DOCUMENT RESUME

ED 087 821 TM 003 479

AUTHOR Helmstadter, Gerald C.
TITLE A Comparison of Bayesian and Traditional Indexes of Test Item Effectiveness.
PUB DATE 74
NOTE 4p.; Paper presented at the Convention of the National Council on Measurement in Education (Chicago, Illinois, 1974)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Achievement Tests; *Bayesian Statistics; *Comparative Analysis; Criterion Referenced Tests; *Item Analysis; Knowledge Level; Norm Referenced Tests; *Statistics; Test Construction

ABSTRACT

Bayes Theorem leads to three indexes of item effectiveness: 1) probability that an examinee knows the content given that the correct response was selected; 2) probability that an examinee does not know that content given that an incorrect response was selected; and 3) probability of making a correct decision about the examinee's knowledge given the performance on that item. These indexes and classical item discrimination were compared for three definitions of "high" and "low" knowledge groups. The results indicated that the way of defining groups does make a difference and that there are two quite distinct characteristics of item effectiveness. (Author)

Presented at the 1974 Convention of the National Council on Measurement in Education, Chicago, Illinois.

A Comparison of Bayesian and Traditional Indexes of Test Item Effectiveness

Gerald C. Helmstadter, Arizona State University

Bayes Theorem leads to three indexes of item effectiveness: 1) probability that an examinee knows the content given that the correct response was selected; 2) probability that an examinee does not know the content given that an incorrect response was selected; and 3) probability of making a correct decision about the examinee's knowledge given the performance on that item. These indexes and classical item discrimination were compared for three definitions of "high" and "low" knowledge groups. The results indicated that the way of defining groups does make a difference and that there are two quite distinct characteristics of item effectiveness.

Presented at the 1974 Convention of the National Council
on Measurement in Education, Chicago, Illinois

A Comparison of Bayesian and Traditional Indexes of Test Item Effectiveness

Objectives

Recent emphasis on the use of criterion-referenced concepts as contrasted with normative-referenced concepts in designing tests to assess classroom performance of students has led some authors to question the value of some of the traditional methods for evaluating test items and to suggest that new and more efficient procedures should be developed. The purpose of this study was to apply Bayes Theorem to test item analysis and to compare the resulting indexes of item effectiveness with the traditional index of test item discrimination when "high" and "low" knowledge groups are defined in three different ways.

Method

Bayes Theorem was applied in the item analysis context and led to the development of three separate indexes of item effectiveness as follows: 1) the probability that a subject knows the content material given that the correct response was selected; 2) the probability that a subject does not know the content material given that the incorrect response was selected; and 3) the probability that a correct decision will be made about the examinee's knowledge of the content given the results of performance on that item.

These three item characteristics, together with a classical item discrimination index, were then computed for each item contained in two different final examinations which had been given both as a pre and a post test in two different university courses. Three separate variants of these four item characteristics were obtained by defining "high knowledge" and "low knowledge" groups in different ways. The first way assumed that persons in the top one-third of the class on the post test were "high knowledge" and persons in the bottom one-third of the class on the post test were "low knowledge"; the second way involved combining scores from the pre and post test as if they constituted one large class and then assuming that persons in the top one-third of this doubled class were "high knowledge" and that persons in the bottom one-third of this doubled class were "low knowledge"; and the third way assumed that pre test scores represented a "low knowledge" group and that post test scores represented a "high knowledge" group.

Intercorrelations among the twelve different indexes derived for each item were then computed over the items within each of the two separate tests. Then, to determine the extent to which varying the definition of high and low knowledge groups would influence assessment of the item, the medians of the intercorrelations among ways of defining groups as calculated separately for each type of item index were obtained. Finally, to determine the extent to which varying the type of information about the item would influence assessment of the item, the medians of the intercorrelations among the different item indexes as calculated separately for each way of defining groups were obtained.

Data Sources

The responses to items used as empirical data in this study were obtained from 43 students enrolled in a university course in multivariate statistics and 55 students enrolled in a university course in adolescent psychology. The examination given as both a pre and a post instruction test in the statistics course contained 50 multiple choice questions of five alternatives each, while the examination given as both a pre and a post test in the adolescent psychology course contained 59 such items.

Results

The resulting intercorrelations indicated that assessment of items based on a post test only definition of "high and low knowledge" groups would lead to only moderately similar conclusions (median intercorrelations of .58 and .44) to those obtained when definitions were based on both pre and post test results and that as long as both pre and post tests were used in the definition of "high" and "low" knowledge groups, the assessment of the items would be quite similar (median intercorrelations of .90). The results also indicated that the classical discrimination index comes closest to providing the same item assessment as would the Bayesian probability of making a correct decision (median of the intercorrelations = .82) but that those items which are effective indicators that the examinee does know the material are not necessarily the same items as those which are effective indicators that the examinee does not know the material.

(The median of the intercorrelations between the probability that the examinee knows the materials given that the correct response was selected and the probability that the examinee does not know the material given that the incorrect response was selected was .00).

Importance

These data clearly suggest that the common practice of defining "high" and "low" knowledge groups in terms of scores on the post test only is questionable. They further indicate that there may be two quite distinct types of effective achievement test items: those that indicate that the examinee knows the material and those that indicate that the examinee does not know the material. Thus, another common practice - that of using a single index of item discrimination - is also called into question.