Setting Mastery Test Cut-Scores-an Evaluative Model Approach

Abdelatti A. El-Sayyad

Al-Azhar University, Egypt

H. W. Stoker

Florida State University

Introduction and Review of Literature

The purpose of this study was to investigate procedures for setting cut-scores for mastery tests which would maximize the reliability of mastery classification decision and minimize two types of misclassification errors. An approach for a new model to set cut-scores for mastery tests, especially when a conflict exists between cut-scores for mastery tests, especially when a conflict exists between cut-scores setters, is proposed.

A review of literature regarding the establishment of cut-scores indicates that some doubt exists concerning any realistic establishment of such scores.

Stoker (1976) proposed two classifications of models by which standards have been and are establihsed: Judgemental and empirical. Within the judgmental category, Stoker distinguished between the professional judgment model and the externally imposed standards model. In one, the professionals (e.g., Department of Education staff or university professionals) have agreed on the standards to be met by the learner. The other type of judgmental model applies when some group external to the profession (e.g., legislature, parents, etc.) mandates standards for member of the profession. Empirical models for setting performance standards were defined as the examination of student performance with respect to some criteria.

Jaeger (1976) claimed that all standard-setting is judgmental. However, he identified twò models. One is called the derived model, and the other is called the proximal model or direct model. Jaeger identified ten threats to the validity of inferences based on selected standard-setting models as they are reflected by Lord and Novick (1968), Hambleton and Novick (1973), Nedlsky (1954), Ebel (1974), Novick, Lewis, and Jackson (1973), Millman (1973), Novick and Jackson (1974). Jaeger also claimed that the research that exists on standard-setting procedures appeared to be largely theoretical. He called for an empirical investigation involving human standard setters in real or sim-lated judgmental situations, using real performance data and real descriptions of task domains.

Shepard (1976), in a conclusion similar to Jaeger's (1976), proposed that all standard setting is judgmental. Empirical methods may facilitate judgment making, but they cannot be used to ferret out standards as if they existed independent of human opinions and values. Shephard also pointed out that expert judgment should consider the relative costs of errors of misclassification and adjust the standards for protection against them.

Criterion-Referenced Reliability Indices

Brewer (1978) reviewed issues of the Journal of Educational Measurement, Educational and Psychological Measurement, American Educational Research Journal and Review of Educational Research for a four-year period in a search for reliability indices for criterion-referenced tests. He found that there are two basic types of indices:

1. Those which required two or more administrations of the test or equivalent forms. These methods are found in Carver (1970), Hambleton and Novick (1973) and Millman (1974).

2. Those which can be computed with a single testing administration. These methods are found in Huynh (1976), Marshall and Haertel (1976) (in Subkoviak [1978]), Subkoviak (1976), Brennan and Kane (1977), and Livingston (1972).

In his conclusion, Brewer recommended that the Brennan and Kane(1977) dependability index be used in computing reliability indices since it has a good level of interpretability and, in addition, it is found in generalizability theory rather than in classical test theory. It provides, for the criterion-referenced testing situation, an easily understood indicator of reliability. The larger the index, the easier it is to detect that a student (selected at random) is truly above or below the cut-off score. It can be computed with any available analysis of variance computer sub-routine, eliminating the need for any sophisticated computer programming.

Hambleton, Swaminathan, Algina and Coulson (in Subkoviak, 1978) categorized criterion-reliability indices differently. They distinguished three different concepts of reliability which arise in the context of criterion-referenced testing:

1. Reliability of mastery classification decisions are discussed in Huynh (1976), Marshall and Haertel (1976), Subkoviak (1976), and Swaminathan, Hambleton, and Algina (1974). These articles refer to the degree of consistency with which individuals are designed as master or non-master over repeated testing of the same group.

2. Reliability of criterion-referenced test scores is discussed in Livingston (1972), Brennan and Kane (1977), Brennan (1978). These authors refer to the consistency of deviations from the criterion across parallel forms.

3. Reliability of domain score estimates as explained in Hambleton, Swaminathan, and Coulson (1978). This concept is relevant when the purpose of the test is to estimate the number or porportion of such items that each student can correctly answer, without setting a criterion score and without distinguishing masters from non-masters.

The reliability concept of interest in this study falls under number one above, and, in particular the Subkoviak method (1976). Subkoviak (1978), demonstrated that the method produces estimates having relatively small standard errors for classroom size samples and requires only one test administration. However, when using short tests, it produces biased estimates for cut-scores near the center of the score distribution; over-estimates near the tails of the score distribution also occur. This method is computationally tedious; however, a FORTRAN IV subrouting has been written to eliminate this problem.

Wilcox (1977) and Subkoviak (1978) argued that, in a real world testing situation, it may be of more value to know the extent to which a test leads to correct decisions, with some degree of precision, than to know the extent it leads to consistent, but possibly incorrect, decisions.

To find the optimal cut-score, an index such as Subkoviak's P_0 an index of agreement, would be of little help. In some cases, it could also be misleading, especially where P_0 displays a symmetric relationship with the cut-score and/or the test is short. If the problem is to find the cut-score which maximizes the index, one could set $X_0 = 0$, i.e., everyone passes the test, or set X_0 = maximum, in which case, nearly everyone fails the test.

This is not to say the indices such as P_0 have no value. It is required to make consistent decisions across test repetitions (Wilcox, 1979). It is necessary to compare proposed cut-scores for tests using an index like P_0, but it is not sufficient. Even if one is satisfied with P_0 as a criterion for judging a given cut-score, assuming that the value of P_0 indirectly reflects the seriousness of the misclassification of two type errors α and β (to be defined later), the exact relationship of α and β to P_0 is not known (Wilcox, 1979).

False-Positive and False-Negative Errors

Let X represent a student's observed score, X_0 represent the judgmental cut-score, T represent the student's true score, and T_0 represent the true cut-score. There are two types of errors which might occur in classifying a student: false-positive error, α, which occurs when $X \geq X_0$ and $T < T_0$; and false-negative error, β , which occurs when $X < X_0$ and $T \geq T_0$. Let,

$$\alpha = \Pr(X \geq X_0, T < T_0), \quad (1)$$

$$\rho = \Pr(X < X_0, T \geq T_0) \quad (2)$$

It should be noted that α and β are defined in terms of the group of students (i.e., samples under study) and g(T), the distribution of T is the probability density function of true scores over the population of the students.

Emrick (1971), in an attempt to describe an evaluation model for mastery-testing, used one minus the square root of the average inter-item reliability coefficient as an estimate of ($\alpha + \beta$), given that α and β are defined as conditional probabilities. However, Wilcox and Harris (1977) showed that Emrick's (1971) model suffered great problems, and, so, the usability of this model in setting standards is, at best, questionable.

Wilcox (1976) related Fhaner's (1974) approach for determining the appropriate length of a mastery test to Millman's (1973) approach for determining passing scores and test length, in search of a routine which could be used in setting the appropriate passing score. His routine makes the assumption that, for a given positive constant, say C, we are indifferent as to how an examinee is classified when the level of functioning is in the open interval (T_0 - C, T_0 + C). This routine treats the cut-score determination problem as if it were situation independent, conflicting with the belief that tests and associated cut-score problems are situation dependent.

Wilcox (1977) succeeded in estimating the likelihood of committing α and β errors given that g(T), the true score distribution, is a beta distribution. This assumption, however, is not easy to meet. Also, Wilcox's (1977) model is difficult to implement, since it requires a considerable amount of computer time.

An acceptable solution to α and β estimation problem has been given by Wilcox (1979). Without any assumption regarding the shape of g(T), Wilcox succeeded in estimating an upper and lower bound to α and β , given that the first two moments of g(T) exist.

Methodology and Results

The strategy in this study was to investigate a dynamic process through the use of sample test data from the administration of the 1978 - 1979 Florida Secondary School Achievement Test-II, mathematics section (SSAT).

In 1976, the Florida legislature enacted a law requiring eleventh grade students to pass a functional literacy test before graduation. The definition of functional literacy, approved by the Department of Education on February 17, 1977, is as follows:

> For the purposes of compliance with the Accountability Act of 1976, functional literacy is the satisfactory application of basic skills in reading, writing, and arithmetic, to problems and tasks of practical nature as encountered in everyday life.

In order to pass this test, students were required to attain scores equal to or greater than a 70% cutoff score. This requirement has been attacked by educators and by the general public. Some people thought the 70% level was unrealistic; others questioned the method of establishing the cut-score (Fisher, 1978; Glass, 1978; Fremer, 1978; Anderson and Lesser, 1978).

The sample used in this study can be considered representative of the eleventh grade student body in the State of Florida. The random sample selected contained observations from 148 schools in 59 districts; 72.8% of the sample (728 students) were classified as white, non-Hispanic. In general, the characteristics of the sample matched those of the population.

The test which was administered contained 60 items. The frequency distribution for the sample, N = 1,000 observation, is shown in Table 1, Columns 1 and 2. Column 1 contains possible scores (x_i, i = i, ... 60) and Column 2 contains the frequencies N_{xi} for each of these scores.

The sample scores distribution has a mean, $\hat{M}_x$ = 46.46; variance, σ^2 = 108.91; mode = 57; median = 49.14; krurtosis = 1.55; skewness = 1.18, and range = 60.

The distribution of item difficulty indexes had a mean = .78; variance, σ^2_π = .025; mode = .86; median = .81; and range = .74. The item difficulty index, π , was computed by using the formula π = R/T (Mehrens and Lehaman, 1973), where R = number of students who answered the item correctly, and T = total number of students who tried it. Kuder-Richardson coefficient, α_{20} ,for the test was reported as .921 (1978/79 technical report).

The aim in this study was to attempt to find that cut-score (standard) which would maximize the reliability of mastery classification decisions and minimize the two types of misclassification errors from among those cut-scores proposed by cut-score setters.

Estimating the Reliability of Mastery Classification Decision

The following is a summary of the steps used for computing the Subkoviak reliability of mastery classification decisions index, P_0 for the sample under study with cut-score X_0 = 42 (70%). Procedures and results are illustrated in Table 1.

Table 1

Raw Scores, frequencies, proportions and values

for Subkoviak's method

1	2	3	4	5	6	7	8
X_i	N_{xi}	$\hat{P1}_{xi}$	$Z_{xi}(0,1)$	$P2_{xi}$	1-2 $(P2_{xi}-P2_{xi}^2)$	(2)x(6)	(2)
00	04	.061	20.650	0.000	1.000	4.000	0.00
01	00	.076	18.170	0.000	1.000	0.000	0.00
02	00	.099	16.130	0.000	1.000	1.000	0.00
03	00	.107	14.8400	0.000	1.000	0.000	0.00
04	00	.123	13.640	0.000	1.000	0.000	0.00
05	00	.138	12.630	0.000	1.000	0.000	0.00
06	00	.153	11.760	0.000	1.000	0.000	0.000
07	00	.169	10.990	0.000	1.000	0.000	0.000
08	00	.184	10.320	0.000	1.000	0.000	0.000
09	00	.199	09.710	0.000	1.000	0.000	0.000
10	01	.215	09.160	0.000	1.000	1.000	0.000
11	00	.230	08.650	0.000	1.000	0.000	0.000
12	01	.245	08.180	0.000	1.000	1.000	0.000
13	00	.261	07.750	0.000	1.000	0.000	0.000
14	00	.276	07.350	0.000	1.000	0.000	0.000
15	01	.291	06.960	0.000	1.000	1.000	0.000
16	03	.307	06.610	0.000	1.000	3.000	0.000
17	01	.322	06.260	0.000	1.000	1.000	0.000
18	02	.337	05.940	0.000	1.000	2.000	0.000
19	04	.353	05.630	0.000	1.000	4.000	0.000
20	04	.368	05.330	0.000	1.000	4.000	0.000

Table 1 (Continued)

1	2	3	4	5	6	7	8
X_i	N_{xi}	$\hat{P}1_{xi}$	$Z_{xi(0,1)}$	$P2_{xi}$	1-2 $(P2_{xi} - P2_{xi}{}^2)$	(2)x(6)	(2)x(5)
21	03	.384	05.040	0.000	1.000	3.000	0.000
22	07	.399	04.760	0.000	1.000	7.000	0.000
23	05	.414	04.490	0.000	1.000	5.000	.000
24	03	.430	04.230	0.000	1.000	2.999	.000
25	04	.443	03.980	0.000	1.000	4.000	.000
26	04	.460	03.730	0.000	1.000	3.999	.000
27	11	.476	03.480	0.000	1.000	10.994	.00276
28	12	.491	03.240	0.000	.999	11.986	.007
29	03	.506	03.000	0.001	.997	2.991	.005
30	13	.522	02.770	0.003	.994	12.927	.036
31	10	.537	02.530	0.006	.989	9.887	.057
32	08	.552	02.300	0.010	.979	7.830	
33	10	.568	02.070	0.019	.962	9.623	.192
34	16	.583	01.840	0.033	.936	14.982	.526
35	20	.98	01.610	0.054	.898	17.967	
36	19	.614	01.370	0.085	.844	.034	
37	21	.629	01.140	0.127	.778	16.339	2.670
38	22	.644	00.899	0.184	.700	15.392	4.049
39	18	.660	00.657	0.255	.620	11.167	4.583
40	24	.675	00.411	0.341	.551	13.215	8.18167
41	17	.691	00.159	0.436	.508	8.637	7.419
42	24	.706	00.997	0.540	.503	12.076	12.956
43	37	.721	00.367	0.644	.542	20.041	23.839

Table 1 (Continued)

1	2	3	4	5	6	7	8
X_i	N_{xi}	$\hat{P}1_{xi}$	$Z_{xi}(0,1)$	$P2_{xi}$	1-2 $(P2_{xi} - P2_{xi}{}^2)$	(2)x(6)	(2)x
44	20	.737	00.643	0.739	.614	12.283	14.776
45	31	.752	00.931	0.824	.710	22.001	25.538
46	27	.767	01.230	0.891	.805	21.741	24.049
47	27	.783	01.550	0.939	.886	23.927	25.365
48	37	.798	01.890	0.971	.943	34.890	35.913
49	39	.813	02.250	0.988	.976	38.058	38.523
50	38	.829	02.650	0.996	.992	37.685	37.847
51	44	.844	03.070	0.999	.998	43.906	43.953
52	39	.859	03.550	0.999	.999	38.485	38.992
53	49	.875	04.090	0.999	1.000	48.998	48.999
54	59	.890	04.710	1.000	1.000	59.000	59.000
55	57	.905	05.440	1.000	1.000	57.000	57.000
56	54	.921	06.330	1.000	1.000	54.000	54.000
57	62	.936	07.480	1.000	1.000	62.000	62.000
58	42	.951	09.060	1.000	1.000	42.00	42.000
59	30	.967	11.590	1.000	1.000	30.000	30.000
60	12	.982	16.510	1.000	1.000	12.000	12.000
						897.57	717.26

1. As noted earlier, Columns 1 and 2 of the table contain the score X_i and the frequency of each score N_{xi} for the sample under study.

2. Assuming that the 60 items of the test represent a sample of items from an actual or hypothetical universe of such items, Column 3 contains an estimate of the proportion of items in that universe, $\hat{P}1_{Xi}$, that a student with test score X_i would be expected to correctly answer. In other words, $\hat{P}1_{Xi}$ could be considered as the probability of a correct item response. The values of $\hat{P}1_{Xi}$ in the table are computed via the following formula:

$$\hat{P}1 = \hat{\alpha}_{20}(X_i/n) + (1 - \alpha_{20})(\hat{\mu}_X/n) \qquad (3)$$

where

$\hat{\alpha}_X$ is the sample mean - 46.46

n is the test length - 60 items

$\hat{\mu}_{20}$ is the Richardson coefficient α_{20}= .921; for example, if X_i = 0.0, then $\hat{P}1_{Xi}$ - .06.*

3. Knowing that the probability of the correct response to a single item is $\hat{P}1_{Xi}$ = .06, we can calculate the probability that the student will correctly answer 42 or more items on a 60-item test and classify as a master. If the items can be considered as a trial in a binomial process, the probability of 42 or more success in n = 60 trials is $\hat{P}2_{Xi}$ = 0.0 for such a student, as it is indicated in Column 5.

The probabilities in Column 5 are computed using a normal probability table (Gilford, 1954). Column 4 contains the normal z values with mean 0 and standard deviation 1. Of

* In actual computation, decimal accuracy beyond the two decimals reported was maintained. Results have been rounded for simplicity in reporting.

course, as the distribution of X_i scores departs from normality, one would expect the accuracy of $\hat{P}2_{xi}$'s estimates to decrease. (See Figure 2 for the frequency distribution of the observed score.)

4. The probability that the above student will be consistently classified as a master on two independent testings is $\hat{P}2_{xi}$, and conversely, the probability that this student will be consistently classified as a non-master is $(1 - \hat{P}2_{xi})$. The probability of consistent classification for this student is:

$$\hat{P}2_{xi} + (1 - \hat{P}2_{xi})^2 = 1 - 2(\hat{P}2_{xi} - \hat{P}2^2_{xi}) \qquad (4)$$
$$= 1 - 2(0 - 0) = 1$$

as indicated in Column 6.

5. The probability of consistent classification across the entire group P_0 is obtained from the total of Column 7,

$$\hat{P}_0 = \Sigma N_{xi}[1 - 2(\hat{P}2_{xi} - \hat{P}2_{xi})] /N \qquad (5)$$
$$= 897.57/1000 = .898$$

6. The chance probability of consistent classification, $\hat{P}_{ch}$, is obtained from the total of the last column,

$$\hat{P}_{ch} = 1 - 2\left[\frac{\Sigma N_x \hat{P}2_{xi}}{N} - \left(\frac{\Sigma N_x \hat{P}2_{xi}}{N}\right)^2\right] \qquad (6)$$
$$= 1 - 2[(717.26/1000) - (717.26/1000)^2]$$
$$= .59$$

7. Cohen's (1960) kappa coefficient, $\hat{k}$, was calculated as suggested by Swaminathan, Hambleton, and Algina (1974):

$$k = \frac{\hat{P}_0 - \hat{P}_{ch}}{1 - \hat{P}_{ch}}$$

$$= \frac{.89757 - .59441}{1 - .59441}$$

$$= .747$$

The $\hat{k}$ can be interpreted as a proportion of consistent classification beyond that expected by change.

These seven steps were repeated for different cut-scores. Table 2 contains the value X_0, P_0, P_{ch} and $\hat{k}$ for different cut-scores X_0's. Figure 2 displays the observed empirical relationship between X_0, P_0 and $\hat{k}$.

As shown in Figure 2, $\hat{P}_0$ displays a curvilinear relationship with cut-scores X_0. It reached its maximum at $X_0 = 0$ and $X_0 = 60$. The graph also suggests that P_0's curve might be a mirror image for the observed score frequency distribution. Similar findings have been reported by Subkoviak (1978).

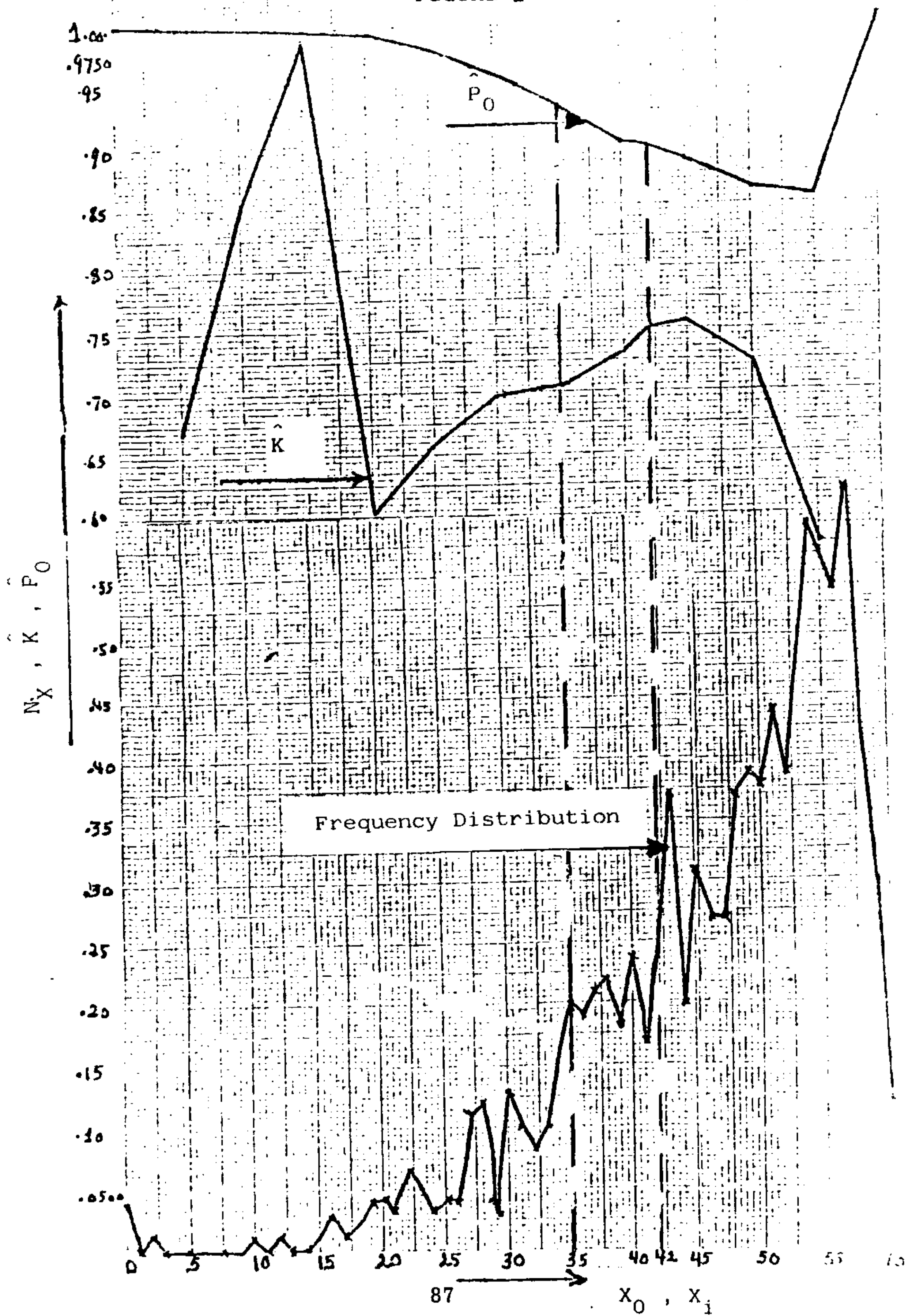
The $\hat{k}$ behaves in an unpredictable manner for the low cut-scores. However, in general, it has an inverse relationship with $\hat{P}_0$, and is undefined at $X_0 = 0$, and $X_0 = 60$. For more discussion about $\hat{k}$ and its relation to some criterion-referenced test indices, the reader is referred to Subkoviak (1978) and Logsdon (1979). It appears that the shape of

Table 2

$\hat{P}_0$, $\hat{P}_{ch}$ and $\hat{k}$ for Different Cut-Scores, X_0

Cut-Score X_0	$\hat{P}_0$	$\hat{P}_{ch}$	$\hat{k}$
0	.999	.998	
5	.997	.992	.671
10	.998	.989	.852
15	.995	.985	.985
20	.987	.967	.608
25	.975	.927	.658
30	.958	.862	.697
35	.931	.763	.710
40	.903	.639	.732
42	.898	.594	.747
45	.886	.536	.755
50	.862	.514	.725
55	.849	.670	.572
60	.983	.983	

FIGURE 1



1.00
.9750
.95
.90
.85
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
.15
.10
.0500
N_X , $\hat{K}$, $\hat{P}_O$
$\hat{P}_O$
$\hat{K}$
Frequency Distribution
0
5
10
15
20
25
30
35
40
42
45
50
55
60
87
X_O , X_i

the observed score frequency distribution has a great impact on $\hat{P}_0$ and $\hat{k}$.

Cut-Scores and α and β Errors

The model to be developed in this study assumes that the cut-score will first be determined judgmentally. This judgment represents the input to the model.

Suppose that a professional judgment group (PJG) decides that the cut-score X_0 should be 42 (70%). Further, suppose that a non-professional judgment group (NPJG) decides that the cut-score X_0 should be 35 (58.34%). Now, one can assume that a conflict exists between the two groups. Using the $\hat{P}_0$ - X_0 relationship as a criteria, the lower cut-score should be preferred since $\hat{P}_0$ (35) = .93119 and P_0 (42) = .89757.

From a decision-maker's point of view, the two judgment cut-scores are different not just because one is preferred over the other, because of high $\hat{P}_0$, but, for a defined loss function, the cut-scores will also differ.

The task is not only to accept an index such as P_0, but, also, to consider the associated estimates of α and β . In this study, the Wilcox (1979) model was used in establishing α and β upper and lower bounds. Then, an attempt was made to find an optimal cut-score which would minimize the loss function:

$$L\ (\alpha,\beta) = L_1 = L_2 \qquad (7)$$

where L_1 and L_2 are constants representing the losses (costs) associated with α and β , respectively.

Suppose A_1 = the event $X \geq X_0$,

A_1^C = the event $X < X_0$,

A_2 = the event $T \geq T_0$, and

A_2^C = the event $T < T_0$.

Let μ and ${\sigma_x}^2$ represent the mean and variance of the true scores of the student. In practice, these can be estimated from therandom sample observed scores X_i, i = i, 2, ..., N, examinees. To get more accurate estimations for α and β one must assume a two-term approximation to the compound binomial distribution has been implemented.

Given the above considerations, the Wilcox (1979) model will be clarified and summarized in the following:

1. Set the known constant true criterion cut-score T_0.
2. Compare $\hat{\mu}$,$\hat{\sigma}^2$, Pr(A1), m, where

$$\hat{\mu} = (Nn)^{-1}\Sigma X_i \text{ , } i = 1, 2, \ldots N \ (=1000), \tag{8}$$

$$\hat{\sigma}^2 = {}_x^2 - (n - 2d)\ \hat{\mu}\ (1 - \hat{\mu})/[n(n - 1) + 2d], \tag{9}$$

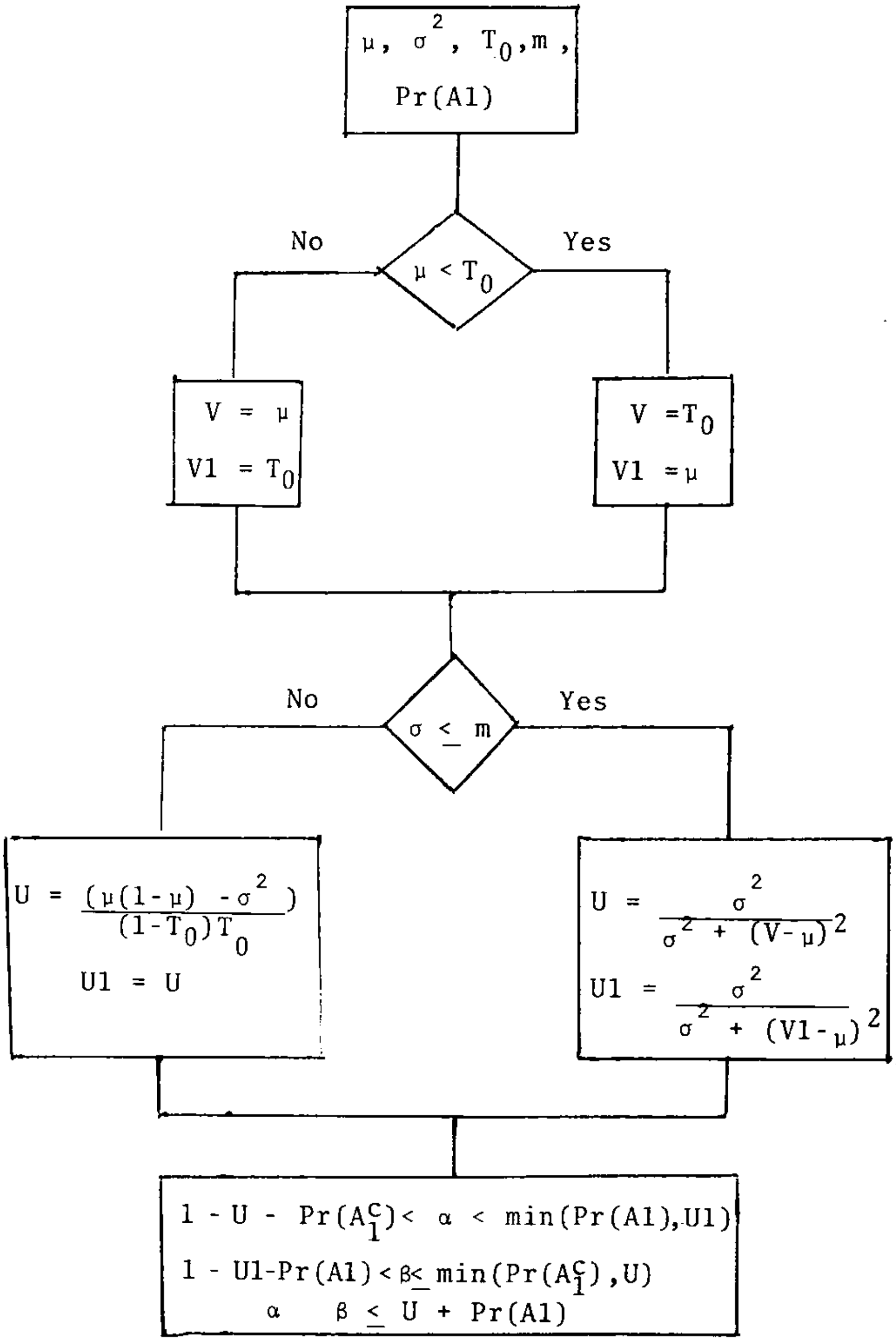
$$d = \frac{n^2\ (n - 1)\sigma_\pi^2}{2[\mu_x\ (n - \mu_x) - {\sigma_x}^2 - n\sigma_\pi^2\ ]} \tag{10}$$

${\sigma_x}^2$ and μ_x are the variance and mean of observed scores and where $\sigma\pi^2$ is the variance of item difficulties, Pr(A1) could be estimated as if it is the proportion of examinees passing the test, and

$$m = \max\ [\mu\ (T_0 - \hat{\mu}\ ),\ (\hat{\mu} - T_0)\ (1 - \hat{\mu})] \tag{11}$$

3. Follow the flowing flowchart (Figure 2) to compute α and β upper and lower bounds.

Figure 2


μ, σ^2, T_0, m ,
Pr(A1)
No
$\mu < T_0$
Yes
$V = \mu$
$V1 = T_0$
$V = T_0$
$V1 = \mu$
No
$\sigma \leq m$
Yes
$U = \frac{(\mu(1-\mu) - \sigma^2)}{(1-T_0)T_0}$
U1 = U
$U = \frac{\sigma^2}{\sigma^2 + (V-\mu)^2}$
$U1 = \frac{\sigma^2}{\sigma^2 + (V1-\mu)^2}$
$1 - U - Pr(A_1^c) < \alpha < \min(Pr(A1), U1)$
$1 - U1 - Pr(A1) < \beta \leq \min(Pr(A_1^c), U)$
$\alpha \quad \beta \leq U + Pr(A1)$

One major difficulty with the above model is setting the fixed value, T_0. A solution must be found. Before we suggest solutions, it is important to conisder what would be an acceptable definition for the true percentage, T.

Wilcox (1976, 1977) defined T as "the percentage of items in the domain of items that an examinee would answer correctly, if all items were administered. With respect to the domain of items, an examinee is to be considered "master" if $T \geq T_0$ and "non-master" if $T \leq T_0$, where T_0 is the known constant with a value between zero and one." This means that T_0 might be considered as one of the T values.

Nunnally (1967) defined the true score T' as "the unbiased scores [which] are the average scores people would obtain if they were administered all possible tests from a domain, holding constant the number of items randomly drawn for each." Their true scores might be estimated as follows (Nunnally, 1967; Novick, 1973):

$$T' = Xr_{xx} + \hat{\mu}_x (1 - r_{xx}) \qquad (12)$$

where

X is the observed score, r_{xx} is the reliability coefficient estimation, and $\hat{\mu}_x$ is the sample observed mean score.

Equation (11) could then be written as

$$\frac{T}{n} = \frac{X}{n} r_{xx} + \frac{\hat{\mu}_x}{n} (1 - r_{xx})$$

n is the number of items in the test, hence

$$T_0 = X_0 r_{xx} + \hat{\mu}_{xp} (1 - r_{xx}) \qquad (13)$$

Equation (12) might now be used to estimate the true proportion T for a given observed proportion X_0 and μ_{xp}, the mean (in proportion) of observed scores.

Now, for the assumed proportion X_0, it would be possible to estimate its corresponding T_0, using the above equation. Note that X_0 represents a group judgment cut-score which, according to our model, will be considered as an input.

One way to set T_0 is described as follows: Assume there are two groups interested in setting a cut-score. Group 1 believes that X_{01} should be the test cut-score. Group 2 believes that X_{02} should be the test cut-score, $X_{01} \neq X_{02}$, and a conflict exists between the two groups. One method for evaluating this situation is to consider T_{01} as it is estimated by equation (13). This might be called *self group true cut-score*. Determine and then compute and for this cut-score, X_{01}. The same could be done for Group 2. One then compares the two group standards in a manner to be described later.

An alternate method is to treat X_{01} as if it were T_{02} and X_{02} as if it were T_{01}. In this way, one can evaluate a given group's observed cut-score in the light of the other group's suggested cut-score. This approach might be called *other group cut-score as a true score*.

An example should clarify the methods. Let Group 1 represent the professional judgment model, and Group 2 represent the external judgment model. Let X_{01} = 42 (70%), the recent SSAT cut-score and $X_{02} = 35$ (58.4%. It is important to know the risks associated with each judgment cut-score.

Self-Group True Cut-Scores -- Using $X_{01} = .70$ and $X_{02} = .584$, solve equations (8) and (10).

Group 1	Group 2
$\hat{\mu} = .774$	$\hat{\mu} = .774$
$d = 5.118$	$d = 5.118$

Since $d > 4$, set $d = 4$ as suggested by Wilcox (1979) and solve equation (9).

$\hat{\sigma}^2 = .028$	$\hat{\sigma}^2 = .028$

The true proportion can now be estimated by equation (13).

$T_{01} = .706$	$T_{02} = .599$

Using equation (11),

$M_{01} = .015$	$M_{02} = .043$

Pr(A1) might be estimated as the proportion of examinees who passed the test.

$Pr(A1)_{01} = .728$	$Pr(A1)_{02} = .869$

Following the flow chart in Figure 3, the upper and lower bounds for α and β are calculated for each cut-score. These values appear in Table 3.

Other Group Cut-Score as a True Cut-Score

$X_{01} = .70$	$X_{02} = .58$
$T_{01} = X_{02} = .584$	$T_{02} = X_{02} = .70$
$\hat{\mu} = .774$	$\hat{\mu} = .774$
$\hat{\sigma}^2 = .028$	$\hat{\sigma}^2 = .028$
$Pr(A1)_{01} = .728$	$Pr(A1)_{02} = .869$

Again, following the flow chart in Figure 2, the upper and lower bounds for α and β are calculated for each cut-score. These values appear in Table 4.

The above four sets of results are summarized in Tables 3 and 4, which, in combination, show that for a given judgmental cut-score, X_0, β error remains the same regardless of any change on the known fixed value, T_0. However, α error is a function of both X_0 and T_0. For a given X_0, decreasing T_0 will decrease ; for a given X_0, increasing T_0 will increase α .

Table 3 can be used in answering such questions as: In terms of α and β values, what is the optimal cut-score to be selected from among different cut-scores proposed by different independent groups (i.e., no interaction assumed between groups)? The answer to this question, according to Table 3, is X_{02} = .584 (35 items correct from 60) would be the optimal cut-score since it has minimum boundaries for α and β errors. It is interesting to note that X_{02} has a higher reliability of mastery classification decision ($\hat{P}_{02}$ = .9311) than X_{01} ($\hat{P}_{01}$ = .89761). This means, according to the Table 3 strategy, that the optimal cut-score X_0 has the higher $\hat{P}_0$.

Table 3 clearly shows the perference of X_{02} over X_{01}, through the magnitudes of the α and β values. The slightly

Table 3

Upper and Lower Bounds for α and β Self Group True Scord

Judgmental Cut-Score X_0	Self T_0 as Estimated by Equation	α		β	
		Lower Bound	Upper Bound	Lower Bound	Upper Bound
X_{01} = .70	T_{01} = .70	.0217	.70630	0	.272
X_{02} = .58	T_{02} = .60	0	.47812	0	.131

Table 4

Upper and Lower Bounds for α and β Other Group Cut-Score as a True Score

Judgmental Cut-Score X_0	Other Group, X_0 as an Estimation for T_0	α		β	
		Lower Bound	Upper Bound	Lower Bound	Upper Bound
X_0 = .70	T_0 = .58	0	.4373	0	.272
X_{02} = .58	T_{02}= .70	0	.69229	0	.131

higher $\hat{P}_0$ supports this preference. Since differing cut-scores could have the same $\hat{P}_0$, it appears that α and β values should be reported as important characteristics of the test. In using the strategy reflected in Table 3, there is no need to define a loss function in order to choose one cut-score over another. In other words, the costs associated with α and β need not be known.

The results in Table 4 are somewhat different and may be more complicated. In this table, we assume that for the Group 1 observed cut-score, X_{01}, the associated true known cut-score, T_{01}, is the Group 2 cut-score, X_{02}, and vice versa. Each group claims that it is proposed cut-score should be the true one, the one which should be used in decision making. In other words, a conflict exists.

Table 4 shows that if $X_0 > T_0$, this will lower α and raise β , and if $X_0 < T_0$, α will be higher and β will be lower. In terms of α values, X_{01} would be considered the optimal cut-score. However, in terms of β values, X_{02} would be the optimal one. Thus α and β errors should both be considered in searching for an optimal cut-score.

Table 4 allows us to examine the risks associated with proposed cut-scores, but does not necessarily show a preference for one cut-score over another. To establish a preference, one needs to examine the loss function, as defined in equation (7).

Let us assume we are looking for that optimal cut-score which will minimize that loss function, assuming that α and β will be at the maximum, i.e.,

For $X_{01} = .7$, $T_{01} = X_{02} = .58$

$\alpha = .4373$

$\beta = .272$

and

For $X_{02} = .58$, $T_{01} = X_{01} = .7$,

$\alpha = .69829$

$\beta = .131$

Now we have

$$L_{X_{01}} (\alpha,\beta) = L_1 (.4373) + L_2 (.272) \qquad (14)$$

$$L_{X_{01}} (\alpha,\beta) = L_1 (.69819) + L_2(.131) \qquad (15)$$

In order to prefer one cut-score over the other, the values of L_1 and L_2 have to be defined or at least the ratio L_1/L_2 has to be known.

Since $\alpha = \Pr(X \geq X, T < T)$, it might be called the error of awarding a false high school diploma. To some extent, it represents a social loss, and so L_1. Since $\beta = \Pr(X < X, T \geq T)$, it might be called the error of failing a truly competent student (no diploma). To a great extent, it represents a personal loss, and so L_2.

The type of error which might lead to a challenge of the cut-score by some individuals is β error. However, this error, as it is reflected by Table 3 and Table 4 is far less than α . In general, if there is reason to believe that this study sample is a homogeneous sample, the results show that for the proposed two cut-scores α is more than twice β . In particular, for the 70% cut-score, the societal risk (α) is higher than individual risk (β). This does not necessarily mean that the 70% level is optimal.

To see how the values of L_1 and L_2 affect cut-score determinations, we will try different combinations for L_1 and L_2. For example,

Set 1: $L_1 = L_2 = C$

$$L_{X_{01}} (\alpha, \beta) = .71C,$$

$$L_{X_{02}} (\alpha, \beta) = .83C$$

which means that $X_{01} = .7$, given that $T_{01} = .584$ should be preferred.

Set 2:

$$L_1 = .5$$

$$\begin{aligned} L_{X_{01}} (\alpha, \beta) &= .5L_2 (.4373) + L_2 (.272) \\ &= .48L_2 \end{aligned}$$

$$\begin{aligned} L_{X_{02}} (\alpha, \beta) &= .5L_2 (.69829) + L_2 (.131) \\ &= .48L_2 \end{aligned}$$

which means that $X_{02} = .584$, given that $T_{02} = .7$ should be preferred.

Conflict Resolution

Assume, now, that the results of these computations will be presented to the two groups of standard setters. We could hope that one of the two groups will be convinced that the other group's proposed cut-score is the one that should be used. If this happens, the agreed upon cut-score should be called the final output and the conflict is over. If conflict still remains, the two groups could be asked to adjust their initial proposed cut-scores. These new, adjusted cut-scores could be considered new input to the model; the new input would be evaluated and the results presented to the groups in conflict. This would continue until an agreement has been reached between the group setters.

An Approach for a New Model

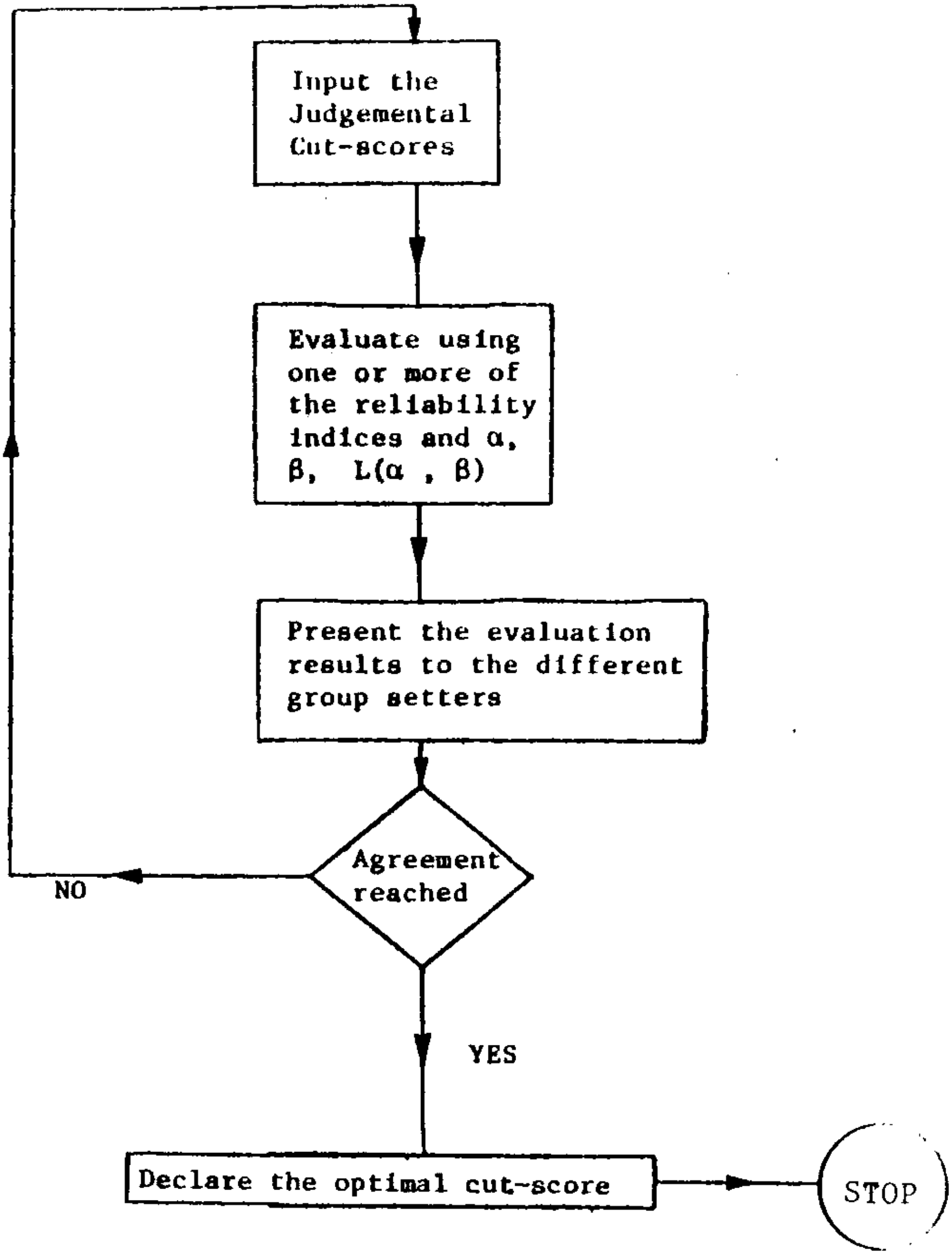
The possibility exists that standard-setting models cannot be classified as either judgmental or empirical. It would be a mistake to believe that professional educators will reflect the values held by parents and legislators. It is also a mistake to believe that parents and legislators (non-professional educators) should dictate what the standards should be. However, every group has the right to share in the formulation of the standards in an interactive way. Tests and testing are situation dependent. They are also time dependent. Hence, no standard should be fixed for all time; evaluation should take place every time the standard is to be used.

In real world testing situations, one might view the standard-setting model as an interactive process, usually beginning with the input of a judgmental standard. This input could be empirically evaluated, and the results presented to standard-setters' groups. Each group could adjust its original proposed standard in light of the evaluation results and then, perhaps, suggest a new standard. The new standard would be considered as new input for the model, and the process continue in this fashion until a kind of stability (agreement) is reached among the standard-setters. The output process is judgmentally assumed, empirically evaluated and has the agreement of the different setters' groups. Note that the process is a dynamic one, i.e., it is time dependent. In other words, cut-scores which seem acceptable this year, for example, may not be acceptable next year.

The model which is proposed appears as Figure 3.

When should this model be used? It could be used after testing, for the purpose of evaluation. In the above example, one cut-score existed for the test. The purpose was to see what values exist for $\hat{P}_0$, α , β , and $L(\alpha,\beta)$, which will tell us what happened in the previous testing. But this use for the model is to be considered a weak one, unless the evaluation might suggest that we should change the existing cut-score. This might change the status (master or non-master) of some students. For example, for the 1978/79 SSAT, the cut-score was 70%. Suppose the model was applied to this cut-score and the results indicated that it should be altered. In this case, students have already been classified according to the 70% cut-score, but our evaluation indicates that 70% should not be the cut score. If students

Figure 3


Input the Judgemental Cut-scores
Evaluate using one or more of the reliability indices and α, β, L(α , β)
Present the evaluation results to the different group setters
Agreement reached
NO
YES
Declare the optimal cut-score
STOP

are reclassified according to the new cut-score, this could be considered an appropriate application of the model.

The model could be used before testing; i.e., for planning for a coming test administration. One could use past years' data to recommend an appropriate cut-score, assuming that the coming year test score distribution will have the same characteristics as past year ones. Or, one could administer the new test, or a parallel form, to a representative sample of students and then apply the model to the sample data in order to recommend an optimal cut-score.

Summary

An evaluative model for setting mastery tests standards has been proposed. This model accepts judgmental standards as input. This input is evaluated using the reliability of mastery classification decisions index $\hat{P}_0$, upper and lower bounds to α and β , and $L(\alpha,\beta)$. The optimal cut-score is the one which minimizes the loss function, $L(\alpha,\beta)$ and maximizes $\hat{P}_0$.

The index of mastery test reliability indices used in this model is the Subkoviak index, $\hat{P}_0$. However, the model is not restricted to this index. It provides an example of how one can use the relationship between a given reliability index (in this case $\hat{P}_0$) and cut-scores in determining an optimal cut-score. The reliability index to be used is left to the model user.

The results showed that the sensitivity of $\hat{P}_0$ to the change in cut-score is very low, especially when n, the test length, is high. However, the sensitivity of α and β to the change in cut-score is much higher than for $\hat{P}_0$.

Many questions regarding the establishment of cut-scores remain to be answered. Included among the many are:

- Is there a "best" criterion-referenced reliability index to be used in evaluating a given test for a given purpose?
- What are the relationships between criterion-referenced reliability index, cut-score, α error and β error and the shape of the observed test score distribution?
- What is the relationship between L_1 and L_2, the two losses associated with α and β error, respectively?

Although these and other questions remain and may take some time to answer, the model proposed could help those responsible for establishing cut-scores make better decisions, particulary in situations where conflicts between groups exist or can be anticipated.

References

Anderson, B. D. & Lesser, P. The costs of legislated minimum competency requirements. Kappan, May, 1978.

Brennan, R. L. & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, Vol. 14, 1977.

Brennan, R. L. Some applications of generalizability theory to the dependability of domain-referenced tests. Unpublished manuscript, American College Testing Program, Iowa City, Iowa, 1978.

Brewer, J. K. A review of criteria-referenced reliability indices, Assessment Section, Bureau of Program Support Services, Florida Department of Education, SDE #780-136, August, 1978.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative Research: Strategies and Methods, Washington American Institutes for Research, 1970.

Cohen, J. A. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Vol. 20, 1960.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N. J., Prentice-Hall Publishing Co., 1974.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, Vol. 8, 1971.

Fhane, V. S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, Vol. 27, 1974.

Fisher, T. H. Florida's approach to competency testing. Kappan, May, 1978.

Fremer, J. In response to Gene Glass. Kappan, May, 1978.

Glass, G. V. Minimum competence and incompetence in Florida. Kappan, May, 1978.

Haladyna, T. Comments: Measurement issues related to performance standards. Florida Journal of Educational Research, Vol. 18, 1976.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, Vol. 10, 1978.

Hambleton, R. K., Swaminathan, H., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, Vol. 48, 1978.

Huynh, H. On the reliability of decisions in domain referenced testing, Journal of Educational Measurement, Vol. 13, 1976.

Jaeger, R. M. Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, Vol. 18, 1976.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, Vol. 9, 1972.

Longsdon, David M. A study of the meaningfulness of criterion and norm-referenced reliability in assessing the validity of mastery rates. Unpublished doctoral dissertation, Florida State University, 1979.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Addison-Wesley Publishing Company, 1968.

Marshall, J. L. & Haertel, E. H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1976.

Mehrens, W. A. & Lehamann, I. J. Measurement and evaluation in education and psychology. Holt, Rinehart and Winston, Inc., 1973.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) *Evaluation in Education*. Berkeley, California: McCutchan, 1974.

Nedelsky, L. Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 1954.

Novick, M. R., Lewis, D. and Jackson, P. H. The estimation of proportions in m groups. *Psychometrika*, 1973.

Novick, M. R. and Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw Hill, 1974.

Nunnally, J. *Psychometric Theory*. New York: McGraw Hill, 1974.

Shepard, Loretta A. Setting standards and living with them.

Florida Journal of Educational Research, Vol. 18, 1976.

Stoker, Howard W. Performance standards in competency-based education. *Florida Journal of Educational Research*. Vol. 18, 1976.

Subkoviak, M. J. Estimating reliability from a single administration of mastery test. *Journal of Educational* Measurement. Vol. 13, 1976.

Subkoviak, M. J. *The Reliability of Mastery Claffification Decision*. Unpublished manuscript, University of Wisconsin, 1978.

Swaminathan, H. & Hambleton, R. K. & Algina, J. Reliability of criterion referenced tests: A decision theoretic formulation. *Journal of Educational Measurement*, Vol. 11, 1974.

Wilcox, R. R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, Vol. 1, Winter, 1976.

Wilcox, R. R. On False-Positive and False-Negative Decisions with a Mastery Test. *Journal of Educational Statistics*, Vol. 4, Winter, 1979.

Wilcox, R. R. & Harris, Chester W. On Emirck's "An evaluation model for mastery testing." *Journal of Educational Measurement*, Vol. 14, 1977.