ED 236 231 TM 830 753

AUTHOR	Ziomek, Robert L.; Szymczuk, Mike
TITLE	A Comparison of Approaches for Setting Proficiency Standards via Monte Carlo Simulations. Research Evaluation Development Technical Report Series, No. 2.
INSTITUTION	Des Moines Public Schools, IA. Dept. of Evaluation and Research.
PUB DATE	83
NOTE	26p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Classification; Computer Simulation; *Error of Measurement; Evaluation Methods; *Minimum Competency Testing; School Districts; *Scores; Secondary Education; *Standards
IDENTIFIERS	*Contrasting Groups Method; Monte Carlo Methods; Repeated Measures Design; *Standard Setting

ABSTRACT

In order to evaluate standard setting procedures, apart from the more commonly applied approach of simply comparing the derived standards or failure rates across various techniques, this study investigated the errors of classification associated with the contrasting-groups procedures. Monte Carlo simulations were employed to produce masters/nonmasters score distributions sampled from normal and left-skewed parent score distribution populations. In addition, three levels of score distribution overlap (noise) between the master/nonmaster subpopulations were simulated to examine the effects of this phenomenon on errors of classification. (Author)

Acknowledgements

We would like to thank the following individuals for their contribution to our manuscript: first, Dr. Thomas Dubloz, Iowa State University, for providing the curve simulation computer program and suggesting several Monte Carlo sample strategies; next, Drs. Frederick Brown, Iowa State University, and Larry Hedges, University of Chicago, for reviewing and giving constructive criticism; and finally , special thanks to our support staff Jean Allen, Des Moines School District word-processor operator, Shirley Jordan, Heartland secretary and Sheila Stewart, Heartland word-processor operator.

Abstract

In order to evaluate standard setting procedures, apart from the more commonly applied approachs of simply comparing the derived standards or failure rates across various techniques, this study investigated the errors of classification associated with the contrasting groups procedures. Monte Carlo simulations were employed to produce masters/nonmasters score distributions sampled from normal and left-skewed parent score distribution populations. In addition, three levels of score distribution overlap (noise) between the master/nonmaster subpopulations were simulated to examine the effects of this phenomenon on errors of classification.

A Comparison of Approaches for Setting Proficiency Standards via Monte Carlo Simulations

School districts, and a variety of other agencies, faced with the responsibility of establishing testing standards leading to the identification of acceptable levels of proficiency are faced not only with the dilemma of deciding which standard setting procedure to employ but are also confronted with the issue of subjectivity often associated with the believability of the results generated by the chosen technique(s). Current practices associated with the setting of standards in educational testing can be broadly placed into one of three categories: (1) comparisons with the performance of others, i.e., the normative approach; (2) considerations of the consequences of misclassification, such as the borderline or contrasting groups techniques; (3) examination of item content, such as the Nedelsky (1954), Ebel (1972), or Angoff (1971) procedures.

Investigations of the variety of standard setting techniques have been limited to comparisons of generated passing scores and/or the number of individuals failing for the procedures studied. Andrew and Hecht (1976), and Schoon, et al. (1970) compared the standards generated by the Nedelsky and Ebel techniques; Skakun and Kling (1980) investigated the comparability of passing scores derived from the Nedelsky, Ebel, and a modification of the Ebel procedure; Poggio, et al. (1981) concentrated on the Angoff, Ebel, Nedelsky and contrasting groups procedures; Koffler (1980) compared the obtained standards from the Nedelsky and the contrasting groups technique;

Saunders, et al. (1980) studied the scores generated by two versions of the Nedelsky approach while Brennan and Lockwood (1979) investigated the variability of passing scores generated by the Angoff and Nedelsky procedures. The one notable outcome of these investigations is that different approaches for establishing a standard produce different standards.

Although the studies noted were conducted under a variety of conditions, the conclusions are restricted to one-time comparisons among the populations and approaches employed. In addition, none of the previously noted studies investigated the errors of classification associated with the derived standards or the stability of the estimates both within and across varying techniques. Furthermore, all of these studies principally focused upon the class of standard setting procedures related to the examination of item content and the probabilities associated with passing a given item, procedures which school district personnel are less accustomed to as compared to judgments made about a student's level of overall performance on a test.

This study employs Monte Carlo simulations to examine the properties of standards derived from the contrasting groups technique. In addition, these standards are compared, on the basis of errors of classification and stability, to the estimates of standards derived from three preselected procedures based upon theory and empirical evidence. Stability, for purposes of this investigation, was studied by simulating pairs of masters and nonmasters subpopulations, randomly generated from a normal and negatively skewed parent population, respectively. The standard associated with the minimum number of misclassifications from the first member of the pair was used as the standard for the second paired simulation, and the corresponding errors of classification tabulated. In addition, three predetermined levels of noise (degree of sample distribution overlap between the masters and nonmasters subpopulations) were simulated in order to study this phenomenon's effects on the stability of errors of classification.

Background

The decision by a school district to employ the contrasting groups procedure as a standard setting technique for a competency testing program, is reasonably based upon two considerations: (1) teachers are more accustomed to judging the overall adequacy of student achievement than to guessing the probabilities of a student's success on a given item; and (2) the contrasting groups method provides a direct assessment of errors of classification associated with a given score (Zieky and Livingston, 1977). As noted by Zieky and Livingston in their manual, Methods for Setting Standard on Criterion-Referenced Tests of Basic Skills, "the idea behind the contrasting groups method is to set a standard at the test score level that best separates the students judged to be masters from the students judged to be nonmasters on the objectives measured by the test."

A sample of teachers, serving as judges, are instructed to identify several students in their classes whom they are certain are either definite masters or nonmasters of the skills measured by the test on which a passing score is being set. Once this process is completed, the test is administered to the population of examinees and the scores for the previously identified groups of masters and nonmasters are examined to determine the standard minimizing the number of errors of classification. Two types of error are associated with this procedure: (1) classifying as master a student who has not adequately mastered the objectives (false master, Type I error); (2) classifying as a nonmaster a student who has adequately mastered the objectives (false nonmaster, Type II error). Raising the standard reduces the number of Type I errors while increasing the number of Type II errors. Lowering the standard produces the opposite results.

The contrasting groups method employed by the school district providing the empirical test data for this study, was utilized over a three year period to set standards for reading and mathematics competency tests. Two salient trends became apparent over this period. First, for both the population of students tested and the subgroups defining the masters and nonmasters, the distribution of scores for the reading competency test exhibited significant negative skewness, while the distribution of scores on the mathematics tests approximated a normal distribution.[1] Second, the degree of overlap between the groups of student masters and nonmasters was consistently greater for the reading than the mathematics competency test. These empirical conditions provided the framework within which the Monte Carlo simulation was pursued.

Monte Carlo Simulation

The Ahrens and Dieter algorithm for beta parameters (Ahrens and Dieter, 1974) was used to simulate normal and negatively skewed population distributions with a raw score range of 1 to 100. For the normal distribution, α and β were set at 10, resulting in a population mean of 49.59 and a standard deviation of 6.82. To generate the negatively skewed distribution, α was set at 10 and β was set at 2 representing a highly negatively skewed distribution modeling the empirical data for the reading tests. This distribution had a mean of 87.02 and a standard deviation of 6.16. Each distribution consisted of 2450 nonzero values representing the average number of students within the school district taking either the mathematics or reading competency test.

A Statistical Analysis System (SAS) program was written to generate samples of masters and nonmasters subpopulations from each simulated parent population. SAS's uniform distribution function was used to randomly sample scores from the tails (greater than ± one standard deviation from the population mean) and middle range (within ± one standard deviation from the mean) of the two populations. Paralleling the recommendations of Zieky and Livingston (1977), the total number of scores comprising the masters and nonmasters samples was maintained at greater than 100 observations per group, respectively.

Sampling from the middle portion of each parent population represented the masters/nonmasters score distributions overlap, noise. Table 1 presents the proportions used to sample from the high and low score tails of each parent population, as well as the proportion and range of the number of cases falling within the overlap region associated with the three noise levels. As an example, refer to Table 1, normal distribution, low noise. Twenty percent of the scores, one standard deviation above and below the mean, were randomly sampled from the overall population of scores and allocated to the masters and nonmasters groups respectively. Of the scores lying within plus-or-minus one standard deviation from the mean, two and one-half percent were sampled and randomly assigned to either the masters or nonmasters group. Varying the percentage sampled from the middle portion of the parent population served to define the three noise levels.

Table 1

Predetermined Proportions Used to Generate the Masters and Nonmasters Subpopulations by Degree of Sample Overlap

	Distribution						
	Normal			Left Skewed			
Noise Level	Low Tail	Middle	High Tail	Low Tail	Middle	High Tail	Noise Level
Low [41][1]	.20	.025	.20	.145	.025	.240	[29] Low
		(18-46)[2]			(21-39)		
Middle [43]	.16	.050	.16	.110	.050	.195	[30] Middle
		(49-99)			(45-73)		
High [44]	.14	.075	.14	.090	.075	.170	[30] High
		(82-124)			(67-100)		

[1] Figures in brackets represent the number of simulated pairs generated at each noise level.

[2] Figures in parenthesis reflect the range of cases falling in the areas of overlap between the master and nonmaster subpopulations.

Standard Setting Procedures

For the contrasting groups procedure the score that resulted in the minimum number of errors of classification was considered to be the optimal standard for a given sample[2]. For purposes of this study, the errors of classification were tabulated by counting the number of scores in the masters distribution which fell below the derived standard and adding this result to the number of scores from the nonmasters group which fell above the standard. This total was subsequently divided by the total number of scores within the overlap region of the two groups and designated the error rate. This was done to standardize the errors of classification across samples and noise levels for analysis purposes.

Alternative standard setting strategies included in this study and, thus, providing a basis of comparison for the contrasting groups technique, were: (1) the linear discriminant function (LDF) applied to the normal distribution, and defined as:

$$[\bar{X}_1 - \bar{X}_2)/S^2]\ [Z - (\bar{X}_1 + \bar{X}_2)/2] \qquad (1)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the masters' and nonmasters' test scores respectively, S^2 is the pooled variance, and Z the test score to be classified; (2) the quadratic discriminant function (QDF) for the left skewed distribution, defined as:

$$Z[\bar{X}_1/S_1^2-\bar{X}_2/S_2^2]-Z^2/2[1/S_1^2-1/S_2^2]-1/2[\bar{X}_1^2/S_1^2-\bar{X}_2^2/S_2^2]+1/2LN[S_2^2/S_1^2] \qquad (2)$$

where $\bar{X}_1$, S_1^2 and $\bar{X}_2$, S_2^2 are the means and variances of the masters' and nonmasters' test scores, respectively, and Z is the test score to be classified; (3) a third solution, referred to as an "empirical solution," assumed each sample of masters' and nonmasters' scores was normally distributed, regardless of the distributional form of the parent population; and (4) each parent population's respective mean.

The LDF is suggested as an appropriate technique to utilize when the populations of masters' and nonmasters' test scores are normally distributed with equal but unknown variances and means; whereas, the QDF applied to the ranks of the raw scores is a recommended approach for skewed distributions (Conover and Iman, 1978). For both the LDF and QDF procedures, the standard which minimized the probability of misclassification was the smallest score, such that either equation was greater than the constant:

$$\mathrm{LN}\ (C_{12}Q_2/C_{21}Q_1) \tag{3}$$

where C_{ij} is the cost (either monetary, psychological or a combination) of misclassifying an observation belonging to population j into population i (i,j = 1,2), and Q_i (i = 1,2) is the prior probability of group membership (Anderson, 1951). In this study C_{12} and C_{21} were assumed equal. The proportions of cases in the sample determining the masters (Q_1) and nonmasters (Q_2) were used as estimates of group membership. (See Koffler, 1980, for a discussion of the QDF and LDF procedures.)

The third standard setting procedure employed for comparison purposes, and referred to as t'e empirical solution, involved the simple equating of two normal density functions yielding Equation 2 (Lachenbruch, et al., 1973). Sample values for the means and standard deviations for the masters and nonmasters simulated distributions are substituted into the equation as noted previously. The two major differences between this procedure and the LDF/QDF strategies is that the empirical solution was used to determine a standard for the masters/nonmasters samples regardless of the distributional form of the parent populations; and, secondly, the solution did not employ Equation 3. Empirical evidence gathered from the results of the administration of the reading and mathematics minimum competency tests within the school district suggested that the test's overall mean was a reasonable approximation to the contrasting groups standard. Hence, the errors of classification associated with the population mean for each sample were tabulated and included for comparison.

Tables 2 and 3 present the descriptive results of the Monte Carlo paired simulations for the master/nonmaster samples generated from the parent normal and skewed distribution, by noise level and technique respectively.

Analysis and Findings

A repeated measures design was used to compare the various standard setting procedures (P), the effect of noise level (N), and the repeated measure (R), for each population. The dependent variable was the paired errors of classification. Tables 4a and 4b, 5a and 5b present the ANOVA results and descriptive statistics for the normal and left skewed simulations, respectively.

Table 2

Descriptive Statistics of the Monte Carlo Simulation for the Normal Population Distribution

Noise Level	Paired Simulation	Error Rates	Technique[a] I	II	III	IV
Low	First	Range	.30-.72	.30-.83	.29-.78	.30-.78
		Mean	.42	.49	.49	.51
		SD	.08	.10	.09	.10
	Second	Range	.30-.64	.27-.65	.30-.65	.27-.63
		Mean	.46	.47	.47	.47
		SD	.12	.08	.08	.08
Medium	First	Range	.29-.51	.31-.56	.31-.57	.29-.60
		Mean	.41	.44	.45	.45
		SD	.05	.06	.07	.07
	Second	Range	.36-.66	.38-.62	.36-.66	.36-.60
		Mean	.49	.49	.49	.48
		SD	.06	.06	.06	.06
High	First	Range	.18-.52	.32-.59	.36-.60	.36-.60
		Mean	.43	.48	.49	.49
		SD	.06	.05	.05	.05
	Second	Range	.36-.60	.38-.55	.38-.57	.25-.57
		Mean	.47	.46	.46	.46
		SD	.05	.05	.04	.06

[a] I equals contrasting groups; II equals LDF; III equals empirical solution; IV equals population mean.

Table 3

Descriptive Statistics of the Monte Carlo Simulation for the Left Skewed Population Distribution

			Technique[a]			
Noise Level	Paired Simulation	Error Rates	I	II	III	IV
Low	First	Range	.23-.47	.28-.56	.27-.67	.23-.67
		Mean	.36	.41	.44	.45
		SD	.06	.08	.08	.09
	Second	Range	.28-.61	.30-.62	.27-.52	.28-.58
		Mean	.45	.44	.44	.45
		SD	.08	.09	.07	.08
Medium	First	Range	.29-.48	.31-.57	.35-.63	.33-.63
		Mean	.40	.45	.48	.47
		SD	.05	.07	.07	.07
	Second	Range	.31-.85	.25-.70	.30-.85	.30-.85
		Mean	.47	.45	.47	.47
		SD	.11	.09	.11	.17
High	First	Range	.34-.53	.34-.56	.37-.61	.37-.60
		Mean	.42	.46	.50	.49
		SD	.04	.05	.05	.05
	Second	Range	.30-.59	.31-.59	.36-.60	.37-.56
		Mean	.44	.43	.45	.44
		SD	.06	.05	.06	.05

[a] I equals contrasting groups; II equals QDF; III equals empirical solution; IV equals population mean.

Table 4a

Repeated Measures ANOVA of Errors of Classification for the Normal Distribution

Source of Variation	df	ms	F
Between			
Noise (N)	2	.008	.8
Procedure (P)	3	.091	9.10*
N x P	6	.009	.90
Residual	500	.010	
Within			
Repeated Measure (R)	1	.016	7.27*
R x N	2	.116	52.73*
R x P	3	.053	24.09*
R x N x P	6	.0015	.68
Residual	1000	.0022	

* $p < .01$

Table 4b

Means and Standard Deviations of Error Rates for the Normal Distribution Repeated Measures ANOVA

Repeated Measure		Procedure Contrasting Group	Discriminant Function	Empirical Solution	Population Mean
1	Mean	.42	.48	.48	.48
	SD	.06	.08	.07	.08
2	Mean	.48	.48	.48	.47
	SD	.07	.06	.06	.07

Repeated Measure		Noise Low	Medium	High
1	Mean	.48	.44	.48
	SD	.10	.06	.06
2	Mean	.47	.49	.46
	SD	.08	.06	.05

Table 5a

Repeated Measures ANOVA of Errors of

Classification for the Left Skewed Distribution

Source of Variation	df	ms	F
Between			
Noise (N)	2	.115	9.58*
Procedure (P)	3	.134	11.17*
N x P	6	.002	.17
Residual	344	.012	
Within			
Repeated Measure (R)	1	.008	2.00
R X N	2	.065	16.25*
R x P	3	.057	14.25*
R x N x P	6	.0005	.125
Residual	688	.004	

* $p < .01$

Table 5b

Means and Standard Deviations of Error Rates for the Left Skewed Distribution Repeated Measure ANOVA

Repeated Measure		Procedure Contrasting Group	Discriminant Function	Empirical Solution	Population Mean
1	Mean	.39	.44	.47	.47
	SD	.06	.07	.08	.07
2	Mean	.45	.44	.45	.45
	SD	.09	.08	.08	.08

Repeated Measure		Noise Low	Medium	High
1	Mean	.41	.44	.47
	SD	.09	.07	.06
2	Mean	.44	.46	.44
	SD	.08	.10	.06

For the normal simulation, a significant main effect resulted for procedure (P) and repeated measure (R), while significant interactions appeared for repeated measure by noise level (R x N) and repeated measure by procedure (R x P). Referring to Table 4b, the R x P interaction occurring for the contrasting groups procedure shows an increase in the average errors of classification over repeated measures while for the three remaining procedures the average errors of classification remain quite stable. Across repeated samplings, the contrasting groups procedure does produce, relatively, a lower average error rate. The fact that the main effect, noise level, is not significant suggests that, in the case of the normal distribution, the amount of master/nonmaster overlap has little influence on errors of classification. However, the R x N interaction reveals the instability of error, especially at the medium noise level, when examined over repeated samplings.

Referring to Tables 5a and 5b, for the left skewed population simulation results, both main effects, noise and procedure, are significant, with significant interactions occurring for R x N, and R x P. The significamce of P is due to the lower average error rate, across repeated samples, for the contrasting groups technique. While the error rates for the three comparison procedures remain relatively stable, the R x P interaction is due to the instability of the errors, across repeated samplings, for the contrasting groups technique. Interestingly, although N was not significant for the normal simulations, it is significant for the left skewed population, with the lowest average error rate, across repeated measures, occurring at the low noise level. Likewise, the repeated measures factor, although significant for the normal simulations and suggesting error rate instability, is not significant for the left skewed simulations. The explanation of the significance of the R x N and R x P interactions for the left skewed simulations is due to the significance of the difference for N and P, favoring the low noise level and the contrasting groups procedure; while the R x N and R x P interactions reported for the normal distributions are due to R and P, instability of the error rates at the medium noise level and a low overall average error rate for the contrasting groups technique. Regardless of the population simulated, procedure or noise levels, all error rates are high and approaching the chance level.

Discussion

In their review of minimum competency testing and the accompanying standard setting problem, Linn, et al. (1982) concluded that "there is no good basis for judging one procedure for setting the passing score superior to another." This statement was based upon a comparison of the differences among the derived passing scores and the varying number of student failures for the standard setting procedures investigated. Our investigation approached this apparent dilemma by assuming that a reasonable method for judging the superiority of a standard setting procedure was to investigate the errors of classification associated with the techniques selected for this study. Regardless of the resulting standard, it seems apparent that the procedure with an "acceptably low" misclassification error rate would be the most appealing strategy.

All standard setting procedures require an investment in time on behalf of expert judges and other personnel. Notwithstanding the concerns plaguing school administrators about the availability of monies to support remediation programs, more consideration should be given to the "accuracy" of a given technique. Based upon the experiences of the school district participating in this study and Educational Testing Service with minimum competency tests, this study attempted to fill an information void in this area.

It is clear from the results that regardless of the standard setting procedure and/or level of overlap, the misclassification rates are extremely high and approaching the 50 percent chance level in many instances. From both a psychometric and administrative viewpoint, it is apparent that determining competency (and the eventual allocation of funds for remediation) on the basis of a one time administration of a minimum competency test is a highly risky undertaking. Thus, Linn's concern about different judges and/or different standard setting procedures producing different standards is accompanied by the potential problem of unacceptably high error rates of classification.

Based upon the results presented in the previous section, it is clear that the contrasting groups technique shows the greatest average change in error rates across repeated samplings (see Tables 4b and 5b). A brief examination of how the standards are computed for each of the techniques studied should shed some light on this phenomenon. Both the QDF and LDF procedures take into account the sample statistics and prior membership probabilities of the groups involved. These statistics remained "relatively" stable across individual members of a pair as well as across simulated pairs. Although the empirical solution did not incorporate the criterion of prior membership probabilities, ($LN(Q_1/Q_2)$ was approximately zero for this study), the sample statistics for the master/nonmasters groups were, as noted for the QDF and LDF techniques, essentially stable. The parent population mean was "fixed" as the standard across simulations, and since it was not conceived as an estimate of a standard in terms of a minimum error rate, the errors of classification were free to vary, and yet, reflect good agreement with the LDF, QDF and empirical solutions. This could be a function of the overall parent population characteristics, however, the contrasting groups procedure does not exhibit a similar agreement across repeated samplings.

The derived standard of the contrasting groups procedure, although based upon the minimum number of classification errors, does not take into account master/nonmaster sample statistics, or prior membership probabilities. Each standard is established for a given masters/nonmasters frequency distribution, shifts in the score distributions across simulated pairs can result in potentially large differences in the derived standard, as well as potentially large changes in the accompanying error rates.

As noted by Divgi (1982), "Standards are set because they have to be, in situations where it is believed (at least by those in authority) that imperfect standards are better than none. No standard can satisfy everybody. One can only ask that the standard be reasonable, and that those who set it be aware of what theya re doing and why." From a strictly psychometric viewpoint and consistent with Divgi's statement, we would argue against making decisions concerning competency, based upon a single test administration, and opt for a more carefully delineated school district testing program. Decisions regarding standard setting, competency and remediation should be based upon a combination of a student's logitudinal history of testing and classroom performance tempered by teacher input.

Reference Notes

1. Samuel Livingston, via a telephone conversation, noted similar trends in district results associated with the Basic Skills Assessment Tests developed by ETS, and employed as competency tests in reading and mathematics.

2. Initially, if several different standards resulted in an equal number of minimum errors of classification, the lowest score was labeled as a Type I estimate, more false masters; whereas, the higher score was designated as a Type II estimate, more false nonmasters. Subsequent analyses produced very similar results for these two classifications, consequently, it was decided to present only the Type I findings in the report.

References

Ahrens, J. H. & Dieter, V. Computer methods for sampling from gamma, beta, poisson, and binomial distributions. Computing, 1974, 12, 223-246.

Anderson, T. W. Classification by multivariate analysis. Psychometrika, 1951, 16, 31-50.

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

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

Brennan, R. L. Lockwood, R. E. A comparison of two cutting score procedures using generalizability theory. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

Conover, W. J. & Iman, R. L. The rank transformation as a method of discrimination with some examples. Albuquerque, New Mexico: Sandia Laboratories, 1978.

Divgi, D. R. The logic of standard setting: some issues and questions. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1982.

Ebel, R. L. Essentials of Educational Measurement, Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1972.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

Lachenbruch, P. A., Sneeringer, C. & Revo, L. T. Robustness of the linear and quadratic discriminant function to certain types of non-normality. Communications in Statistics, 1973, 1, 39-56.

Linn, R. L., Madaus, G. F. & Pedulla, J. J. Minimum competency testing: cautions on the state of the art. American Journal of Education, 1982, 91, 1-35.

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

Poggio, J. P., Glasnapp, D. R. & Eros, D. S. An empirical investigation of the Angoff, Ebel and Nedelsky standard-setting methods. Paper presented at the annual meeting of the American Education Research Association, Los Angeles April 1981.

Saunders C., Ryan, J. P. & Huynh, H. A comparison of two ways of setting passing scores based on the Nedelsky procedure. Paper presented at the Eastern Educational Research Association Conference, Norfolk, Virginia, 1980.

Schoon, C. G., Gullion, C. M. & Ferrara, P. Bayesian Statistics, Credentialing examinations, and the determination of passing points. Evaluation and the Health Professions, 1979, 2, 181-201.

Skakon, E. N. & Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

Zieky, J. L. & Livingston, S. A. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, New Jersey: Educational Testing Service, 1977.