Academic Achievement; Achievement Tests; *Criterion Referenced Tests; Cutting Scores; Equivalency Tests; Intellectual Development; Measurement Techniques; *Models; Predictor Variables; Scores; Test Construction; *Testing; *Testing Problems; Test Reliability; Tests; True Scores; Verbal Development

ABSTRACT

Two models were identified for criterion-referenced tests, one based on the assumption of a continuous achievement variable and the other assuming a dichotomous or binary variable. Several test characteristics were examined and contrasted for the two models, including the distribution of scores, establishment of a cutting score, test length, item difficulty, and reporting of test information. In addition, the appropriateness of each model for measuring learning tasks involving verbal information or intellectual skills was discussed. (Author)

ED106347

AN EXAMINATION OF CRITERION-REFERENCED TEST CHARACTERISTICS IN RELATION TO ASSUMPTIONS ABOUT THE NATURE OF ACHIEVEMENT VARIABLES.

Darol L. Graham, Ph.D.
University of Texas Health Science Center at Dallas

and

Constance Bergquist
Florida State University

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the annual meeting of the American
Educational Research Association, Washington, April, 1975

TM 004 463

ERIC

An Examination of Criterion-Referenced Test Characteristics in Relation to Assumptions About the Nature of Achievement Variables.

Darol L. Graham, Ph.D.
University of Texas Health Science Center at Dallas

Constance Bergquist
Florida State University

Since introduction of the term "criterion-referenced" by Glaser (1963), a wide variety of definitions and interpretations of the term, as well as alternative terms for similar concepts, have appeared in the literature. Many controversies have arisen over various characteristics of criterion-referenced tests. Much of the disagreement can be traced to differences in underlying assumptions, often unstated, about the nature of the achievement variable being measured. Once these assumptions are made public, it often becomes evident that opposing proponents are discussing different situations and that both may be correct. Many of the differences concerning the nature and use of criterion-referenced tests can be abated by considering more than one type of achievement variable.

It is the contention of this paper that assumptions concerning the continuous or dichotomous nature of an achievement variable substantially affect the characteristics and use of a criterion-referenced test developed to measure the variable. It is further contended that different assumptions may be desirable for measurement of different domains of learning outcomes (Gagne, 1971). In particular, the assumption of continuity may be most appropriate in measuring verbal information outcomes whereas the assumption of dichotomy may be most appropriate in measuring outcomes described as intellectual skills.

The Nature of Achievement Variables

Problems have arisen in criterion-referenced measurement because of variation in the manner that different types of achievement can be demonstrated. As noted by Popham & Husek (1969), "Some criterion-referenced tests yield scores which are essentially 'on-off' in nature, that is, the individual has either mastered the criterion or he hasn't . . . more commonly, however, a range of acceptable performance exists [p. 7]." Unfortunately, these differences in types of observable performance are often ignored and tests of both types of performance are treated similarily.

Most criterion-referenced test users assume that achievement is distributed as a continuous variable and that all levels of proficiency relative to an objective can exist. This assumption was first expressed by Glaser (1963) in his discussion of a "continuum of knowledge acquisition ranging from no proficiency at all to perfect performance [p. 519]."

A few users of criterion-referenced tests consider achievement as a binary variable and assume that all examinees are either masters or nonmasters of a specified objective. For example, Emrick (1971) stated, "mastery of each unitary skill is assumed to be an all or none variable [p. 322]." Regardless of whether achievement is considered as a binary or continuous variable, nearly all test users attempt to dichotomize scores to provide mastery and nonmastery classifications of examinees.

It appears that most developers of criterion-referenced tests consider all types of human performance to be similar. Gagne (1974) suggested, however, that five different classes of performance are readily distinguishable from each other. If Gagne is correct, it may be appropriate to employ different measurement models with different types of learning outcomes. The following discussion focuses upon two of Gagne's domains, verbal information and intellectual skills.

According to Gagne and Briggs (1974) the verbal information domain encompasses the learning of labels, single facts, and organized information or knowledge. One might argue that single units of verbal information such as labels or single facts are recalled in an all or none manner. Even if this is true, the measurement of single units of information is probably a trivial operation in most instances. Seldom is a single unit of information considered of sufficient importance to be tested separately. More commonly, a collection of information, preferably interrelated to comprise a body of organized knowledge, is tested simultaneously. A collection of information forms a content domain from which items are randomly sampled. Performance of an examinee relative to the entire domain depends upon the number of discrete units of information that have been acquired and remembered. If it is assumed that achievement of each of the discrete units of information is demonstrated independently, any proficiency from 0-100% might be demonstrated on a test. Thus, achievement of verbal information measured by a domain-referenced test would be demonstrated as a continuous variable.

A stronger case can be established for the measurement of single intellectual skills than for the measurement of single units of verbal information. While a single verbal proposition represents only one behavior, a single intellectual skill encompasses an entire class of behaviors. If the research on learning hierarchies is valid, the intellectual skill may constitute a prerequisite for a number of other skills, whereas the verbal information may have limited utility for other learning. In addition, the measurement of a collection of intellectual skills may present serious scaling problems. If hierarchical dependencies exist, combining scores from different levels of the hierarchy may be analogous to adding feet and inches.

Since an intellectual skill defines an entire class of behaviors, a large number of parallel items could be generated to measure a single skill. Theoretically, a learner who acquires the intellectual skill would be able to demonstrate the entire class of behaviors while the learner who has not acquired the skill would be unable to perform any of the behaviors. Accordingly, achievement of an intellectual skill would be demonstrated as a binary variable.

In a previous paper, the author (Graham, 1974) adopted the terms competency test and proficiency test to differentiate between tests constructed to measure the two different types of achievement variables. The term competency test was used to describe a criterion-referenced test of achievement that is demonstrated as a binary variable, while the term proficiency test was reserved for a criterion-referenced instrument constructed to measure a learning variable which can be achieved to any degree. It seems appropriate to consider a continuum of proficiencies but only two states of competency, mastery and nonmastery. This restricted usage of the terms competency test and proficiency test is followed in the remainder of the present paper.

Basic Assumptions

The discussion above presents a case for two different criterion-referenced measurement models for the assessment of learning outcomes. A binary model would be necessary for the measurement of intellectual skills, while a continuous model would be more appropriate for assessing achievement of verbal information. Let us look briefly at the assumptions and corollaries of these two models.

Binary Model

The critical assumption in the binary model is that certain capabilities enable an individual to perform an entire class of behaviors, and if the capability is not acquired, the individual cannot perform any of the class of behaviors. Since a series of items sampled from a domain representing the class of behaviors are measuring the same learned capability, responses to the items are expected to be highly intercorrelated. Theoretically, true scores for individuals relative to the item domain representing the class of behaviors will be either zero or 100%. Deviations from these all or nothing scores are caused by measurement error and do not accurately reflect the true capability of the individual.

It was previously stated that achievement in the domain of human performance referred to as intellectual skills appears to provide an appropriate situation for application of the binary model. In the study noted earlier (Graham, 1974) the author employed a strict item-sampling model to generate domain-referenced tests of intellectual skills. The tests displayed the characteristics expected for a binary achievement variable. Horwitz (1974) and Bergquist and Horwitz (1975) also demonstrated the viability of the binary model with tests constructed to measure unitary, explicitly defined intellectual skills. Performance on tests constructed in these studies was essentially all or none resulting in high interitem correlations.

It might be useful to examine an example of a competency test of the intellectual skill domain. In developing a test to measure the skill of adding negative integers, Bergquist and Horwitz (1975) randomly selected 10 items from the total domain of addition problems comprised of two negative integers. The test was administered to 67 eighth grade students. More than 94% of the examinees scored outside the range 2-6 with approximately one-fourth of the students failing all items and approximately one-half of the students receiving perfect scores. It seems reasonable that scores falling in the middle of the range actually represented measurement error resulting from such factors as carelessness, fatigue, guessing, and cheating and that the students were either capable or not capable of adding two negative one digit integers.

ERIC
Full Text Provided by ERIC

Continuous Model

In many situations, achievement is expected to be demonstrated as a continuous variable. The model is based on the major assumption that certain learned capabilities exist for which only a single behavior can be demonstrated, and unless that capability is of major importance, it should be measured as part of a collection of behaviors comprising a larger domain. It is further assumed that performance of one capability is independent of performance of the other capabilities in the collection. Relative to a domain of independent capabilities the true score of an individual is determined by the number of individual capabilities that have been acquired and may assume any value from zero to 100% of the capabilities comprising the collection.

Domain-referenced tests of verbal information would warrant consideration of this model. Most educators have considerable familiarity with tests of verbal information. Even tests intended to measure achievement of intellectual skills are often constructed in such a manner that it is possible to provide correct responses through recall of related verbal information without actually demonstrating the skill of interest.

A typical example of a verbal information test can be drawn from the Physician's Assistant Program with which the author is associated. Trainees in the program are expected to learn 214 common medical abbreviations. It is possible for individual students to learn any number of abbreviations from the total collection and thus possess any true proficiency relative to the total collection. The proportion of correct responses provided by a student on a random sample of items from the domain of medical abbreviations would provide an unbiased estimate of the examinee's true proficiency with respect to the domain.

Implications

Many of the controversial issues concerning criterion-referenced measurement are given new perspective when considered in the context of alternative achievement variable models. In the discussion that follows, several characteristics of criterion-referenced tests are examined in relation to binary and continuous achievement variables.

Score Distributions

In the previous section, it was indicated that under the assumption of achievement as a binary variable, only two performance capabilities, mastery and nonmastery, are expected. Theoretically, true scores for all members of the mastery population are 100% while true scores for all nonmasters are zero. Deviation of observed scores from these two levels is attributed to measurement error. When such a test is administered to a group comprised of both masters and nonmasters, the scores would be expected to be distributed bimodally. In the studies by Graham (1974), Horwitz (1974), and Bergquist and Horwitz (1975), quite pronounced bimodal characteristics were obtained for score distributions on the tests of intellectual skills. In addition, Graham obtained a trimodal score distribution for a test constructed to measure two intellectual skills simultaneously.

The anticipated distribution of scores on proficiency tests would be quite different than for competency tests. Since all true proficiencies would theoretically exist, the score distribution on a given test administration would be determined by the level of attainment of the sample tested. With a large random sample of individuals in a traditional, time-based learning environment, scores would likely be normally distributed. On the other hand, pretest and posttest scores in a mastery learning situation would no doubt be highly skewed. A bimodal score distribution for a single administration of a proficiency test, however, would be highly unusual.

A major point of controversy in discussions of criterion-referenced test characteristics nas been the issue of score variance. The introduction of two achievement variable models does not directly address this issue. For competency tests, however, considerable score variability would exist except in the special case when only masters or nonmasters are included in the test sample.

Test Homogeneity

One of the most important implications of a dual concept for achievement variables concerns the homogeneity of an item set. Some advocates of criterion-referenced measurement believe that a test of a single behavioral objective should be homogeneous in form, content, and difficulty while others argue that a highly homogeneous test measures an overly restricted item domain. This controversy should be examined in reference to the alternative measurement models.

An information objective can be stated to describe a single behavior or a collection of behaviors. In most instances, the measurement of a single behavior is probably a trivial or at least an inefficient operation. It is usually advantageous to define a collection or domain of similar behaviors and to draw inferences about capabilities relative to the entire domain through item-sampling procedures. In this situation, item homogeneity would depend upon the similarity of the behaviors comprising the domain. To the extent that increasing the size of the domain would tend to exhaust the supply of similar behaviors, item homogeneity would be dependent upon domain size. Since performance on one item is assumed to be independent of performance on another, items would be expected to display a range of difficulty values. Thus, a test of a domain of verbal information would not necessarily be homogeneous in content or difficulty.

ERIC Full Text Provided by ERIC

On the other hand competency test items are not independent of each other. Since an intellectual skill domain defines a class of behaviors, each item provides a repeated measure of the same skill or behavior. Consequently, item homogeneity would be a necessary characteristic of a good competency test. Deviations from a high degree of item homogeneity indicate confounding of measurement with other skills or verbal information. A test that simultaneously measures more than one class of performance would not possess the characteristic described by Gagne (1968) as distinctiveness.

It is seldom possible to construct a test, or even a single item, which is so distinctive that it measures only one intellectual skill. The measurement of intellectual skills is always confounded with the simultaneous measurement of other capabilities. If all members of the test population have mastered the extraneous capabilities, however, the confounding does not interfere with measurement of the specific skill defined by an objective, and the test possesses the quality of distinctiveness.

There is at least one situation in which differential capabilities of examinees to perform supplementary skills cannot be detected. This situation exists whenever the supplementary skill or skills are uniformly required for all items in a test. An example is the need for prerequisite reading skills for solution of any verbally stated mathematics problem. In such situations, the supplementary skills should either be specified as part of the objective or should be measured independently in a separate pretest to ascertain their influence upon misclassification of certain examinees. Intensive investigation into the effects of measurement confounding and into appropriate means of handling this problem appears warranted.

Item difficulty values for a competency test are actually average difficulty values that depend upon the composition of the test sample. For such a test, item difficulty is actually a function of the learning state of the examinee. Hypothetically, the difficulty values for the mastery and nonmastery populations should be one and zero respectively. Thus, whenever an examination sample is comprised of both masters and nonmasters of an intellectual skill, the magnitude of the difficulty value for an item depends upon the relative representation of the two competency populations in the test administration sample.

In a discussion of reliability, Stanley (1971) demonstrated that the only time dichotomously scored items can be perfectly intercorrelated, resulting in the maximum value of one for KR-20, is when all items have equal difficulty. The author (Graham, 1974) repeatedly obtained KR-20 estimates of reliability well above 0.9 for 10-item tests of intellectual skills for which items were randomly generated. In the study, item-test correlation coefficients above 0.7 were the rule rather than the exception. Instances in which single items deviated in difficulty value from other items of a test could be explained by differences in the supplementary capabilities required for correct responses to the items. This investigation provided strong evidence for the binary nature of intellectual skill achievement.

Passing Scores

For a variety of reasons, educators often wish to establish a minimum standard of acceptable performance on a domain-referenced achievement test. Kriewall (1969) suggested that such standards should be formulated as part of the design specifications outlined during curriculum development. At the present time, Hambleton and Novick (1973) believe that, "the establishment of proficiency levels is primarily a value judgement [p. 163]." To assist in this judgement, Millman (1973) discussed five factors that should be considered in the determination of performance standards. Once a performance standard has been established, it must then be translated into a passing score for a given sample of items from the domain. Factors other than the performance standard that influence passing scores are test length and the relative seriousness of the two types of classification error.

For situations in which achievement is demonstra [illegible] binary variable, there is no need for establishing a performance standard. Since only two performance capabilities are assumed to exist, it is unnecessary to operationally define the mastery state. A passing score is established at a level that tends to minimize the number of examinees that are misclassified due to measurement error. Figure 1 presents the frequencies of scores obtained on a 10-item test administered by the author (Graham, 1974). With bimodal score distributions of this type, it is most convenient to establish a passing score simply by inspection. Since less than 15% of the examinees received scores in the range 1-8, selection

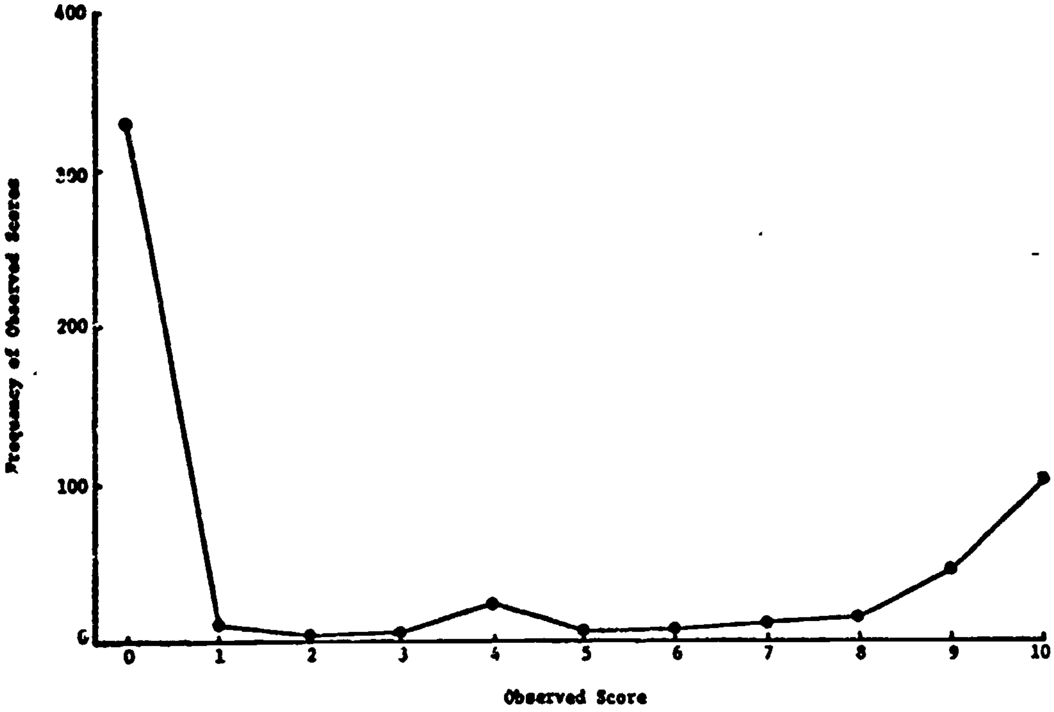

Figure 1. Distribution of obtained on a domain-referenced test of an intellectual skill (Graham, 1974).

of any score within this range as a passing score would not substantially alter the classification results. The nature of the consequences of misclassifying true masters or true nonmasters would influence the selection of a specific passing score within this range.

Test Length

The determination of test length is closely related to passing score and errors of classification. For situations in which a continuum of proficiencies is assumed to exist, Kriewall (1969) employed acceptance sampling procedures based upon the classical binomial model for establishing test length. Millman (1972) used the binomial model to construct tables that relate test length to classification accuracy for various passing scores. By assuming prior information about an examinee's level of functioning, Novick and Lewis (1974) introduced a more precise method of determining test length based upon a Bayesian model. These procedures appear useful for determining the number of items required for a domain-referenced proficiency test.

The binomial and Bayesian procedures are appropriate for proficiency tests because they make no assumptions about the homogeneity of items. For a competency test, however, the number of items required to provide reliable mastery classifications of examinees is closely related to item homogeneity. Without a homogeneous item set, the bimodal characteristics of the score distribution would not be pronounced. Unless the distinctiveness of a measure can be increased to produce a more homogeneous set of items, a greater number of items will be necessary to minimize the amount of classification error.

Figure 1 indicates that ten items were more than enough for classifying examinees on the behavior of interest. For tests of unitary intellectual skills, in which there is little confounding with subordinate skills or related information three-five items are probably sufficient for providing reasonably reliable classifications. If this is true, considerably fewer items are necessary for measuring a homogeneous class of behaviors than is necessary for measuring a collection of behaviors such as verbal information.

Reporting Results

The final consideration involving the relationship between criterion-referenced test characteristics and assumptions about the nature of achievement variables concerns the reporting of test results. Sensible reporting of the results on a competency test should probably be binary (e.g., master-nonmaster or pass-fail). The purpose of the test is to determine in which category the student actually belongs. Deviations from these categories are assumed to be attributable only to error and need not be included in the score reporting.

Scores from a test of an achievement continuum are expected to reflect the underlying range of capabilities. These scores are more meaningfully expressed as a percentage passed or a proficiency level. Even if the information from a proficiency test is used to divide the group into mastery and nonmastery classifications through an established passing score, it appears unjustifiable not to inform the students of the obtained estimate of his true level of proficiency.

Summary

It was suggested that different measurement models may be required for assessing different types of learning outcomes. In particular, intellectual skills apparently encompass classes of behaviors that are demonstrated in an all or none manner, while a collection of verbal information can be achieved to varying degrees. If this is true, mastery seems more relevant to skill learning and proficiency is a more important concept for information. By considering alternative measurement models for these two situations, a new perspective is provided for viewing the contradictions and controversies related to criterion-referenced measurement theory. Table 1 summarizes some of the implications that alternative achievement variable models may have for different characteristics of criterion-referenced tests.

Table 1

Relation of Criterion-Referenced Test Characteristics to Assumptions about the Nature of Achievement Variables

Criterion-Referenced Test Characteristics	Achievement Variable Model	
	Binary	Continuous
Name	Competency Test	Proficiency Test
Application	Intellectual Skills	Verbal Information
Type of Performance	Class of Behaviors	Collection of Behaviors
Score Distribution	Bimodal	Variable (Depends on item domain and test administration sample.)
Test Homogeneity	Desirable (Characteristic of a good test.)	Unnecessary (Often indicates an overly-restricted item domain.)
Passing Score	Established by determining point of minimal overlap of distribution.	Established to maintain performance standard (judgment) in conjunction with test length and error probability (Binomial or Bayesian methods).
Test Length	Determined by homogeneity and importance of correct classification.	Determined by passing score and error probability (Binomial or Bayesian methods).
Reporting Results	Dichotomous (Pass-Fail or Mastery-Nonmastery)	Proficiency estimate

Many domain-referenced tests have been constructed, either intentionally or unintentionally, to measure collections of several intellectual skills and a variety of verbal information simultaneously. In such situations it is virtually impossible to draw inferences concerning what the examinee can and cannot do. If items are randomly sampled from a domain of clearly defined verbal information it is possible to infer the examinee's capability or proficiency relative to the entire domain. Likewise, measurement of a unitary intellectual skill permits conclusions concerning whether or not the skill has been mastered. Combining of skills with other skills or verbal information results in confounding of measurement that makes any conclusion tenuous.

References

Bergquist, C. C. & Horwitz, S. P. A preliminary study of test characteristics for criterion-referenced tests of intellectual skills. Unpublished paper, Florida State University, 1975.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 4, 321-326.

Gagne, R. M. Instructional variables and learning outcomes. SCEIP Occasional Report No. 16, September, 1968, University of California, Los Angeles, California.

Gagne, R. M. Domains of learning. Presidential address presented at the annual meeting of the American Educational Research Association, February, 1971, New York City, New York.

Gagne, R. M. Observing the effects of learning. Paper presented at the annual meeting of the American Psychological Association, New Orleans, September, 1974.

Gagne, R. M. and Briggs, L. J. Principles of instructional design. New York: Holt, Rinehart, and Winston, 1974.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Graham, D. L. An empirical investigation of the application of criterion-referenced measurement to survey achievement testing. Unpublished doctoral dissertation, Florida State University, 1974.

Hambleton, R. K., and Novick, M. R. Toward an integration of theory and method for criterion referenced tests. Journal of Educational Measurement, 1973, 3, 159-170.

Horwitz, S. P. Effects of amount of immediate and of delayed practice on retention of mathematical rules. Unpublished paper, Florida State University, 1974.

Kriewall, T. E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report No. 103, October, 1969, Wisconsin Research and Development Center, Madison, Wisconsin.

Millman, J. 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.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-215.

Novick, M. R. & Lewis, C. Prescribing test length for criterion-referenced measurement. I ACT Technical Bulletin, No. 18, January, 1974.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 1, 1-9.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement. (2nd Ed.) Washington: American Council on Education, 1971.