DOCUMENT RESUME

ED 356 263 TM 019 690

AUTHOR Hester, Yvette
TITLE A Review of Strategies for Standard Setting and Identifying Cutoff Scores.
PUB DATE Jan 93
NOTE 17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 28-30, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; Classification; *Cutting Scores; *Evaluation Methods; Evaluators; *Mathematical Models; *Scoring Formulas
IDENTIFIERS Decision Theory; Empirical Research; *Experts; Minimax Procedure; *Standard Setting

ABSTRACT

Some of the different approaches to standard setting are discussed. Brief comments and references are offered concerning strategies that rely primarily on the use of expert judgment. Controversy surrounds methods that use expert judges, as well as those using test groups to set scores empirically. A minimax procedure developed by H. Huynh, an empirical procedure that invokes evaluation of the mathematical properties of various cutoffs through the application of decision theory, is illustrated. Minimax procedures are useful in minimizing probabilities of misclassification (i.e., the optimal minimization of false negatives and positives). (Author/SLD)

A Review of Strategies for Standard Setting and Identifying Cutoff Scores

Yvette Hester

Texas A & M University, 77843-3368

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 29, 1993

Abstract

The purpose of this paper is to discuss some of the different approaches to standard setting. Brief comments and references are offered concerning strategies that rely primarily on the use of expert judgment. A minimax procedure, an empirical procedure that invokes evaluation of the mathematical properties of various cutoffs through the application of decision theory, is illustrated. Minimax procedures are useful in minimizing probabilities of misclassification; i.e. the optimal minimization of false negatives and positives.

The setting of standard cutoff scores, commonly called standard setting, is preliminary to interpreting many test performances. Examples of situations which require these types of scores are those in which the examinee is required to meet or exceed a previously set score before advancing to the next level or unit of learning, or certification for practice in a particular profession or occupation. After an appropriate cutoff score has been determined, say x_0 , an examinee must meet or exceed this score on the test to pass. In this case, the cutoff score of x_0 was applied directly to the observed scores, which is the most common way to use a cutoff score. A less common way to use a cutoff score is to apply it to the true (or domain) scores. The domain cutoff score divides the true score scale into two regions which are referred to as mastery states. Examinees who have a true score at or above the cutoff score are called true masters and the examinees who have a true score below the cutoff score are called true nonmasters. Note that some standard-setting situations may call for the use of more than one cutoff score, but in practice it is most common to use a single score (Crocker & Algina, 1986). The purpose of this paper is to discuss some of the different approaches to standard setting. Most strategies that rely primarily on the use of expert judgments are based on one of three categories: 1) holistic impressions of the item pool, 2) content of each test item, and 3) examinee's performance on the test (Crocker & Algina, 1986).

Methods where judgments are based on holistic impressions of the item pool are widely used, but frequently criticized. Alternate panels of judges might set the standard cutoff score at different levels. A test

developer could perform replication studies, but the number of judges available for each study decreases by 1/k, where k is the number of replications performed. This, in turn, causes more fluctuation in the cutoff scores from sample to sample.

Methods where judgments are based on the content of each test item have been the most studied approaches to standard setting. Crocker and Algina (1986) consider three well-known procedures in this category.

Nedelsky (1954) designed a procedure for multiple choice items and was primarily concerned with setting standards of minimum competency for university-level examinations.

Angoff (1971) developed a method based on an individual judges' concept of the proportion of individuals from a minimally competent group who could answer a given item correctly. Summing across items would give a minimum passing score per judge. A general average or consensus of ratings across judges is the cutoff score.

Ebel (1972) uses a categorizing technique employing a two-dimensional grid. With one dimension as relevance and the other as difficulty, this system takes into account the possible influence of these two dimensions on the perception of the judges.

Comparison studies of these three methods have not shown one method to be superior over the others, although a well-cited study by Andrews and Hecht (1976) found large differences in standards set by the same panel of judges using the Ebel and Nedelsky methods. A more recent study by Behuniak, Archambault and Gable (1982) supported findings of large differences in standard scores set by different panels of judges using the same method. Saunders, Ryan and Huynh (1981) found that a judge's

content knowledge was a factor in producing a cutoff score using the Nedelsky method.

A final problem to be considered is that of intrajudge inconsistency (van der Linden, 1982). A judge may assign a lower probability to passing an easier test item and a higher probability to passing a harder test item. Using item response theory, van der Linden offers an index of discrepancy and demonstates its application.

Methods where judgments are based on examinees actual performance during some trial administration of a test have many critics. Logic for support of these methods comes from Shepard (1979), who explains that judges' standards are swayed by their concepts of howknown individuals would perform on a test or given test item and their judgments are confined to finite perceptions of these abilities. Empirical studies have shown that different cutoffs might be set by judges that have varying characteristics. One approach in this area is to use a test group of examinees lower in ability than the target population and set a standard using the "average" of the group.

Jaeger (1982) combines features of all three categories in a method called *iterative structured item judgment process.* A study illustrating its application was conducted for the North Carolina High School Competency Tests. Differences in standards set were reported depending upon group membership, e.g., the group of teachers recommended different cutoff scores than did the group of citizens.

To try and obtain an estimate of the contribution of differences in judges to a final cutoff score, Brennan and Lockwood (1980) examined the possibility of using generalizibility theory. Continued exploration in this area using this procedure is suggested (Crocker & Algina, 1986).

The psychometric problem of standard setting naturally invokes the question of legitimacy and justification. There are proponents for and against the practice. Crocker and Algina suggest the following steps to help answer these questions:

1. Question whether there exists a legitimate need to set standards.
2. Identify the likely threats to invalidity of the inferences to be made.
3. Use two or more different procedures.
4. Examine empirical evidence of how a typical sample of examinees perform on the test.

Recall that a domain cutoff score divides the true score scale into two regions; true masters and true nonmasters. Adopting the notation of Crocker and Algina, (1986), the domain cutoff score is denoted τ_0, and may be determined by one of the previously described methods. The observed score cutoff, denoted x_0, is to be determined.

Let τ be the true score for a particular examinee and let x be the observed score for the same examinee. The examinee will be classified in one of the following four ways:

Classification Grid

	$\tau \geq \tau_0$	$\tau < \tau_0$
	True	False
$x \geq x_0$	positive	positive
	False	True
$x < x_0$	negative	negative

Examinees on the off-diagonal are misclassified. Minimizing the probability of an examinee falling into one of these two categories, false positive or false negative, is the goal of these empirical procedures.

Hunyh (1976) developed a procedure for determining a cutoff score based on the assumption that the bivariate distribution of the domain and observed scores is beta-binomial Hunyh's procedure is quite accurate, but complicated to apply. A less complicated procedure was shown by Hunyh to be a good estimate if the number of test items is greater than or equal to 20 and the domain scale cutoff score is in the interval from .5 to .8 (Hunyh & Saunders, 1980). A formula for estimating the observed cutoff score is given.

A practical problem with these procedures can occur when used in ongoing testing programs which put psychometric and legal principles in possible conflict (Crocker & Algina, 1986). Estimates of observed cutoff scores may vary from year to year indicating samples from different populations. If a lower cutoff score is used in one year for psychometric reasons, legal problems could arise since previous examinees at that level would have failed the test.

One approach to examining empirical evidence of how a typical sample of examinees perform on a test that avoids this problem is to use a minimax procedure. A minimax procedure invokes the evaluation of the mathematical properties of various cutoffs through the application of decision theory. These procedures are useful in minimizing the probabilities of misclassification, i.e., false negatives and false positives, in a systematic way. A minimax procedure developed by Hunyh (1980) will be illustrated in this paper.

As in Hunyh's more complicated procedure, assumptions about the distribution of scores must be made before probabilities can be calculated. Most specifically, the distribution of the scores of the misclassified examinees is of concern. The binomial distribution is the most common assumption for these scores since scores are based on the proportion of items answered correctly. Proportions can be interpreted as probabilities for randomly chosen items. By specifically applying the theory of Bernoulli trials, the probability of answering x out of n randomly chosen items correctly can be computed.

Bernoulli Trials and the Binomial Distribution

A Bernoulli trial is a type of experiment with two possible outcomes, each the complement of the other. One is designated as a success and the other as a failure. For example, suppose the experiment of interest is the roll of a fair die. Suppose the outcome of interest is rolling a 6. A success (S) for this experiment is a 6 turning up on one roll of the die and a failure (F) is anything else turning up on the one roll of the die. Then the probability of a success, denoted p, is $\frac{1}{6}$ and the probability of a failure, denoted q (= 1 - p), is $\frac{5}{6}$. This experiment is classified as a Bernoulli trial.

Barnett and Ziegler (1990) afford us the following definition:

A sequence of experiments is called a sequence of Bernoulli trials, or a binomial experiment, if:

1. *Only two outcomes are possible on each trial.*
2. *The probability of success p for each trial is a constant (probability of failure is then q = 1 - p).*
3. *All trials are independent.*

The probability of x successes in n Bernoulli trials is computed by

$$(1) \quad C_{n,x}\, p^x\, q^{n-x}$$

where $C_{n,x} = \dfrac{n!}{(n-x)!\ x!}$ yields all possible combinations of n things taken x at a time without regard to order.

Clearly, then, the matter of correctly answering an item on a particular test is also a Bernoulli trial where S is a correct response and F is an incorrect response. A test would constitute a sequence of these trials.

A brief review of the binomial formula will lend itself to understanding the relationship between the binomial distribution and a sequence of Bernoulli trials. In general, it can be shown that a binomial expansion is given by

$$(a + b)^n = C_{n,0}\, a^n + C_{n,1}\, a^{n-1}\, b + C_{n,2}\, a^{n-2}\, b^2 + \ldots + C_{n,n}\, b^n$$

where n is a natural number.

Consider now an experiment consisting of a sequence of threeBernoulli trials. The number of successes for this sequence of experiments are the values of the random variable X. The probabilities associated with each of these values have been computed using formula (1) and shown in the following table.

X (= possible # of successes in 3 trials)	P(X)
0	$C_{3,0}\, p^0\, q^3 = q^3$
1	$C_{3,1}\, p^1\, q^2 = 3q^2p$
2	$C_{3,2}\, p^2\, q^1 = 3qp^2$
3	$C_{3,3}\, p^3\, q^0 = p^3$

Expanding $(q + p)^3$ using the binomial formula, we obtain

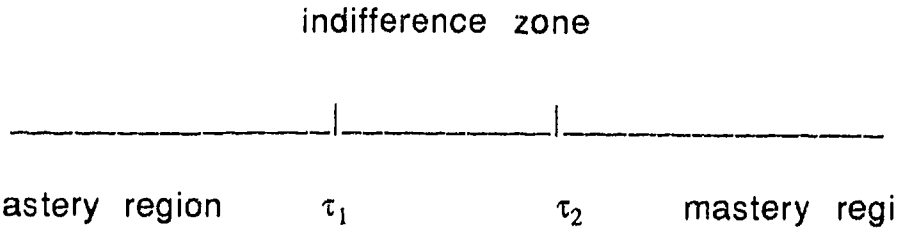
$$(q + p)^3 = C_{3,0}\, q^3 + C_{3,1}\, q^2\, p + C_{3,2}\, qp^2 + C_{3,3}\, p^3$$
$$= q^3 + 3q^2p + 3qp^2 + p^3.$$

Note that the probabilities in the second column of the above table are the terms in the binomial expansion of $(q + p)^3$.

Reasoning in the same way for the general case, the probability of each value of the random variable X is a term in the binomial expansion of $(q + p)^n$. Thus, the probability of x successes in n trials calculated by $C_{n,x}\, p^x\, q^{n-x}$ where x is an element of {0,1,2,...,n}, yields the Binomial Distribution.

Example

In using Hunyh's minimax procedure, τ_0 is not the domain cutoff score of interest. Instead, an *indifference zone* is created by setting domain cutoff scores τ_1 and τ_2 where τ_1 is the upper bound for the nonmastery region and τ_2 is the lower bound for the mastery region.



Examinees with scores in the indifference zone are close to both regions. Thus, the probability of misclassifying an examinee in this range is of no concern. The probabilities of concern and those we wish to minimize are the probabilities that an examinee at the τ_1 level will be misclassified as a master (false positive) and the prcbability that an examinee at the τ_2 level will be misclassified as a nonmaster (false negative).

Suppose that there are 5 items on a hypothetical test and that $\tau_1 = .6$ and $\tau_2 = .8$. The probability of correctly answering exactly x items out of the 5 item test must be caluculated using the formula for Bernoulli trials. The result will be a binomial proabability distribution for the two domain scores τ_1 and τ_2. A taole summarizing the probability of misclassification at each possible cutoff score based on the binomial distribution of the domain scores will allow the minimax cutoff score to be determined. Note that the values for the random variable X in the two probability distributions range from 0 to 5, inclusive, which is also the number of possible correct items on the test. The Bernoulli trial formula uses $p = .6$ and $p = .8$, respectively, which are the domain cut scores τ_1 and τ_2.

X (= possible # of correct items out of 5)	Domain Score (probability of exactly x items correct) $\tau_1 = .6$	$\tau_2 = .8$
0	$C_{5,0}\,(.6)^0\,(.4)^5 = .01024$	$C_{5,0}\,(.8)^0\,(.2)^5 = .00032$
1	$C_{5,1}\,(.6)^1\,(.4)^4 = .0768$	$C_{5,1}\,(.8)^1\,(.2)^4 = .0064$
2	$C_{5,2}\,(.6)^2\,(.4)^3 = .2304$	$C_{5,2}\,(.8)^2\,(.2)^3 = .0512$
3	$C_{5,3}\,(.6)^3\,(.4)^2 = .3456$	$C_{5,3}\,(.8)^3\,(.2)^2 = .2048$
4	$C_{5,4}\,(.6)^4\,(.4)^1 = .2592$	$C_{5,4}\,(.8)^4\,(.2)^1 = .4096$
5	$C_{5,5}\,(.6)^5\,(.4)^0 = .07776$	$C_{5,5}\,(.8)^5\,(.4)^0 = .32768$

Note that all probabilities satisfy the conditions for a probability distribution, i.e., each probability is in the range 0 to 1, inclusive and the probabilities for each distribution sum to 1.

To illustrate the use of the above table, suppose 3 is chosen as the cutoff score. Then an examinee with a domain score of .6 is misclassified as a master (false positive) if that examinee answers 3, 4 or 5 items correctly. The probabilities for those scores are taken from the probability distribution above and added to reach the total probability of a nonmaster being misclassified at that cutoff score. An examinee with a domain score .8 will be misclassified as a nonmaster (false negative) if that examinee answers 2 or less items correctly. Thus, the probabilities for 2, 1 and 0 are taken from the probability distribution above and added to reach the probability of misclassification. These values have been calculated for each possible cutoff score and summarized in the table below. An asterisk has been placed by the maximum probability for each cutoff score. The minimum of these maximum probabilities will be the

probability of the optimal cutoff score. In this example, the observed minimax cutoff score should be set at 4 items with maximum probability of misclassification at .33696. It follows, then, that all other examinees have a misclassification probability smaller than .33696.

Possible Cutoff Score (i)	Domain Scores	
	Nonmasters	Masters
	$\tau_1 = .6$	$\tau_2 = .8$
0	1 *	0
1	.98976 *	.00032
2	.91296 *	.00672
3	.68256 *	.05792
4	.33696 *	.26272
5	.07776	.67232 *

If the probability associated with the minimax cutoff score is unacceptable, increasing the test length will reduce this probability. Set first a maximum acceptable probability of misclassification and construct new tables for a first choice test length. Increase the length of the test item by item, calculating the new probabilities until the acceptable probability is reached. A more detailed treatment of increasing test length to reach a desired maximum probability can be found in Fahner (1974) and Wilcox (1976).

Summary

Many different approaches to setting standard cutoff scores are currently used in research and in practice. Controversy surrounds methods which employ expert judges and also those using test groups to empirically set scores. A minimax procedure developed by Hunyh was illustrated to provide insight into an empirical procedure that invokes the

evaluation of the mathematical properties of various cutoff scores. The example provided is an application of decison theory with underlying binomial probability distribution. The minimax procedure helps to minimize the probabilities of misclassification for false negatives and false positives.

References

Andrews, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36,* 45-50.

Angoff, W. H. (1971). Norms, scales, and equivalent scores. In R. L. Thorndike (ed.). *Educational Measurement* (2nd Ed.). Washington, D.C.: American Council on Education.

Barnett, R. & Ziegler, M., (1990). *Finite Mathematics for business, economics, life sciences, and social sciences.* Riversided, N.J.: Dellen Publishing Company.

Behuniak, P., Jr. Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42,* 247-252.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219-240.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Reinhart and Winston, Inc.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd Ed.). Englewood Cliffs, N.J.: Prentice-Hall.

Hunyh, H. (1976). Statistical considerations of mastery scores. *Psychometrika, 41,* 65-78.

Hunyh, H. (1980). A nonrandomized minimax solution for passing scores in the binomial error model. *Psychometrika, 45,* 167-182.

Hunyh, H., & Saunders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 17,* 351-358.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Saunders, J. C., Ryan, J. P., & Hunyh, H. (1981). A comparison of two approaches to standard setting based on the Nedelsky procedure. *Applied Psychological Measurement, 5,* 209-218.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda and J. R. Sanders (Eds.). *Practices and problems in competency-based measurement.* National Council of Measurement in Education.

van der Linden, Wim J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard-setting. *Journal of Educational Measurement, 19,* 295-380.