DOCUMENT RESUME

ED 205 604 TM 810 514

AUTHOR	Yeh, Jennie P.; Moy, Raymond
TITLE	Comparison of Program Effects: The Use of Mastery Scores.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
REPORT NO	CSE-R-149
PUB DATE	Nov 80
NOTE	25p.
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Classification: Criteria: *Cutting Scores: English (Second Language): Higher Education: Language Tests: Mastery Tests: *Models: *Testing Problems: Test Norms: True Scores

ABSTRACT

The setting of a cut-off score on a mastery test usually involves a consideration of one or more of the following elements: (1) the distribution of observed test scores: (2) the type of mastery criterion used: (3) the level of acceptable risks of mis-classification: (4) the loss of functions of mis-classifications: and (5) the distribution of true scores. In order to illustrate several procedures for setting cut-off scores, and how various considerations may change the cut-off score value, a data set was obtained consisting of 99 foreign engineering graduate students' test scores on a sample of 87 items from the UCLA English as a Second Language proficiency test, their grade point average, the number of university courses failed, and percentile scores from the Graduate Record Examination. The methods examined include a norm-referenced vs. a theoretical criterion: cut-off scores based on acceptable risks of mis-classification: Huynh's optimal decision rule model: Wilcox's optimal cut-off score based on observed scores and an external criterion: and Wilcox's method for approximating true score distribution. (Author/BW)

COMPARISON OF PROGRAM EFFECTS: THE USE OF MASTERY SCORES

Jennie P. Yeh

Raymond Moy

CSE Report No. 149

November, 1980

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education, UCLA
Los Angeles, California 90024

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Departmentof Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

INTRODUCTION

Classical psychometric theory is based on the notion that the purpose of educational and psychological assessment is to sort students or grade them from excellent to poor (Tyler and White, 1979). Recent developments and interest in adaptive instructional systems such as Individually Prescribed Instruction (Glaser, 1968), and minimum competency testing call for new procedures focusing on the evaluation of individual performance in terms of mastery. A test is purposely constructed to give scores that reflect what a student can or cannot do. Based on a student's observed test score, he or she is classified, in a simple two category case, in either the "mastery" or the "non-mastery" group for a skill. For example, as a master he or she may proceed to the next unit or receive a diploma, and as a non-master he or she may receive remedial work. Decision procedures tend to fall into two categories: mastery status is granted if either the subject's observed test score exceeds a minimum level, or the probability is reasonably high that his or her true score is beyond a given standard. In both cases, the dividing line between masters and non-masters is called the cut-off score, mastery score, or criterion. In making decisions about an examinee's mastery status, how far the examinee is from the cut-off score is of no concern. Instead, the main concern is whether the examinee is above or below the cut-off score. Therefore, one essential task in competency testing is to locate a <u>valid</u> cut-off score which will classify individuals into categories representing their true mastery status.

Cut-Score Models

At this stage of development, the setting of a cut-off score on a mastery test usually involves a consideration of one or more of the following elements: (1) the distribution of observed test scores; (2) the type of mastery criterion used; (3) the level of acceptable risks of mis-classification; (4) the loss of functions of mis-classifications; and (5) the distribution of true scores.

Perhaps the most ad hoc method of setting a cut-off-score is to look at the distribution of observed scores and pass either some upper proportion of the examinees or select a cut-off point at some reasonable break in the distribution (such as between two modes or above or below one tail of a skewed distribution). Over a succession of test administrations, these procedures may lead to impressions of expected performance and a substantive feel for what such a cut-off score standard means. However, this method of setting a cut-score is basically a norm-referenced decision and actually avoids the mastery/non-mastery decision problem.

True mastery can only be determined in terms of a criterion which has been established on an empirical or a theoretical basis or both. For example, a theoretical criterion proposed by Nedelsky (1954) for multiple choice tests is established in the following manner: distractors which the lowest passing student should be able to reject are identified for each item and the reciprocal of the remaining distractors if the minimum passing level (MPL). A summation of these MPL's is a theoretical minimum passing score for the overall test.

Alternatively, one can identify a criterion such as observable success in a closely related task and a cut-off-score can be chosen so that the number of mis-classifications is minimized. Such mis-classifications can be of two types: (1) false positives, reflecting those who are non-masters on the criterion but are classified as masters by the test; and (2) false negatives, reflecting those who are masters on the criterion but who are classified by the test as non-masters. If one uses observed scores and a criterion has been selected in terms of mastery ability θ, where $0 \leq \theta \leq 1$, one would want to adjust the cut-off score according to the level of acceptable risks associated with each of the two types of misclassification. For example, a school may be willing to admit non-masters to its program--but only up to 10% of the overall enrollment--while it does not wish to turn away more than, say, 20% of the true masters who apply for participation. A cut-off score could then be chosen such that the compound binomial probability of mis-classification for a given ability parameter of true mastery would not exceed the established risk levels. A solution to this problem, of course, depends on having a sufficient number of test items. Stig Fhaner (1974) poses the problem as follows.

Find the critical score C such that

$$P(x>C|\theta_1) = \sum_{x=C+1}^{n} \binom{n}{x} \theta^x(1-\theta_1)^{n-x} \leq \alpha \qquad (1)$$

$$P(x\leq C|\theta_2) = \sum_{x=0}^{C} \binom{n}{x} \theta_2^x(1-\theta_2)^{n-x} \leq \beta$$

where θ_1 = universe score definitely insufficient for passing

θ_2 = universe score definitely sufficent for passing

α = tolerable risk of accepting a non-master

β = tolerable risk of rejecting a master

n = number of test items

x = observed score

Related to these risk levels are measures of loss associated with each type of mis-classification. Losses can be specified in terms of time or costs. For example, the losses associated with admitting a non-master might be loss of training costs or time wasted in pursuing a non-successful endeavor. Losses associated with rejecting a master might involve postponement of societal benefits, loss of institutional revenue, or time wasted on needless remedial training. If the losses can be specified, then the mastery score problem becomes one of finding that score which will minimize them. Huynh (1976) incorporates the probability of success on a referral task into determining a rule allowing for an optimal decision. He specifies the loss function $(R(C))$ to be minimized as follows:

$$R(C) = \int_{\Omega}\int_{x \geq 0} C_f(\theta)[1-S(\theta)]p(\theta)f(x|\theta)dxd\theta + \int_{\Omega}\int_{x \geq 0} C_s(\theta)S(\theta)p(\theta)f(x|\theta)dxd\theta$$

where

$C_f(\theta)$: loss of granting mastery status to a failure

$C_s(\theta)$: loss of assigning non-mastery status to a success

$S(\theta)$: probability of success on a criterion

$f(x/\theta)$: probability density function of observed scores given θ

θ: universe score of ability $0 \leq \theta \leq 1$

c: Cut-Score c

$P(\theta)$: probability density function of θ

The minimization of the double integral and solutions for the cut-score c can be approximated if a beta distribution is assumed for the ability θ and the binomial distribution of observed scores is approximately described by the normal distribution (large n and parameter θ not near 1 or 0). Also, the loss ratio C_s/C_f must be constant and the functions S(θ) close to a 0-1 form. The solution can then be expressed as,

$$c = (n+\alpha+\beta-1)t_o + z\sqrt{(n+\alpha+\beta-1)t_o(1-t_o)} \; -\alpha+.5$$

where,

α, β: are parameters of the beta distribution

t_o: the value of θ associated with true mastery

z: 100/1+Q percentile of the unit normal distribution

Q: C_s/C_f

In summary, many different approaches to setting cut-off scores have been advanced. The purpose of the present research was to compare the results derived from the various approaches.

Applications of Models

In order to illustrate several procedures for setting cut-off scores, and how various considerations may change the cut-off score value, a data set was obtained consisting of 99 foreign engineering graduate students' test scores on a sample of 87 items from the UCLA English as a Second Language proficiency test, their GPA, the number of university courses failed, and GRE percentile scores (Table 1). Since the ESL test was administered to determine if remedial English courses were required for successful performance in graduate work, GPA and number of courses failed were used as external criteria of English mastery. However, it is acknowledged that, in addition to language profic[illegible]y, achievement in graduate work is highly dependent on other factors such as previous preparation in related work, amount of effort, quality of instruction.

TABLE 1

Means and Standard Deviations of ESL Data

Variable Name	$\overline{X}$	σ
1. General GPA	3.45	0.38
2. YR 1 GPA	3.43	0.38
3. GRE Verbal	15.55	17.45
4. GRE Quantitative	88.89	9.74
5. GRE Advanced	58.55	26.71
6. ESL Score	59.84	157.50

Norm-Referenced vs. Theoretical Criterion

Based on the past few years' records, approximately 26 percent to 30 percent of the students taking the ESL exam each year are declared proficient enough to take university courses without remedial English courses. For the 87 item test considered here, the upper 30th percentile corresponds to a test score of 69. This percentile score was based on a total of 1150 students, university wide, of which the 99 engineering graduate students were a sub-group. Although no theoretical mastery cut-off score is explicitly stated by the test-makers, it does appear that exemption status is associated with at least the ability to answer 75 percent of the items correctly. If such a proportion of correct answers is used as the theoretical mastery criterion, then minimal competency is associated with a score of 66 or above. These different criteria result in different classifications of mastery/non-mastery status according to normed placement (cut-score set as the 26th and 30th percentiles) or theoretical criterion (Table 2 and Table 3).

TABLE 2

Cross Tabulations of Mastery/Non-Mastery by Normed Placement (Upper 30th Percentile)

Theoretical Criterion, 75th Percentile c=66

ESL Normal Placement Upper 30th Percentile c=68		Mastery	non-Mastery	
	Mastery	31	0	31
	non-Mastery	7	61	68
		38	61	99

TABLE 3

Cross Tabulations of Mastery/Non-Mastery by Normed Placement (Upper 26th Percentile)

Theoretical Criterion, 75th Percentile c=66

ESL Normal Placement Upper 26th Percentile c=69		Mastery	non-Mastery	
	Mastery	29	0	29
	non-Mastery	9	61	70
		38	61	99

The results of these cross tabulations indicate that if the theoretical criterion were taken as the true mastery standard, then mis-classification only occurred when true masters were put in the non-mastery category, implying that a false-negative type of error was seen as less serious than passing a non-master into mastery status.

Cut-Scores Based on Acceptable Risks of Mis-classifications

In applying Stig Fhaner's method of incorporating acceptable risk levels in the setting of cut-off scores, the normal approximation was used to compute the cut-scores which would result in α<.01 and β<.10. Given that the length of the test is fixed at 87 items and the α error must be very small, then the cut-off score becomes a function of the value one uses for ability which is definitely sufficient for success or definitely insufficient for success. If one were to use .75 and .60 respectively for these values, then:

$$\frac{x_1+.5 - 87(.75)}{(87(.75)(.25))} = -1.281 \quad \rightarrow x_1 = 59.58$$

$$\frac{x_2-.5-87(.60)}{87(.60)(.40)} = 2.33 \quad \rightarrow x_2 = 63.34$$

Since there is a discrepancy in the cut-off scores (x_1, x_2), then the only solution is either to increase the number of test items or relax the α risk level. If the α level is relaxed to .05, then 1.645 is substituted for 2.33 and x_2 is computed to be 60.21. This would result in a cut-off score of 61 which corresponds to being able to answer over 70 percent of the items correctly. A cross-tabulation cable of the theoretical criterion of .75 by this risk-incorporated cut-off score is shown in Table 4.

TABLE 4

Cross Tabulations of Mastery/non-Mastery by Tolerable Risk Placement (α=.05, β=.10) versus Theoretical Mastery Ability .75

		Theoretical Criterion=75 Percent Items Correct c=66 Mastery	non-Mastery	
risk incorporated cut-score c=61	Mastery	38	8	46
	non-Mastery	0	53	53
		38	61	99

By this standard then, the number of false masters is increased over the norm-referenced procedures and the number of false non-masters goes to zero. However, if α is set to .01 and β is allowed to go to .25, then the cut-off score would become 63 (see Table 5).

TABLE 5

Cross Tabulations of Mastery/non-Mastery by Tolerable Risk Placement (α=.01, β=.25) versus Theoretical Mastery Ability .75

		Theoretical Criterion=75 Percent Items Correct c=66 Mastery	non-Mastery	
c=63	Mastery	38	3	41
	non-Mastery	0	58	58
		38	61	99

Since most of the students in the engineering sample were exempted from English courses or only had to take one remedial course, the distribution of ability is probably skewed. As a result, there is still a greater number of false masters than false non-masters. It is clear, however, that the types of mis-classification increase or decrease according to how the risk levels are set.

Huynh's Optimal Decision Rule Model

An application of Huynh's model (1976b) was applied to the data using the approximation formula which assumes a constant loss ratio and a 0-1 referral success. The α, β parameters of the beta distribution were estimated to be (Huynh, 1976a):

$$\hat{\alpha} = \left(-1+\frac{1}{\hat{\alpha}_{21}}\right)\hat{\mu} = 7.25$$

$$\hat{\beta} = -\alpha + \frac{n}{\hat{\alpha}_{21}} - n = 3.29$$

Where,

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left[1-\frac{\hat{\mu}(n-\hat{\mu})}{n\sigma^2}\right]$$

$$\hat{\mu} = 59.84$$

$$\hat{\sigma}^2 = 157.50$$

When t_0, true mastery, is assumed to be 75 percent correct, and the loss ratio is one, then, the cut-score with Huynh's model is 65.66. A comparison of classifications using the theoretical criterion and the cut-off score derived from Huynh's model is shown in Table 6.

TABLE 6

Theoretical Criterion

c=66

		Master	non-Master	
Huynh's Model c=66	Master	38	0	38
	non-Master	0	61	61
		38	61	99

In this situation, where the probability of false positive and false negative mis-classification is assumed equal, no errors of classification are observed. If, however, classifying a failure as a success is twice as serious as a false non-master, a cut-off score of 67.48 or 68 is found. The cross-tabulation would then be the same as Table 2 where the norm-referenced cut-off score is used.

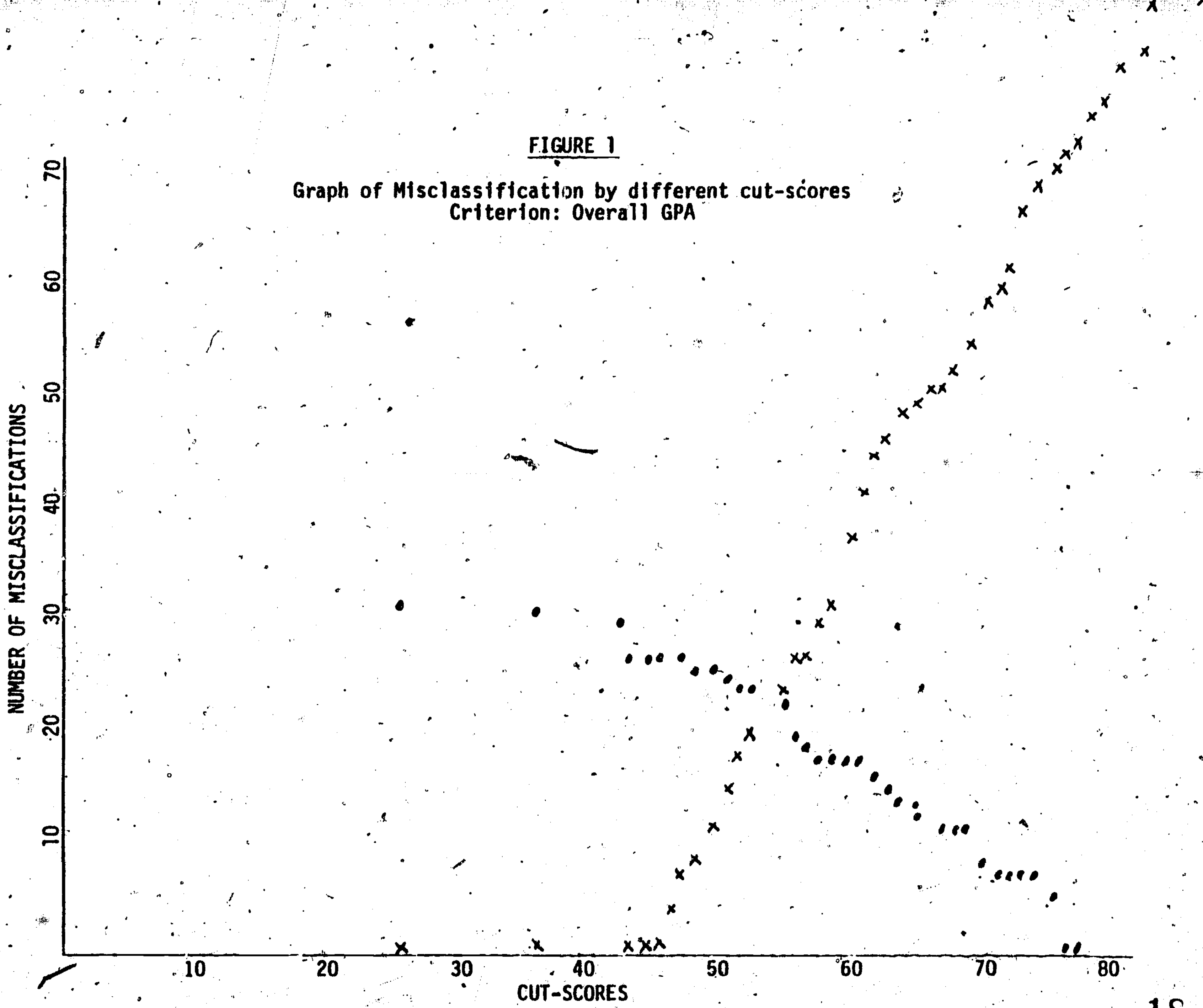
<u>Wilcox's Optimal Cut-Off Score Based on Observed Scores and an External Criterion</u>

Wilcox (1979) proposed a procedure that simply classifies examinees into masters and non-masters on the basis of some external criterion and then finding the test cut-off score which minimizes the number of mis-classifications. For example, if GPA were taken as the external criterion, the classification of masters/non-masters would depend on the GPA needed to remain in good standing as a graduate student, namely a 3.25 or above. Plotting the various cut-off score possibilities along the X-axis and the number of classification errors on the Y-axis, a graph such as the one in Figure 1 is obtained. The minimum number of mis-classifications occurs at a cut-off score of 43. This same score is obtained when a similar graph is drawn using the number of failed courses as the external criterion, and a non-master is defined as one who fails more than one course in the first year of graduate study.

Figure 1 shows that the optimal cut-off score is considerably lower than the cut-off scores of the other illustrated methods, probably indicating that the proficiency test is best for determining the <u>minimum</u> language standard needed for successful academic performance, whereas the high cut-off scores of the other methods are more concerned with a standard at which one is reasonably sure of successful performance. In fact, this interpretation is fairly consistent with UCLA's remedial English placement practices for foreign students. The upper cut-off score of 68 is associated with exempting students from all ESL course requirements, and a score of about 30 is associated with the heaviest ESL course requirements while still allowing enrollment in regular university classes.

FIGURE 1

Graph of Misclassification by different cut-scores
Criterion: Overall GPA



Wilcox's Method for Approximating True Score Distribution

Methods proposed to approximate true score distributions can also be used to examine the problems of setting cut-off scores. Let θ be the percent correct true score of an examinee, x be an observed score having as possible values 0, 1, 2, ... n, where n is the number of dichotomously scored items on a test, and $f(x|\theta)$ be the conditional probability density function of true scores over a population of examinees. Keats and Lord (1962) proposed a strong true-score model based on the assumption that $f(x|\theta)$ is the binomial probability function

(3) $$\binom{n}{x}\theta^{x}(1-\theta)^{n-x}$$

It is further assumed that the distribution of θ over the population of examinees is given by

(4) $$g(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\ \theta^{r-1}(1-\theta)^{s-1}$$

where Γ is the usual gamma function and where r and s are unknown parameters that can be estimated via the examinees' observed test scores. This is the family of beta distributions that is typically used in conjunction with (3).

Wilcox (1979) suggests replacing (4) with a more general family of distributions given by

(5) $$g(\theta) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!}\ \frac{\Gamma(r+j+s)}{\Gamma(r+j)\Gamma(s)}\ \theta^{r+j-1}(1-\theta)^{s-1}$$

where λ, r and s are unknown parameters that are estimated using observed test scores. This is the family of non-central beta distributions which contains the family of beta distributions (λ=0) as a special case.

The motivation for (5) is that we obtain a better approximation to $g(\theta)$ which in turn can have an effect on the choice of a passing score.

Using Wilcox's method, we need only the first three moments of the true score distribution in order to approximate λ, r and s. The number of examinees receiving an observed score of x on the 87 item ESL test is presented in Table 7.

TABLE 7

Frequency Distribution of Total Scores on the ESL Test

N=99

Total Test Score	Frequency	Total Test Score	Frequency
17	1	61	3
25	1	62	2
35	1	63	2
42	2	65	1
43	1	66	2
44	2	67	5
45	3	68	2
46	2	69	2
47	2	70	4
48	4	71	2
49	3	72	3
50	3	73	4
41	3	74	1
53	5	75	3
54	1	76	1
55	3	77	3
56	2	78	1
57	5	80	3
58	3	81	1
59	4	84	1
60	2		

The first three moments of the true score distribution were estimated to be .688, .491 and .362 respectively.

Setting $\lambda=0$ and using the method of moments, we estimate r and s with

$$\hat{r} = \frac{(\hat{\mu}_1)^2(1-\hat{\mu}_1)}{\hat{\mu}_2-\hat{\mu}_1} - \hat{\mu}_1$$

$$\hat{s} = \frac{\hat{\mu}_1(1-\hat{\mu}_1)^2}{\hat{\mu}_2-\hat{\mu}_1} + \hat{\mu}_1 - 1$$

(e.g., Huynh, 1976; Wilcox, 1977) yield $\hat{r}=7.784$ and $\hat{s}=3.533$. From standard results on the beta distribution, these values of r, s and λ imply that $\mu_1=.688$, $\mu_2=.499$, $\mu_3=.360$. In order to find the best estimates of μ_1, μ_2, μ_3, different λ values are estimated and presented in the following table (Table 8):

TABLE 8

Estimated Values of the First Three Moments

	Using AUG15EQ		Using AUG15μμ		
	r	s	μ_1	μ_2	μ_3
$\lambda=0$	7.7838	3.5331	.6878	.49051	.36038
$\lambda=.3$	6.9031	3.2895	.68780	.49227	.36308
$\lambda=.4$	6.9196	3.3068	.6878	.49188	.36303
$\lambda=.5$	7.1440	3.4109	.68781	.49179	.36173
$\lambda=1.0$	6.4792	3.3560	.6878	.49207	.36335
$\lambda=2.0$	6.3069	3.6992	.6878	.49101	.36124
$\lambda=3.0$	5.4857	3.7448	.6878	.49144	.36199

Notice that for λ equals 3 and solved for $\hat{r}$ and $\hat{s}$ yielding r= 5.4857 and s=3.7448. These values of r, s and λ imply that $\mu_1=.6878$, $\mu_2=.4914$ and $\mu_3=.3620$. Thus, these values of r, s, and λ are in reasonably good agreement with the estimated values of μ_1, μ_2 and μ_3. Assuming these approximations to the true score distribution $g(\theta)$, the probability of committing a false-positive (A) and false-negative error (B) can thus be estimated using:

$$A = \sum_{x=x_0}^{n} \sum_{j=p}^{100} \frac{e^{\lambda}\lambda^{j}}{j!} \binom{n}{x} \frac{r(r+s+j)}{j} \int_{0}^{\zeta_0} \zeta^{x+r+j-1} (1-\zeta)^{n-x+s-1}$$

$$B = \sum_{x=0}^{x_0-1} \sum_{j=0}^{p} \frac{e^{\lambda}\lambda^{j}}{j!} \binom{n}{x} \frac{r(r+s+j)}{j} \int_{\zeta_0}^{1.0} \zeta^{x+r+j-1} (1-\zeta)^{n-x+s-1}$$

When the cut score is set at 66 on this 87 item ESL Test, the probabilities of committing a false positive and false negative error are .010 and .152 respectively. When the cut-off score is set at 65, the probabilities are .015 and .126. Therefore, the total probability of mis-classification is less than when the cut-off score is 66. Using 42 and 43 as the cut-off scores, as computed based on Wilcox's method of choosing an optimal passing score with an external criterion, the probability of Type A error is .408 and .397 respectively and Type B errors become minimal, .202E-6 and .548E-6.

Discussion and Recommendations

Since the purpose of the ESL test is to identify students who lack the language skill required to go through graduate school successfully, it appears that a number of other factors are also needed to be considered in selecting a cut-off score. The first factor--which has been the major consideration for all illustrated methods--is the loss associated with mis-classification. Millman (1973) stated that although there are multiple methods for setting cut-off scores, none of them eliminates the element of judgment that occurs at some stage of their execution; this statement is still true. Recent developments on the topic of standard setting, however, enable us to make more informed decisions. How much risk are we willing to take? (Very little? Ten percent? Fifty percent?) What type of risk are we more willing to take? (Promote students who have attained proficiency?) Depending on the levels of risk one is willing to take, a different cut-off score can be chosen accordingly.

Another factor of concern is the predictive and construct validity of the test content with respect to the chosen external criteria. The intercorrelation between the ESL test score and overall GPA is .22 (Table 9), and it is slightly more positively correlated with the first year's GPA. This finding is expected since, after an initial stage, students all acquire a certain level of proficiency in English. The overall GPA, as well as first year GPA, shows the highest correlation with scores on the Advanced Graduate Record Examination, which is an achievement test.

TABLE 9

Correlation Matrix of ESL Data

	1 Overall GPA	2 Year 1 GPA	3 GRE Verbal	4 GRE Quantitative	5 GRE Advanced	6 ESL
1	1.00					
2	.98	1.00				
3	.18	.20	1.00			
4	.38	.34	.20	1.00		
5	.57	.55	.23	.51	1.00	
6	.22	.25	.33	.57	.27	1.00

The multiple correlation coefficient of scores on the advanced GRE and ESL with overall GPA was .56 (R^2=.31). The relatively low correlation between the ESL test and performance in graduate study may indicate that for Engineering majors, the skills tested by the ESL test have a low impact on achievement. Therefore, a lower cut-off score, such as 42 or 43, may serve screening purposes adequately. By studying the relationships between English competency and performance in subject areas for various fields of study, e.g., the humanities, sciences, social sciences, we may decide that different cut-off scores are needed to insure a given level of risk. The problem then becomes one of gathering the appropriate data to obtain estimates for the parameters used in the various cut-off score models. No matter how sophisticated these models may be in describing such things as a true score distribution, the decision makers must still take into account substantive issues unique to their own applications of the models.

References

Birk, R. A. Determination of optimal cutting scores in criterion-referenced measures. Journal of Experimental Education, 1976, 45, 4-9.

Fhaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 24, 172-75.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13(4), 253-64.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41(1).

Keats, J. A. and Lord, F. A. A theoretical distribution for mental test scores. Psychometrika, 1962, 27(1), 59-72.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43.

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

Tyler, R. W. and Sheldon, W. Education objectives and educational testing: Problems now faced, testing, teaching and learning. Report of a Conference on Research on Testing, National Institute of Education, Washington, D.C., 1979.

Wilcox, R. A lower board to the probability of choosing the optimal passing score for a mastery test where there is an external criterion. Psychometrika, 44(2), 1979.

Wilcox, R. Toward better approximation of the true score distribution with extensions to the Dirichlet-multinomial model. CSE Report, 1979.