DOCUMENT RESUME

ED 083 298 TM 003 263

AUTHOR Woodson, M. I. Charles E.
TITLE Classical Test Theory and Criterion-Referenced Scales.
NOTE 13p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; *Item Analysis; *Statistics; Technical Reports; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS Classical Test Theory

ABSTRACT

The item (difficulty and discrimination) and test (reliability and validity) statistics in classical test theory are highly dependent upon the calibration sample of individuals used. The estimates of item and test parameters in classical test theory is valid within a range of interest along the characteristic measured. Generally, this range of interest is the distribution of the characteristic in some population, and the calibration sample used is intended to be a random sample from that population. In such populations, the extremes usually are poorly represented, and the parameter estimates are relatively poor at these extremes. For criterion-referenced scales, the range of interest is defined by a range of the characteristic rather than the distribution of that characteristic in some population. The calibration sample must be representative of that range of interest. When the range of interest is appropriately defined, an appropriate calibration sample may be selected, and classical test theory applies directly to criterion-referenced scales. (Author/DB)

CLASSICAL TEST THEORY AND CRITERION-REFERENCED SCALES

M. I. Chas. E. Woodson

University of California, Berkeley

ABSTRACT

Many authors have thought classical test theory was invalid for criterion-referenced tests. The item (difficulty and discrimination) and test (reliability and validity) statistics in classical test theory are highly dependent upon the calibration sample of individuals used. We may speak of the estimates of item and test parameters in classical test theory as valid within a range of interest along the characteristic measured. It has generally been the case that this range of interest is the distribution of the characteristic in some population and the calibration sample used is intended to be a random sample from that population. In such populations, it is usually the case that the extremes are poorly represented and the parameter estimates are relatively poor at these extremes.

For criterion-referenced scales the range of interest is defined by a range of the characteristic rather than the distribution of that characteristic in some population. The

calibration sample must be representative of that range of interest. When the range of interest is appropriately defined, an appropriate calibration sample may be selected, and classical test theory applies directly to criterion-referenced scales.

Classical Test Theory and Criterion-Referenced Scales

M. I. Chas. E. Woodson

University of California, Berkeley

A wide variety of definitions of "criterion-referenced test" have been suggested (e.g., Glaser, 1963; Glaser & Nitko, 1971; Harris & Stewart, 1971; Ivens, 1970; Kriewall, 1969; Popham & Husek, 1969; Hively, Patterson & Page, 1968). Common to these definition is an emphasis on the interpretation of test outcomes in terms of behavior. We shall take the position that "criterion-referenced" is not a property of the test but of a scale for interpreting the test (Woodson, 1973a), although it seems likely that the kind of scale one has in mind using with a test will have an impact on test construction procedures. Our definition of "criterion-referenced" is close to that of Glaser and Nitko (1971): "A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specific performance standards." We would modify this to refer to scales, and rather than limit ourselves to a cutoff score associated with a standard, place individuals on a scale interpretable in terms of behavior. Therefore, "a criterion-referenced scale is one that yields measurements directly interpretable in terms of some specific dimension of behavior." Note that it is not designed to most effectively rank individuals within a population.

In our judgment, there has been an over emphasis on determining whether an individual has exceeded a standard in order to stop instruction on that objective. Instructors need to know where the individual is on a dimension of learning. For example, Woodson (1973c) found the effectiveness of instructional steps differed considerably at different degrees of learning. If other studies find this, degree of learning will be a significant parameter in instructional models.

It has been argued that classical test theory does not apply to criterion-referenced tests (Popham & Husek, 1969) because under some common circumstances criterion-referenced test items and the tests themselves are likely to have no variance and a lack of variance makes the common statistics (item difficulty, item discrimination, test reliability and test validity) invalid or undefined. Woodson (1974) has argued that this argument is falacious as all items and tests must have variance within the range of interest for which they are calibrated in order to provide any useful information.

The above argument suggests that classical test theory may therefore be relevant to criterion-referenced test and item analysis. The present paper argues that this is the case.

In classical test theory, item and test parameters are estimated by statistics from a calibration sample of individuals. For classical test theory the calibration of a test or item must be done within a population of testees with appropriate variability on the characteristic measured. The distribution of the characteristic of interest in the population sampled, and therefore the distribution of the characteristic in the sample, determines in part the parameter estimates. These statistics are known to be sensitive to restriction of the sample.

Item difficulty within a population, estimated by difficulty within the sample, obviously depends upon the distribution of the characteristic in the sample. To skillful individuals, an item is much easier than to less skillful individuals.

Estimates of item discrimination within a population are also sensitive to the characteristics of the calibration sample used. If the calibration is restricted in some way, the estimates may be unreliable.

Test reliability within a population, the most commonly used parameter to evaluate a test, is known to be sensitive to the characteristics of the calibration sample.

These classical test theory statistics are referred to here as "within the population" to emphasize the characteristic that they are bound by the population which the sample represents. In most cases random sampling from the population is assumed, so the statistics apply for a specified population of testees (e.g., 4th, 5th and 6th graders).

Another way of conceptualizing this situation is to refer to these statistical estimates of the parameters involved as valid within a range of interest. For norm-referenced scales, the calibration sample is a random sample of a population, the distribution of the characteristic in this population defines the range of interest. Such a scale is norm-referenced in that the scale is dependent upon the population represented by the calibration-sample for its meaning.

Criterion-referenced scales are scales whose meaning refers to the characteristic measured rather than the distribution of the characteristic in some population. It is therefore necessary to estimate item and test parameters with statistics valid for the range of interest within which the test will be used. This can be done by specifying the characteristics of the population for which the test is to be calibrated.

In the case of criterion-referenced scales, the items or test statistics also apply to a "range of interest", that is, a range of the characteristic for which data is available and the item and test are calibrated. In the norm-referenced test, this is specified as the range of the characteristic in the population.

The same item and test statistics of classical test theory used for norm-referenced scales apply to criterion-referenced scales, provided the range of interest is appropriately specified. One way of specifying this range of interest (W. E. Coffman, personal communication) is to include in the calibration sample equal numbers of individuals who have received and have not received relevant instruction. A more general procedure is to choose a calibration sample which contains an adequate representation of the range of the characteristic to be measured.

Difficulty within the range of interest is therefore a relevant characteristic for item analysis. Discrimination within the range of interest is the most useful statistic for the selection of items. Test reliability and validity also have the same meaning for criterion-referenced scales as they do for norm-referenced scales.

Note, however, that no matter what the type of scale is being used, if the calibration sample is highly restricted, or not representative of the range of interest, the item and test statistics are not valid estimates of the parameters of interest. For example, if the ability of the calibration sample is so high that the items of a particular test are trivially easy, this restriction of the sample makes the statistics invalid for any range of interest other than the one upon which the test was calibrated, and trivial within that range because the item (or test) does not discriminate.

The empirical characteristics of items within a calibration sample are used to select items for a test and thereby contributes to the determination of what a test measures. If a norm-referenced approach is taken, items which do not measure a characteristic which varies within the calibration sample tend to be discarded, and items which vary greatly within the calibration sample tend to be selected. For criterion-referenced tests, the reference is not a population but a range of a dimension of behavior.

Examples

Consider the problem of the development of a spelling test and related norm-referenced and criterion-referenced scales. The characteristic involved is spelling ability within the 500 most frequently misspelled English words.

The norm-referenced approach to construct a 10 item test would be:

1. Select a sample (not necessarily randomly) of the items,
2. Administer to a calibration sample of individuals, randomly sampled from the population which defines the range of interest within which the item and test parameter estimates will be valid,
3. Compute item difficulties within the sample which are estimates of the difficulty in the range of interest (about .5 is desirable),
4. Compute item discriminations within the sample (the higher the better),
5. Select the 10 items with the best discrimination estimates,
6. Norms are prepared for the population for which the test is designed (7th, 8th, 9th graders),
7. Individual performance would be described in terms of how individuals compared to the distribution of scores in the standardization sample

(e.g., rank order within 7th graders, or grade-equivalent scores),

8. The resulting scale may be referred to as spelling ability relative to the distribution of abilities of a particular population of persons on those items which these persons differ most frequently.

The purpose of this scale is to discriminate among persons on spelling ability, therefore items selected will tend to be ones on which persons differ the most. In other words, differences among persons in the calibration sample will contribute to the definition of what is measured.

The criterion-referenced scale approach to construction of a 10 item test:

1. Select a sample (not necessarily randomly) of the items.
2. Administer the items to a calibration sample. The calibration sample is selected to be representative of the population of observations (range of interest) for which the items and test are to be calibrated. If an instructional program is being assessed, this would include appropriate proportions of persons to represent every value of the characteristic in question in the range of interest.
3. Compute item difficulties within the calibration sample, (.5 would give most effective measurement near the center of the range of interest, other values are needed for the extremes).
4. Compute item discriminations within the calibration sample, (the higher the better).
5. Select the 10 items with the best discrimination estimates within the calibration sample.
6. Scores of individuals are in terms of the selected items within the range of interest. An individual score on this scale does not rank

him with respect to others, but places him on a scale defined by the items.

7. The resulting scale may be referred to as spelling ability on the 500 most misspelled words.

This scale is not dependent upon the distribution of the characteristic in a population.

Note that for the criterion-referenced scale, items are eliminated for being inappropriate within the range of interest, which is not necessarily the distribution of the characteristic in some standardization population. It may well be a range at the extreme of some natural population of individuals. It may also include observations on an individual various levels of learning.

In the limiting case of a very short range of interest, item discrimination and test reliability go to zero.

In the limiting case of a very broad range of interest, observations may be difficult to obtain to reliably estimate parameters. This is quite reasonable, we cannot calibrate a test by classical test theory for a range of a characteristic of which we have few or no instances.

In the limiting case where a population of individuals is randomly sampled, we have, of course, the classical norm-referenced situation.

In short, the range of interest and therefore the calibration sample, in which a test is developed and calibrated defines the range of the characteristic for which the test is useful.

This paper has taken the approach of using a calibration sample representative of the range of interest of the characteristic and using classical test theory to develop and evaluate a test. This is necessary because the estimates of item and test parameters used in classical test theory are sensitive to the calibration sample. Modern test theory may well free us of the burden of sample-bound calibration. The two-parameter logistic model yields sample-free calibration in theory (Rasch, 1966) and in practice (Wright, 1968). There is also evidence (Woodson, 1973; Samejima, 1973) that the three-parameter normal-ogive model gives relatively sample-free calibration. Sample-free calibration may not require the specification of a range of interest.

Pending fortunate developments in test theory, the developer of criterion-referenced scales is best advised to select a calibration sample representative of the range of interest of the characteristic to be measured, and use the item statistics and test statistics of classical test theory, bearing in mind that the estimates of the parameters obtained are valid only for a particular range of the characteristic in question.

References

Ebel, R. L. Criterion-referenced measurements: Limitations. *School Review*, 1971, *69*, 282-288.

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

Glaser, R. & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement*. Washington: American Council on Education, 1971, pp. 625-670.

Harris, C. W. An index of efficiency for fixed length mastery tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972. (b)

Harris, M. L. & Stewart, D. M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hively, W., Patterson, H. L., & Page, S. A. "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275-290.

Ivens, S. H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Kriewall, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Samejima, F. Item analysis of a non-verbal reasoning test in terms of two subject groups of different nationalities. Proceedings: American Psychological Annual Meeting, 1973.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, in press.

Woodson, M. I. C. E. Criterion-referenced scales by a normal-ogive model, 1973b, in press.

Woodson, M. I. C. E. How scales get their meaning, 1973a, in press.

Woodson, M. I. C. E. Optomizing the Instruction of Meaningful Paired-Associates: Degree of Learning and Degree of Prompt. Journa´ of Experimental Psychology, 1973c, in press.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1968, 85-101.

AUTHOR

Woodson, M. I. Chas. E. Address: Institute of Human Learning, University of California, Berkeley, CA 94720. Title: Assistant Professor and Assistant Research Psychologist. Degrees: Ph.D., University of California. Specialization: Psychometrics and Instruction.