DOCUMENT RESUME

ED 397 077 TM 025 074

AUTHOR Dimitrov, Dimiter M.
TITLE On the Cutting Score Determination in Dichotomous Classifications.
PUB DATE Oct 95
NOTE 8p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 11-15, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Chi Square; *Classification; *Criterion Referenced Tests; *Cutting Scores; *Research Methodology; Test Results; Validity
IDENTIFIERS *Dichotomous Scoring

ABSTRACT

The choice of a cutting score for criterion-related tests influences decisions related to classifying people into dichotomous categories. This paper proposes an empirical methodology for determining the best cutting score when there is information about the test score frequency distribution of test-takers defined as actually successful and actually unsuccessful on some criterion. The method is based on two statistics calculated for each possible cutting score. The first is a pure hit rate, representing the proportion of correct classifications above those expected by chance. Second is a chi-square statistic for testing the significance of the difference between the population frequencies of the two types of misclassifications errors. A cutting score summary table is developed based on the information about the test score frequency distributions of two validation samples based on actually successful and actually unsuccessful samples. Cutting scores are divided into those that yield equal frequencies of the two types of misclassification errors and those in which the frequency of one type of error is higher than that of the other. The cutting score summary table facilitates the determination of the best cutting score in each category. (Contains 2 tables and 19 references.) (SLD)

On the Cutting Score Determination in Dichotomous Classifications

Dimiter M.Dimitrov

Kent State University - Kent, Ohio

A paper presented at the Annual Meeting of
The Mid-Western Educational Research Association
Chicago, IL October 11-15, 1995

ON THE CUTTING SCORE DETERMINATION IN DICHOTOMOUS CLASSIFICATIONS

INTRODUCTION

The choice of a **cutting score** for criterion-related tests influences decisions related to classifying people into dichotomous categories - for example, decisions based on tests for admitting students to a college, hiring job applicants, prescription of preventive psychopathological therapy, etc. In general, choosing an appropriate cutting score is an essential issue in setting standards on educational, psychological, and occupational tests which explains the considerable amount of publications related to this topic. All major works on determining optimal cutting scores focus on the estimation of two possible types of misclassifications errors by using different models for the true score distribution - mainly Bayesian models and binomial models (e.g. Hambleton & Novick, 1973; Klein & Cleary, 1967; Huynh, 1976; Lord & Stocking, 1976; Wilcox, 1977). The sum of the estimated misclassification error probabilities, multiplied by judgmentally specified misclassification "losses", define the expected loss and, most frequently, the test score that minimizes the expected loss is taken as the best cutting score. However, factors like need for testing model assumptions, judgmental nature of the misclassification losses, and relatively difficult calculation of the expected losses still keep the door open for a search of technically simple procedures which do not include assumptions about the true score distribution.

In an attempt to make a step in this direction, the present paper proposes an empirical methodology for determining the best cutting score when there is an information about the test score frequency distribution of test-takers defined as actually successful and actually unsuccessful on some criterion (educational, clinical, professional, etc.).

METHOD

The approach proposed here is methodologically based on the following two statistics calculated for each possible cutting score:

1) A "**pure hit rate**", **PHR**, representing the proportion of correct classifications above the expected by chance.

2) A χ^2 - statistic for testing the significance of the difference between the population frequencies of the two types of misclassifications errors.

Table 1 represents the general form for the two-way classification frequency distribution yielded by a given cutting score. Cell **A** is the frequency of correct classifications of the type **"predicted successful - actually successful" (PS-AS)**, and cell **D** is the frequency of correct classifications of the type **"predicted unsuccessful - actually unsuccessful" (PU-AU)**. Cell **B** is the frequency of misclassiffications of the type **"predicted unsuccessful - actually successful" (PU-AS)**, and cell **C** is the frequency of the other type misclassifications: **"predicted successful-actually unsuccessful" (PS-AU).** The proportion of the correct classification is called **hit rate**:

(1) $\quad HR = \frac{A+D}{N}, \quad$ where $N = A + B + C + D.$

Table 1 **Predicted classification**

		Successful	Unsuccessful	Total
Actual classification	Successful	**A**	**B**	A + B
	Unsuccessful	**C**	**D**	C + D
	Total	A + C	B + D	N

The **hit rate**, as calculated by (1), is taken into account in many empirical approaches for determining optimal cutting scores (e.g. Berk, 1976 ; Allen & Yen, 1979, p. 104). However, its value includes a proportion of correct classifications that may occur by chance. In order to avoid this problem and increase the reliability of the cutting score determination, we propose the use of the **Cohen's kappa,** κ, (see Cohen, 1960). In the context of the present study, κ will represent the proportion of correct classifications, **PS-AS** and **PU-AU**, above that expected by chance, i.e. the **"pure hit rate", PHR,** and is calculated by the respective formula:

(2) $$PHR = \frac{HR - P_c}{1 - P_c}$$

where: HR is the hit rate calculated by (1) ;

P_c is the proportion of correct classification expected to occur by chance and, in terms of the cell frequencies in Table 1, is: $P_c = \frac{(A+B)(A+C)/N + (C+D)(B+D)/N}{N}$.

The question about the equality of the misclassifications in both "directions" **PS-AU** and **PU-AS**, is related to testing the following null hypothesis: **For a given sample of test-takers, the entries B and C in Table 2 differ only as a result of chance sampling.** If this is true, the expected number of **PS-AU** misclassifications equals the expected number of **PU-AS** misclassifications and is given by the average of **C** and **B**, i.e. **(B + C)/2.** Hence, the null hypothesis can be tested by the use of a χ^2-statistic, which is the sum of the squared differences between the observed and expected frequencies, each divided by the expected frequency:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{\left(C - \frac{B+C}{2}\right)^2}{\frac{B+C}{2}} + \frac{\left(B - \frac{B+C}{2}\right)^2}{\frac{B+C}{2}} = \frac{(B-C)^2}{B+C} .$$

Hence, the calculation of the χ^2- statistic for testing the significance of the difference between the population frequencies of the two types of misclassification errors is given by the formula:

(3) $\chi^2 = \frac{(B-C)^2}{B+C}$.

This is, in fact, an application of the **McNemar** test for significance of changes for the situation represented by Table 1. In this case, 2 x 2 tables, the degrees of freedom are df = 1 which makes the use of χ^2 suspicious when the expected frequencies $\frac{B+C}{2}$ are less than 5. The **Yates'** correction for continuity leads to the following corrected form:

(4) $\chi^2 = \frac{(|B-C|-1)^2}{B+C}$; (see **McNemar**, 1969, pp. 260-263).

For example, if a given cutting score leads to the following cell frequencies in Table 1: **A**=45, **B**=15, **C**=10, and **D**=30, by using formula (3), we calculate: $\chi^2 = \frac{(B-C)^2}{B+C} = \frac{(15-10)^2}{15+10} = 1.00$. This number is less than the critical value, $\chi^2 = 3.841$, at level of significance $\alpha = .05$ and degrees of freedom df = 1. Hence, in this case, the cutting score yields equal population frequencies of the two types of misclassifications errors, **PS-AU** and **PU-AS**. On the other hand, the **pure hit rate** yielded by this hypothetical cutting score will be **PHR** = .49, after applying formulas (1) and (2) for the calculation of the hit rate, **HR**, and the pure hit rate, **PHR**, respectively.

Thus, the χ^2 statistic and the **PHR** index answer two very important questions related to each possible cutting score:
1) Does the cutting score yield equally serious misclassification errors, **PS-AU** and **PU-AS** ?
2) What is the proportion of correct classifications above that expected by chance?

Cutting score Summary Table (CTS)

Proposed here is a table that summarizes the χ^2-values, the **PHR** values, and the cell frequencies **A**, **B**, **C**, and **D** from Table 1 yielded by each possible cutting score. This table, called "**Cutting score Summary Table**" (**CST**), is based on the information about the test score frequency distributions of two validation samples of people defined as **actually successful** and **actually unsuccessful** . Table 2 represents a **CTS** for hypothetical data including a test scale, given in column **C1**, and the frequencies over this scale of actually successful (**AS**) and actually unsuccessful (**AU**) test-takers, given in columns **C2** and **C3**, respectively. The calculation of the numbers in columns **C4, C5, ..., C10** is straightforward:

C4 = the number of correct **PS-AS** classifications (cell **A** in Table 1), obtained as cumulative frequencies from column C2;
C5 = the number of **PS-AU** misclassifications (cell **C** in Table 1), obtained as cumulative frequencies from column **C3**;
C6 = the number of **PU-AS** misclassifications (cell **B** in Table 1), obtained as N_s- À, i.e. by subtracting the column **C4** numbers from the total number of successful people, N_s ;

C7 = the number of correct **PU-AU** classifications (cell **D** in Table 1), obtained as N_U - **C**, i.e. by subtracting the column **C5** numbers from the total number of unsuccessful people, N_U ;
C8 = the **hit rate**, **HR**, calculated by formula (1);
C9 = the **pure hit rate**, **PHR**, calculated by formula (2);
C10 = the χ^2 statistic, calculated by formula (3).

Table 2

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Test score	**Actually Success.**	**Actually Unsucc.**	**PS-AS (A)**	**P S-AU (C)**	**PU-AS (B)**	**PU-AU (D)**	**HR**	**PHR**	**Chi-sq..**
9	9	0	9	0	75	46	.423	.078	75.00
8	18	2	27	2	57	44	.546	.218	51.27
7	26	7	53	9	31	37	.692	.393	12.10
6	11	10	64	19	20	27	.700	.347	0.02 * *
5	10	8	74	27	10	19	.715	.321	7.81
4	4	4	78	31	6	15	**.715**	.290	16.89
3	2	7	80	38	4	8	.677	.152	27.52
2	4	6	84	44	0	2	.661	.055	44.00
0 or 1	0	2	84	46	0	0	.646	.000	46.00

The ** in column **C10** indicates a χ^2- statistic which is less than the critical $\chi^2 = 3.841$, at the $\alpha = .05$ level of significance. The respective cutting score yields equally serious misclassification errors, **PS-AU** and **PU-AS,** in the sense that it yields equal frequencies of the two types of errors over the entire population of test-takers.

As one can see from Table 2, the test score of **6**, if taken as a cutting score, is the only one that yields equally serious misclassification errors, **PS-AU** and **PU-AS,** because its χ^2-statistic (=.02) is the only one which is less than the critical $\chi^2 = 3.841$ (with $\alpha = .05$ and df = 1). Hence, under the assumption of equally serious misclassification errors, we can choose the cutting score of **6**. One can also see that all cutting score above the cutting score of **6** yield higher frequency of the **PU-AS** error compared to the frequency of the **PS-AU** error. Hence, if we prefer more **PU-AU** errors over the entire population of test-takers, we can choose the cutting score of 7 as the best cutting score because it yields the highest pure hit rate (**PHR**=.393) among all cutting scores above the cutting score of 6. Finally, if we prefer more **PS-AU** errors over the entire population of test-takers, we can choose the cutting score of **5** as the best one because it yields the highest pure hit rate (**PHR**=.321) among all cutting scores below the cutting score of **6.**

CONCLUDING REMARKS

The method proposed here for determining the best cutting score is based on the idea of the **pure hit rate** (**PHR** = proportion of correct classifications above that expected by chance) and on the fact that the **McNemar χ^2 test** in the context of dichotomous classification tables (see Table 1) divides the set of all possible cutting scores into three categories:

A) Cutting scores that yield equal frequencies of the two types of misclassification errors, **PS-AU** and **PU-AS**, over the entire population of test-takers. These cutting scores yield χ^2 -statistics which are less then the critical χ^2 (e.g. $\chi^2 = 3.841$ at the level $\alpha = .05$). The best cutting score is the one that yields the highest pure hit rate among all cutting score in this category, assuming that the two types of misclassification errors are equally serious.

B) Cutting scores for which the frequency of the **PU-AS** error is higher than this of the **PS-AU** error over the entire population of test-takers. These cutting scores yield χ^2-statistics which are greater than the critical χ^2 (e.g. $\chi^2 = 3.841$ at the level $\alpha = .05$) and they are greater then the cutting scores from the above category, A). The best cutting scores yields the highest pure hit rate among all cutting scores in this category, B), assuming that the **PU-AS** errors are less serious than the **PS-AU** errors.

C) Cutting scores for which the frequency of the **PS-AU** error is higher than this of the **PU-AS** error over the entire population of test-takers. Like the cutting scores in category B), the cutting scores in this category also yield χ^2- statistics greater than the critical χ^2 (e.g. $\chi^2 = 3.841$ at the level $\alpha = .05$), but they are less than the cutting scores in category A). The best cutting score is the one that yields the highest pure hit rate among all cutting scores in this category, C), if it is assumed that the **PS-AU** errors are less serious than the **PU-AS** errors.

The Cutting score Summary Table (**CST**), illustrated by Table 2, facilitates the determination of the best cutting score in dependence of the category, A), B), or C), reflecting the assumption about the seriousness of the misclassification errors. The development of the **CTS** is straightforward for a simple use of a calculator or some statistical software. For example, the description of columns **C4, C5, ..., C10**, given in relation to Table 2, is directly interpretable in **MINITAB** commands. This is an important advantage of the method for either real data manipulations or computer simulations in the process of determining the best cutting score for the purposes of dichotomous classifications.

REFERENCES

ALLEN, M. J., & YEN, W. M. Introduction to measurement theory. Monterey/California: Brooks/Cole Publishing Company, 1979.

BERK, R.A. (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore: The Johns Hopkins University Press, 1980.

BERK, R.A. Determination of optional cutting scores in criterion-referenced measurement. The Journal of Experimental Education, 1976, **45**, No.2, 4-9.

COHEN, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, **XX**, No.1, 37-46.

DARLINGTON, R.B., & STAUFFER, G.F. A method for choosing a cutting point on a test. Journal of Applied Psychology, 1966, **50**, 125-129.

EMRICK, J.A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, **8**, 321-326.

HAMBLETON, R.K. Testing and decision-making procedures for selecting individualized instructional programs. Review of Educational Research, 1974, **44**, 371-400.

HAMBLETON, R.K., & NOVICK, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, **10**, 159-170.

HUYNH, N. Statistical consideration of mastery scores. Psychometrika, 1976, **41**, 65-78.

KLEIN, D.F., & CLEARY, T.A. Platonic true scores and error in psychiatric rating scales. Psychological Bulletin, 1967, **68**, 77-80.

LIVINGSTON, S. A., & ZIEKY, M. J. Passing scores. A Manual for Setting Standards of Performance on educational and occupational tests. Princeton, N.J.: Educational Testing Service, 1982.

LORD, F.M., & STOCKING, M.L. An interval estimate for making statistical inference about true scores. Psychometrika, 1976, **41**, 79-88.

McNEMAR, Q. Psychological statistics. New York: Wiley, 1969.

MILLMAN, J. Reliability and validity of criterion-referenced test scores. In R.E. Traub (Ed.), New directions for testing and measurement, 1979, No.**4**, San Francisco: Jossey-Bass.

SECHREST, L. Incremental validity: A recommendation. Educational and Psychological Measurement, 1963, 23, 153-158.

SUTCLIFFE, J. P. A probability model errors of classification- I: General considerations. Psychometrika, 1965, **30**, 73-96.

SWAMINATHAN, H., HAMBLETON, R. K., & ALGINA, J. J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, **12**, 87-98.

VAN DER LINDEN, W. J., & MELLENBERGH, G. J. Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1977, **1**, 593-599.

WILCOX, R. R. Estimating the likelihood of a false-positive or false-negative decisions with a mastery test: An empirical Bayes approach. Journal of Educational Statistics, 1977, **2**, 289-307.