1980

An Investigation of Two Criterion-Referencing Scoring Procedures for National Board Dental Examinations

Maribeth Hladis
Loyola University Chicago

AN INVESTIGATION OF
TWO CRITERION-REFERENCING SCORING PROCEDURES
FOR NATIONAL BOARD DENTAL EXAMINATIONS

by

Maribeth Hladis

A Thesis Submitted to the Faculty of the Graduate School
of Loyola University of Chicago in Partial Fulfillment
of the Requirements for the Degree of
Master of Arts

May

1980

ACKNOWLEDGMENTS

The author wishes to acknowledge with deep gratitude the assistance given by Dr. Jack Kavanagh and Dr. Ronald Morgan.

A sincere thank-you is also extended to Dr. James R. Little, who served as chairman to both committees, and other committee members: Dr. Harold F. Bosco, Dr. S. Henry Lampert, Dr. Helyn C. Luechauer, Dr. Ricardo R. Perez-Balzac, Dr. Richard J. Reynolds and Dr. John A. Stewart. A special thank-you is extended to the staff of the Commission on National Dental Examinations and unnamed relatives and friends for their unending patience and encouragement.

LIFE

The author, Maribeth Kathryn Hladis, is the daughter of Edward J. Hladis and Beatrice (Malcak) Hladis. She was born May 20, 1949 in Chicago, Illinois.

Her elementary education was obtained in private schools in Chicago and Westchester, Illinois. Her secondary education was obtained in Immaculate Heart of Mary High School, Westchester, Illinois, where she graduated with honors in 1967. In October, 1967, she entered the College of Saint Teresa, Winona, Minnesota, and in June, 1971, she received the degree of Bachelor of Arts with majors in English and psychology.

In October, 1971, she was employed by the Council on Dental Education of the American Dental Association. In January, 1975, she was promoted and transferred to the then Council on National Board Examinations of the same association. Her current position is assistant secretary of the Commission on National Dental Examinations.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF ILLUSTRATIONS
(or Figures)

CONTENTS OF APPENDICES

CHAPTER I

INTRODUCTION

Experts in testing methodology concur that, whether an examination is intended to measure a person's ability to learn a new principle or task or mastery of the principle or task, of most importance are the characteristics of the testing instrument. To be reliable and valid, an examination must measure what it purports to measure and produce results that are a consistent and fair assessment of the examinee's ability or achievement. Throughout the years, nationwide testing agencies have achieved wide acclaim for developing and conducting reliable and valid examinations.

Most nationwide testing agencies use norm-referenced scoring procedures to report performance to examinees. Because of the large numbers participating in each administration of a nationwide examination, test results produce near-normal distributions of raw scores. Conversion to a standard scoring system is accomplished using performance of a predetermined norming group as a base for assigning scores. Equivalency among norming groups is monitored to insure consistency in interpretation of scores from edition to edition of an examination.

In recent years, criterion-referenced scoring procedures have been proposed as a better mechanism for evaluating performance on examinations that test for entry into an occupation or a profession. Unlike norm-referencing, criterion-referencing is distribution-free. Using this method of evaluation, an individual's ability or achievement is assessed by comparing his performance on an examination to criteria established by experts in advance of administration of the examination. The degree to which an individual meets these criteria determines his score on the examination.

Most recently, federal and state legislators have adopted the principles of criterion-referencing certification and licensure examinations. For example, proposed federal legislation (HR3564), which has come to be known as a "truth in testing" bill, requires testing agencies that develop and conduct examinations for entry into an occupation or a profession to evaluate performance and report scores without regard to the distribution of scores produced by the population of examinees. If this legislation is passed, nationwide testing agencies will be forced to begin criterion-referencing examinations.

One testing agency that would be regulated if proposed legislation is enacted is the Commission on National Dental Examinations which is responsible for the development and conduct of National Board dental examinations. National Board dental examinations are recognized in 51 of 53 licensing jurisdictions as fulfilling or partially fulfilling the written examination requirement for dental licensure. The examination battery consists of 11 written examinations that test knowledge of basic biomedical and clinical sciences required for the competent practice of dentistry. Examinations are composed solely of multiple-choice test items and are scored using a system of standard scores and a defined norming group for each new edition of an examination.

Statement of the Problem and Rationale

In anticipating a possible change in scoring procedures, the Commission on National Dental Examinations, like other testing agencies, is faced with identifying a method of criterion-referencing that will ensure consistency in meaning of scores through the period of transition and following implementation of a new scoring system. While due emphasis has been placed on the different procedures by which an examination can be criterion-referenced, comparison of the various methods has been subject to little study. Do different criterion-referencing methods based on similar assumptions elicit the same scoring standards when applied to the same examination? Do different groups of experts using the same criterion-referencing method on the same examination set similar standards for scoring? If different criterion-referencing methods or the judgments of various experts elicit dissimilar scoring standards, can the bases for the differences be determined by studying the components of the methods used? Could not item analysis statistics generated from actual administration of an examination be used in investigating the bases for differences in methods or judgments?

Before criterion-referencing measures can be adopted for use in scoring standardized examinations, further study seems needed to determine the consistency with which different criterion-referencing methods elicit similar standards for scoring. By definition, criterion-referencing involves a source of variation and, therefore, potential error that is not common to norm-referencing procedures. Criterion-referencing procedures require decisions of experts who are assigned the task of establishing criteria. No one method of criterion-referencing has been universally accepted.

Purpose of the Study

The purpose of this study is to investigate the equivalence, stability and other characteristics of two criterion-referencing methods applied to National Board dental examinations by (1) comparing the scoring standards established by two committees of experts using the same criterion-referencing method on the same examination, (2) comparing the scoring standards established by the same committee of experts using different criterion-referencing methods on the same examination, (3) evaluating the stability of each criterion-referencing method, (4) determining the relationships between measurement components common to both criterion-referencing methods and (5) comparing scoring standards established using criterion-referencing methods with performance data collected following administration of the examinations to candidates for dental licensure.

CHAPTER II

REVIEW OF RELATED LITERATURE

Popham and Husek (1969) suggest that Glaser's discussion of instruction and measurement of learning (1963) catalyzed interest among test and measurement experts in evaluation procedures. In his article, Glaser defines two constructs in measurement practice, referring to them as norm-referenced and criterion-referenced approaches to measurement. Since the appearance of Glaser's article, measurement specialists have been drawing distinctions between the two approaches and arguing the advantages and disadvantages of each. Peculiarly enough, Ebel (1971) states that a study of the history of evaluation practices in schools reveals that the trend in educational measurement is one of predictable change. Ebel suggests that the outdated practice of assigning per cent grades is, in fact, one method of criterion-referencing. Seventy-five per cent identified the criterion that a passing student had to attain or surpass. Ebel continues to explain that when per cent scores fell into disfavor among educators, per cents were replaced with converted scores ranging from A assigned to superior performance to F assigned to failing performance. He states that the current trend is one of predictable change--a return to a criterion-referenced approach to evaluation.

Norm-Referencing versus Criterion-Referencing

Ebel (1971) states that the essential difference between norm-referenced and criterion-referenced measures is in the quantitative scale used to express how much an examinee can do. He depicts a scale of norm-referencing as being anchored in the middle on the average level of performance for a group. The units on the scale are functions of the distribution of the group. In a criterion-referenced approach, the scale is anchored at the extremities. Performance at the top indicates complete mastery while performance at the bottom indicates complete lack of abilities. The units on the scale are, then, subdivisions of the total range of the scale.

A review of the literature published since Glaser's discussion of norm-referencing and criterion-referencing illustrates the disparity of beliefs that exists concerning the better approach to evaluation. In his discussion of measurement, Gardner (1962) proposes characteristics of an ideal examination. Among these characteristics are that test items constitute a representative sample of the domain to be tested, that a frame of reference for interpreting scores be provided and that items be such that a score of zero indicates complete lack of ability. While norm-referenced measures are indicative of level of performance on a representative sampling of the test domain, criterion-referenced measures, by definition of complete mastery, dictate that more than a sample of the domain be included in the examination. A frame of reference is provided in both norm-referenced and criterion-referenced measures; however, the nature of the frame of reference differs substantially. With respect to Gardner's third criterion, a score of zero on a criterion-referenced examination indicates lack of ability because the entire domain of behavior is being tested. A score of zero on a norm-referenced examination cannot be interpreted as complete absence of ability if the examination consists of only a sampling of the domain.

Both Lindquist (1953) and Angoff (1962) concur that the best type of measurement scale is one that is divorced as much as possible from any defined norm. With this approach, if norms change, measurement is not contaminated. Cronbach (1971), in his discussion of test validation, also implies support for criterion-referenced measures by stating that the aim of testing is to predict a criterion and the merit of a test is judged by the accuracy with which it predicts, irrespective of the performance of others. Block (1971), also favors absolute measures that are interpretable solely on the basis of predetermined performance standards.

In defense of norm-referencing procedures, Thorndike (1971) distinguishes between using test scores to make an absolute decision and using test scores to indicate relative performance levels. He states that criterion-referencing is appropriate in programmed instruction because the question asked relates only to a specific individual and the materials of instruction. Standardized tests differ in that results should not only reflect an evaluation of an individual's competence, but the evaluation should place the individual in relation to his peers. Millman (1970) identifies key difficulties with criterion-referenced measurement that center around specifying the universe of tasks to be tested and determining proficiency standards on which to base evaluation. Stake (1971) and Hieronymus (1972) recognize that norm-referenced measures are not pure in predicting specific behavior, but believe that criterion-referenced measures are unable to serve as predictors of either specific or general behavior.

Perhaps a clue to solving the issue of which method of measurement is superior is suggested in Klein's discussion of evaluating tests (1970). Klein proposes that norms should be derived for examinations labeled criterion-referenced so that the two evaluation methods could be combined to interpret test results.

Criterion-Referencing Methods

In contrast to norm-referencing which has come to describe a specific evaluation procedure, many procedures have been developed for criterion-referencing. Meskauskas (1976) states that criterion-referencing models are alike in that they require tight specification of content areas; however, the models differ in how they define mastery and, therefore, in how they perform.

Nedelsky: In the late 1940's, Nedelsky (1954) developed an approach to determine an absolute standard for passing scores. He reasons that on a multiple choice examination where each test item has a single correct response, a minimum passing score can be determined by calculating the probability that a minimally competent examinee will answer each item correctly. The procedure he proposes involves having experts determine, in advance of administration of the examination, the choices in each test item that the lowest "D student" should be able to reject as incorrect. The probability that a minimally competent examinee will choose a correct response is the reciprocal of the number of remaining responses. For example, a test item with five choices where two of the choices have been eliminated would be assigned a probability of 1/3 that a minimally competent examinee would respond correctly. The minimum passing score is the sum of all reciprocals.

Nedelsky proposes adding a constant (K) term multiplied by the standard deviation to the average of minimum passing levels of various judges to adjust the distribution of probabilities. Meskauskas (1976) relates that because the constant term seems unjustified, K should always be assigned a value of zero and, therefore, the term can be eliminated from calculations.

Nedelsky's work is significant in that his model is one of few that forces experts to assess individual test items when establishing criteria for acceptable performance. His method requires evaluation of the difficulty level of each test item, while assuming the content of all test items to be of relevance to the ability being tested.

Ebel: Ebel (1972) developed a method for arriving at a minimum passing score by considering the characteristics of test items along two dimensions--relevance and difficulty. He suggests four categories of relevance (essential, important, acceptable, questionable) and three categories of difficulty (easy, medium, hard) that form a 4 x 3 matrix. Experts assign test items to the cells of the matrix that describe the relevance and difficulty levels estimated for a minimally competent examinee. Once all items are classified, judges are asked to assign a percentage to each cell that defines how many test items a minimally qualified examinee should be able to answer correctly. The number of questions in each cell is multiplied by the percentage assigned to that cell. The minimum passing score is the sum of all cross-products.

Ebel's model, like that of Nedelsky, requires that decisions be based on assessment of individual test items. Unlike Nedelsky, Ebel believes that relevance of item content is a significant factor in setting scoring standards. In Ebel's approach, the possibility of answering correctly based on a lucky guess is not accounted for. If judges determine that zero per cent of the items categorized as questionable and difficult would be answered correctly by borderline examinees, the minimum pass level for that cell would equal zero. Yet, it is reasonable to assume that examinees will guess correctly on a certain per cent of these items. In this respect, Nedelsky's model based on probabilities is conceptually more attractive.

University of Illinois: Based on Angoff's model (in Thorndike, 1971), educational psychologists at the University of Illinois (1973) developed yet another method of assessing individual test items to determine minimum passing levels for multiple-choice examinations. This procedure involves using a scale of 0-2 to weigh possible responses of each test item in terms of the likelihood that a minimally competent examinee will choose each response. Judges are asked to assign the correct response a value of 2. All other choices are assigned a value of 0 if a minimally competent examinee is expected to reject the option, 2 if such examinee would find the option exceptionally attractive, and 1 if such examinee may or may not select the choice as being correct. A minimum passing index is calculated for each test item by summing the difficulty weightings and dividing the sum into 2. The minimum passing level is the sum of the minimum passing indices for all test items.

Like Nedelsky's method, this model is based on determining the probability of success for a minimally qualified examinee. Again, relevance of test items is not considered in establishing scoring standards.

The preceding methods of setting standards focus on decisions about individual test items. Other approaches have been developed that are pure mathematically-based techniques. An assumption underlying these methods is that all items in a test are of equal difficulty. Any deviation from this level of difficulty is attributed to random selection of incorrect responses. These models also assume a standard of performance and then evaluate errors of classification into mastery or non-mastery performance to adjust the standard. Because the models are unattractive in their underlying assumption, only limited discussion seems necessary.

Kriewall: Kriewall's model (1972) focuses on categorization of examinees into three groups: master, non-masters and those in-between these extremes. While he suggests three categories of behavior, his model is founded on the assumption that only masters and non-masters exist. Because the likelihood that an individual will select a correct response is fixed across all items, the probability of success on a test item can be calculated using a binomial-based model.

Emrick: Emrick (1971) proposes a mastery testing evaluation model that assumes that the examination is testing a homogeneous content area and that all test items are clustered around the content area. The formula Emrick proposes requires determination of the probability of committing Type I and Type II errors in classifying examinees as masters or non-masters, determination of test length and calculation of a Ratio of Regret. The Ratio of Regret is obtained by evaluating classification errors and noting summed risks. With his formula, the highest percentage of items that should be attained by an individual performing at mastery level is obtained.

Meskauskas (1976) also discusses models developed by Millman (1972, 1973), Novick (1974) and Davis and Diamond (1974). Like those proposed by Kriewall and Emrick, these models are mathematically-based and view mastery as an all-or-none description of an individual's ability with respect to a specific content domain. Because these methods assume that setting scoring standards should ideally be error-free, the focus of study is on determining measurement error and accounting for potential error mathematically.

Comparisons of Pairs of Criterion-Referencing Methods

Andrew and Hecht (1976) investigated two criterion-referencing procedures for establishing scoring standards to determine (1) whether procedures based on similar assumptions would result in similar examination standards and (2) whether different panels of judges set similar examination standards when using the same criterion-referencing procedures on the same examination content. A 180-item examination was divided into equal halves. Two groups of four judges each were asked to determine minimum passing scores using procedures developed by Nedelsky (1954) and Ebel (1972). Both groups applied the same criterion-referencing method to the same half of the examination. Results of the study indicate a significant difference between methods, but no significant difference between committee decisions and no significant interaction effect.

Replication of the study using the same procedures with two different groups and a different examination produced results that indicate significant differences between methods and panels of experts and no significant interaction effect. Andrew and Hecht conclude that applying the Ebel and the Nedelsky models yields significantly different overall examination standards for equivalent samples of test content and that different panels of judges using the same procedure on the same examination content do not necessarily set similar overall examination standards.

Brennan and Lockwood (1979) applied a generalizability theory in an attempt to quantify the variance in minimum passing scores resulting from two different cutting score procedures. In their study, each of five raters used the methods developed by Angoff and Nedelsky to establish minimum pass levels for a 126-item examination. Results indicate that both the cutting scores and the expected variance in scores among raters were considerably different for the two procedures. Brennan and Lockwood postulate that the differences could be explained by (1) differences in the ways probabilities of success are assigned to items using the two procedures and (2) differences in the ways raters conceptualize "minimum competence." They conclude that the differences between these two criterion-referencing procedures may be of greater consequence than their apparent similarities.

Recapitulation

It appears that experts in educational instruction and evaluation are not in agreement as to whether a norm-referencing approach or a criterion-referencing approach to measurement is preferable. Since the appearance of Glaser's delineation of the two constructs, criterion-referenced measures have been labeled by some as the panacea for evaluating an individual's ability or achievement without contamination of a relative scale. But as Ebel suggests, Glaser's review gave a name to the predictable return to criterion-referenced measures. Criterion-referencing has not been proved to be superior to norm-referencing. In fact, criterion-referenced measures have yet to be fully developed as a construct for evaluation and, therefore, have undergone little other than peripheral study.

While norm-referencing procedures connote a universally accepted method by which individuals may be evaluated, the state of the art of criterion-referencing is still in development. Considering the array of criterion-referencing models proposed, it appears that those models most conducive for use in setting scoring standards are those developed by Nedelsky, Ebel and the University of Illinois. If ideal examinations could be constructed to assure equal difficulty across all test items, then models based on binomial distributions or Bayesian theory may seem more applicable to examinations such as National Board examinations. Evaluating difficulty of test item content seems an important factor in determining pass/fail cutoff scores on licensure examinations.

But the issue still exists of which criterion-referencing procedure is most desirable. Those that purport to establish minimum passing levels by assessing individual test items are, by far, most attractive in that they appear to be easily used and easily understood. Yet, limited study has been conducted to validate their assumptions or even to assess consistency in results. In their research, Andrew and Hecht concluded that the methods developed by Ebel and Nedelsky elicit dissimilar examination standards and that the judgments of experts using identical standard-setting procedures result in significantly different minimum passing scores. Brennan and Lockwood also found that while the methods developed by Angoff and Nedelsky are similar, the variance in cutting scores resulting from different raters applying the two procedures is considerable. Yet, neither study assessed the value of particular criterion-referencing methods.

What now seems essential is study of the comparisons and contrasts between various criterion-referencing procedures, not in terms of underlying theories and assumptions, but in terms of practical significance. Because the methods suggested by Ebel and Nedelsky include, as a component, assessment of the levels of item difficulty, it appears that some comparison between how judges determine this criterion using each method could be researched. Also, it seems reasonable to assume that item analysis data generated from administration of an examination could be used to aid in validating at least the requirement of determining levels of difficulty inherent in both methods. Further study into the characteristics of criterion-referenced measures is essential before one is able to determine whether norm-referencing or criterion-referencing is preferable.

CHAPTER III

METHODOLOGY

A study was conducted to investigate the implications of using two criterion-referencing scoring procedures and two committees of experts to set scoring standards for a sample of National Board dental examinations. The initial phase of the study involved testing whether different committees using the same method establish similar scoring standards and whether different methods applied by the same committee produce similar scoring standards. The investigation also involved analyses of the criterion-referencing procedures to determine whether (1) each procedure is internally consistent so that replication of the study would produce similar results, (2) scaling components common to both procedures elicited similar results and (3) criterion-referenced measures produce scoring standards that correlate with actual performance of candidates for dental licensure.

Hypotheses

Ho_1: Two committees of experts using the same criterion-referencing method on the same examination content establish similar standards for scoring.

Ho_2: The same committee of experts using different criterion-referencing methods on the same examination content establish similar standards for scoring.

Partitioned Variables

1. Subjects/Cell Entries: For this study, subjects are defined as the multiple choice test items included in the sample of National Board dental examinations. Three examinations were selected for the study from the battery of National Board examinations that test knowledge of the clinical dental sciences: the oral pathology and oral radiology examination (hereafter referred to as oral pathology); the oral surgery and pain control examination (hereafter referred to as oral surgery) and the operative dentistry examination. The examinations had been administered to over 2,000 candidates for dental licensure during the April 1978 testing session. The examinations were selected on the basis of statistical data collected following the April 1978 administration of the examinations. Analysis of the examinations is presented in Table 1.

2. Criterion-Referencing Methods: Because National Board examinations are licensure examinations, their purpose is to

Table 1

Descriptive Statistics for Selected National Board Examinations

		No. of			Reliability	Minimum
	No. of	Options		Standard	Coefficient	Passing
Examination	Test Items	Per Item	Mean	Deviation	(KR21)	Raw Score
Oral Pathology	97	3-8	74.84	7.33	0.67	56
Oral Surgery	97	3-8	70.60	7.17	0.61	53
Operative Dentistry	97	3-8	75.28	6.61	0.59	54

identify the small percentage of candidates who are not minimally competent to practice dentistry. To this end, of most significance is the pass/fail cutoff score established for each examination. The scoring methods proposed by Ebel and Nedelsky were selected for this study because both methods involve determining minimum passing scores.

Ebel's method arrives at a minimum passing score by assessing the degree of difficulty and the relevance of each test item in terms of performance expected of a minimally qualified candidate. Once all items are cross-categorized, judges assign a percentage to each cross-category that defines how many test items a minimally qualified candidate should be able to answer correctly. The number of items in each cell is multiplied by the percentage assigned to the cell. The sum of all cross-products is the minimum passing score. An example of using Ebel's method on five hypothetical test items is presented in Table 2.

Nedelsky's method arrives at a minimum passing score by determining for each test item the probability that a minimally qualified candidate will select the correct response. Judges are asked to eliminate for each item those distractors that a barely passing candidate would know are

Table 2

An Example of Ebel's Method

Applied to Five Hypothetical Test Items

Item Relevance	Item Difficulty		
	Easy	Medium	Difficult
Essential	Item #1 100%	75%	Item #3 Item #5 60%
Important	70%	Item #2 50%	40%
Acceptable	50%	35%	15%
Questionable	5%	5%	Item #4 0%

Essential x Easy = 1 item x 100% = 1

Essential x Medium = 0 items x 75% = 0

Essential x Difficult = 2 items x 60% = 1.20

Important x Easy = 0 items x 70% = 0

Important x Medium = 1 item x 50% = 0.50

Important x Difficult = 0 items x 40% = 0

Acceptable x Easy = 0 items x 50% = 0

Acceptable x Medium = 0 items x 35% = 0

Acceptable x Difficult = 0 items x 15% = 0

Questionable x Easy = 0 items x 5% = 0

Questionable x Medium = 0 items x 5% = 0

Questionable x Difficult = 1 item x 0% = 0

2.7 =

3.0 = minimum passing score

incorrect. The probability of choosing the correct response is the reciprocal of the number of remaining alternatives for each item. The sum of all reciprocals is the minimum passing score. An example of using Nedelsky's method on five hypothetical test items is presented in Table 3.

3. Committees: Six members of state licensing boards for dentistry were selected to comprise the two three-member committees of experts. In addition, one state board member who is familiar with the structure of National Board examinations was selected to serve on both committees to explain the purpose of the study and the criterion-referencing procedures to be used. Therefore, each committee included four judges.

Committee members were not randomly selected; of more importance was ensuring a representative sample of the geographic areas in which National Board examinations are administered. Geographic distribution of judges seemed important to modify any regional differences that may exist concerning the practice of dentistry. Members of state dental examining boards were selected because of their familiarity with examinations for dental licensure and because most serve dentistry as both examiners and general practitioners.

Table 3

An Example of Nedelsky's Method

Applied to Five Hypothetical Test Items

Options for Each Item	Test Items				
	Item #1	Item #2	Item #3	Item #4	Item #5
Choice A	Eliminate				Eliminate
Choice B	Eliminate			Eliminate	Eliminate
Choice C	Eliminate		Eliminate	Eliminate	
Choice D			Eliminate	Eliminate	

	No. of Choices Remaining	Reciprocal	Probability of Success
Item #1	1	1/1	1.00
Item #2	4	1/4	0.25
Item #3	2	1/2	0.50
Item #4	1	1/1	1.00
Item #5	2	1/2	0.50
			3.25 = 3 = minimum passing score

Procedures for Obtaining Data

Committee Functions: The two committees met independently to apply the criterion-referencing methods. Each committee was assigned the task of determining minimum passing scores for the three National Board dental examinations selected for the study. Each committee employed Ebel's method and Nedelsky's method on either one or two examinations. Committee assignments were determined in advance of the meetings. Committee assignments and the order in which examinations were reviewed are presented in Table 4.

Table 4

Committee X Criterion-Referencing Method Assignments

	Committees	
Examinations	Committee 1	Committee 2
Oral Pathology	Ebel's Method	Ebel's Method
Operative Dentistry	Ebel's Method	Nedelsky's Method
Oral Surgery	Nedelsky's Method	Nedelsky's Method

Conduct of Meetings: Identical explanations and instructions were given by the same individual to both committees. First, committee members were presented an explanation of the purpose of the study. Discussion began with a review of the norm-referenced system currently used to score National Board examinations. Basic assumptions underlying norm-referenced and criterion-referenced approaches to measurement were described to clarify the differences in the approaches. The criterion-referencing procedures of Ebel and Nedelsky were noted as having been selected for the study. Because National Board examinations are licensure examinations, the score that distinguishes those who pass from those who fail is most important. Both Ebel's and Nedelsky's procedures are based on determining a minimum passing score--the point below which failing scores fall.

Next, the conduct of the study was outlined. Each committee was instructed that its task was to superimpose a selected criterion-referenced scoring procedure on each of three National Board examinations that had been administered and scored using the norm-referenced scoring system. Each committee was also instructed to report its results as a consensus judgment rather than as individual member ratings. Each committee was made aware of its assignments and those of the other committee. Members were also informed that results of the study would be reported.

Oral and written descriptions of and instructions for using Ebel's and Nedelsky's procedures were provided. It was explained that both procedures arrive at a minimum passing score through analysis of individual test items. In analyzing items, the reference point is performance expected of a minimally qualified (barely passing) candidate for licensure. Samples of written instructions for using Ebel's and Nedelsky's methods and worksheets distributed to committees are attached as Appendices A and B.

Statistical Analyses

1. Differences Between Methods and Differences Between Committees: To test for statistically significant differences between methods, differences between committees and interaction effects, a repeated measures (split plot) design for a two-way analysis of variance was completed using oral pathology and oral surgery as halves of the same examination. The assumption of equivalent halves was tested with a t-test between means using performance statistics obtained from the April 1978 administration of the examinations. The results of the t-test are provided in Table 5.

Table 5

t-Test Between Means of

Oral Pathology and Oral Surgery Examinations

	Statistics			
Examinations	No. of Items	Mean	Standard Deviation	t
Oral Pathology	97	74.84	7.33	-7.77*
Oral Surgery	97	70.60	7.17	

*significant at p<.01

Because the oral surgery examination produced a lower mean raw score than did the oral pathology examination, results of the t-test between means proved to be statistically significant.

An F_{max} test for homogeneity of variances did not reach statistical significance (F_{max} = 1.05, not significant at .01). Assuming equal variance across all items, an adjustment of scores was planned. The difference between means (4.24) was assumed to be evenly distributed across all 97 test items. The transformation selected involved adjusting the oral surgery items by adding .04 (4.24/97 = .04) to the value assigned to each item by the committees. The transformation adjusted for differences in means while maintaining homogeneous variances.

Using Ebel's method, the value or weighting assigned to a test item was defined as the percentage assigned by a committee to the cross-category (relevance x difficulty) in which the item fell. Using Nedelsky's method, the value or weighting assigned to a test item was defined as the probability (expressed as a decimal) assigned to the item by a committee. The possible ranges of values differ for the two methods. In Ebel's procedure, a weighting of 100% is unlikely while a weighting of 0% is common. In Nedelsky's procedure, a weighting of 100% is common, while a weighting of 0% is impossible. Because of the difference in scales, values assigned to test items were transformed to produce homogeneity of variances among cells of the crossbreak. The transformation found to fit the data best was $\sqrt{X + 0.5}$.

The crossbreak for the repeated measures design for a two-way analysis of variance follows.

Repeated Measures Two-Way ANOVA: Methods x Committees

Methods	Committees	
	Committee 1	Committee 2
Ebel	97 item values	97 item values
Nedelsky	97 item values	97 item values

2. Stability of Each Method: To investigate whether a method is internally consistent in the minimum passing level it produces, each method was broken down into its measurement components and analyzed by component and overall. For this study, measurement or scaling components were defined as the judgments a committee must make to evaluate a test item. For example, in Ebel's procedure, a judgment is made about relevance of an item; in Nedelsky's procedure, a judgment is made about difficulty of a distractor. Because committees applied the same method to identical test items (Ebel's method to oral pathology items and Nedelsky's method to oral surgery items), the decisions of the two committees were matched by item and by distractor and analyzed.

Ebel's Method: Three scaling components were identified for Ebel's method: (1) assignment of an item to a relevance category, (2) assignment of an item to a difficulty level and (3) assignment of a percentage to a block (cross-category of relevance x difficulty).

A Pearson product-moment correlation coefficient was calculated to determine the relationship between the relevance categories assigned to items by the two committees. In Ebel's procedure, the only numbers assigned are percentages to cross-categories. Because percentages reflect relevance _and_ difficulty of test items, a scale of relevance values and a scale of difficulty values were derived for each committee. To define each scale, the "medium" level of difficulty was identified as the center of an axis and assigned an arbitrary value of zero. Twelve equations were constructed based on all possible combinations of the four relevance categories and the three levels of difficulty. By solving the equations (using averages for some subcategories), values that could be correlated were derived. An example of how values were derived using hypothetical percentages assigned to cross-categories is presented in Table 6.

Using values derived for levels of difficulty, a Pearson correlation coefficient was calculated to investigate the consistency of the difficulty component of Ebel's procedure. A comparison of the percentages assigned to

Table 6

An Example of Deriving Values for Relevance Categories and Levels of Difficulty Using Hypothetical Percentages

	Hypothetical Data		
	Difficulty		
Relevance	Easy	Medium	Difficult
Essential	90 %	80 %	70 %
Important	75 %	60 %	50 %
Acceptable	40 %	40 %	30 %
Questionable	10 %	5 %	5 %

EQUATIONS

GIVEN: Medium = 0

Essential + Medium (0) = 0.80	Essential = 0.80
Important + Medium (0) = 0.60	Important = 0.60
Acceptable + Medium (0) = 0.40	Acceptable = 0.40
Questionable + Medium (0) = 0.05	Questionable = 0.05
Essential (0.80) + Easy = .90	Easy = 0.10
Important (0.60) + Easy = .75	Easy = 0.15
Acceptable (0.40) + Easy = .40	Easy = 0
Questionable (0.05) + Easy = .10	Easy = 0.05

0.30/4 = 0.075

Easy = 0.075

Table 6 continued

Essential (0.80) + Difficult	= 0.70	Difficult	= -0.10
Important (0.60) + Difficult	= 0.50	Difficult	= -0.10
Acceptable (0.40) + Difficult	= 0.30	Difficult	= -0.10
Questionable (0.05) + Difficult	= 0.05	Difficult	= 0
			-0.30/4
			= -0.075
		Difficult	= -0.075

SUMMARY OF DERIVED VALUES

Essential	= 0.80	Easy	= 0.075
Important	= 0.60	Medium	= 0
Acceptable	= 0.40	Difficult	- -0.075
Questionable	= 0.05		

blocks by the two committees will be presented in table form.

Item weightings were correlated to assess the overall stability of Ebel's method. As in the two-way ANOVA, item value was defined as the percentage assigned to the block into which the test item was categorized.

Nedelsky's Method: One scaling component was identified for Nedelsky's method: elimination of a distractor. Values for distractors were identified by arbitrarily assigning a +1 to a distractor that was eliminated and a 0 to a distractor that was retained. Distractors were correlated using a Pearson correlation coefficient. To assess overall stability of the method, probabilities assigned to items by the two committees were correlated.

3. Stability of Scaling Components Across Methods: Weightings assigned to test items on the operative dentistry examination were used for this portion of the study. Because each committee applied a different criterion-referencing method to this examination, overall methods and scaling components common to both methods can be compared.

A one-way analysis of variance was completed to test for significant difference in means of assigned item weightings. As in the repeated measures design for a two-way analysis of variance, item values were transformed using a $\sqrt{x + 0.5}$ transformation. The crossbreak for the one-way analysis of variance follows.

One-Way ANOVA: Methods

Ebel's Method Committee 1	Nedelsky's Method Committee 2
97 item values	97 item values

An underlying assumption of both Ebel's and Nedelsky's procedures is that different panels of judges applying a single method of criterion-referencing to well defined test items establish consistent standards for scoring. If this assumption is credible, committees should be discounted as a source of variance. Results of the one-way analysis of variance were used to assess consistency between methods. To further study consistency between methods, Ebel item percentages and Nedelsky item probabilities were correlated.

Item difficulty level was identified as the scaling component common to both methods. For Ebel's method, values for difficulty levels were derived using the procedure described earlier in Table 6. For Nedelsky's method, difficulty level was defined as the probability of success assigned to a test item. A Pearson correlation coefficient was calculated to assess the strength of the relationship between difficulty components assigned through the two methods.

4. <u>Relationship Between Criterion-Referenced Measures and Actual Performance Data</u>: Weightings assigned to test items on the operative dentistry examination were used for this portion of the study, again because both methods were applied to the items. Comparisons were made between assigned item values and actual performance data tabulated following the April 1978 administration of National Board examinations. Because performance data were collected on the high 27 per cent and the low 27 per cent of the population of candidates who took the examination, item difficulty was defined as the average of the per cents of high and low groups who chose the correct response. Difficulty level of a distractor was similarly defined as the average of per cents of high and low groups who selected the distractor as an answer.

Test items were correlated to determine the relationship between values assigned through each method and actual performance data. In Ebel's method, item value was defined as the percentage assigned to the block within which the item was categorized. In Nedelsky's method, item value was defined as the probability of success assigned to the item. Correlations were calculated for each method with performance data. Comparisons of the Ebel difficulty component and the Nedelsky eliminated distractor component with actual performance data will be displayed graphically.

CHAPTER IV

RESULTS

Differences Between Methods and Differences Between Committees

Minimum passing raw scores established by the two committees by applying the same criterion-referencing procedure to the same test items are presented in Table 7. Reported scores are based on the 97 items included in each the oral pathology and the oral surgery examinations.

Table 7

Minimum Passing Raw Scores Established by Two Committees Using Two Criterion-Referencing Methods on 97 Test Items

	Committees	
Methods	Committee 1	Committee 2
Ebel	33	46
Nedelsky	37	34

A repeated measures (split plot) design for a two-way analysis of variance was used to test for differences between methods, differences between committees and interaction effects. Item values were adjusted and transformed as planned to make the two examinations equivalent and to produce homogeneous variances among cells of the crossbreak. Estimated mean squares were calculated using methods and committees as fixed variables and test items as a random variable to identify appropriate error terms. Cell means and a summary of the analysis of variance are presented in Table 8.

Table 8 indicates that differences in committees and the interaction effect are statistically significant at the .01 level. To analyze the interaction effect, graphs of cell means are presented as Figures 1 and 2.

To further analyze the interaction effect, sums of squares were partitioned to test simple main effects and Tukey's test for honestly significant differences was completed. Results of these analyses are presented in Tables 9 and 10 respectively.

Graphs of cell means show interaction across both levels of methods and committees. Tables 9 and 10 indicate that Committee 1 using Ebel's method produced results significantly different from other cell means.

Table 8

Repeated Measures Two-Way ANOVA

Methods X Committees

Methods	Committees	
	Committee 1	Committee 2
Ebel	Mean = 0.8980	Mean = 0.9675
Nedelsky	Mean = 0.9554	Mean = 0.9406

Source of Variance	d.f.	SS	MS	F
Methods	1	0.02250	0.02250	0.47
Items (Methods)	192	9.23149	0.04810	-
Committees	1	0.07269	0.07269	19.44*
Methods x Committees	1	0.17241	0.17241	46.10*
Committees x Items (Methods)	192	0.71783	0.00374	
Total	387	10.21692		

*significant at p<.01

Figure 1

Cell Means by Criterion-Referencing Method

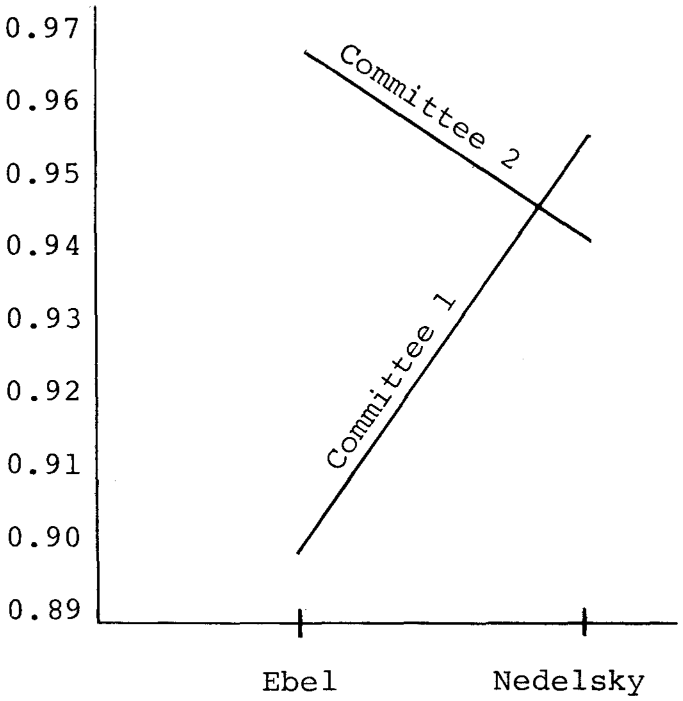


Figure 2

Cell Means by Committee

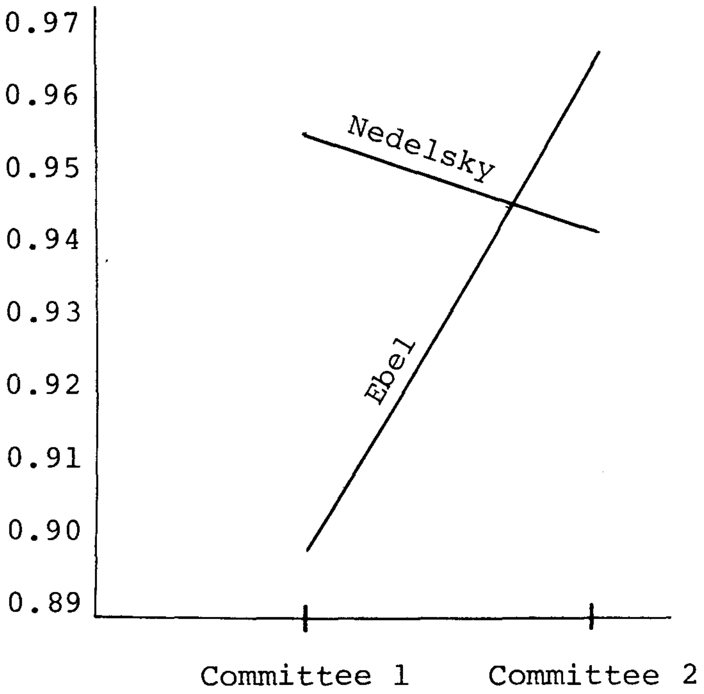


Table 9

Results of Test of Simple Main Effects to Identify Source of Interaction

Error Term for Testing Methods = 0.02591

Error Term for Testing Committees = 0.00374

Source of Variance	d.f.	SS	MS	F
Methods at Committee 1	1	0.15976	0.15976	6.17
Methods at Committee 2	1	0.03517	0.03517	1.36
Committees at Ebel's Method	1	0.23451	0.23451	62.70*
Committees at Nedelsky's Method	1	0.01060	0.01060	2.83

*significant at p<.005

Table 10

Results of Tukey's Test for Honestly Significant Differences

to Identify Source of Interaction

Critical value of HSD = 0.0273 (.01 level of significance)

	Ebel's Method Committee 1	Nedelsky's Method Committee 2	Nedelsky's Method Committee 1	Ebel's Method Committee 2
Ebel's Method--Committee 1	-	0.0426*	0.0574*	0.0695*
Nedelsky's Method--Committee 2		-	0.0148	0.0269
Nedelsky's Method--Committee 1			-	0.0121
Ebel's Method--Committee 2				-

*significant at p<.01

Stability of Each Method

Values assigned by the two committees to the same test items were used to analyze each criterion-referencing method separately. Each method was broken into its scaling components and investigated to evaluate consistency in results by overall method.

Ebel's Method: Three scaling components were analyzed to determine stability of the method: (1) assignment of an item to a relevance category, (2) assignment of an item to a difficulty level and (3) assignment of a percentage to a block (cross-category of relevance x difficulty).

(1) Assignment to Relevance Categories: Table 11 summarizes the agreement between the two committees in assigning the 97 test items to relevance categories. Cell entries represent number of items.

By solving the 12 equations for Relevance Category + Difficulty Level = Percentage Assigned to Block for each committee, the following values for the four relevance categories and the three difficulty levels were derived using the method illustrated in Table 6.

Table 11

Committee Agreement on Assignment of Items to Relevance Categories

	Committee 2			
Committee 1	Essential	Important	Acceptable	Questionable
Essential	16	3	1	0
Important	6	7	4	1
Acceptable	6	5	7	8
Questionable	1	2	12	18

COMMITTEE 1		COMMITTEE 2	
Relevance Categories		Relevance Categories	
Essential	= 85	Essential	= 85
Important	= 50	Important	= 80
Acceptable	= 15	Acceptable	= 40
Questionable	= 2	Questionable	= 0
Difficulty Levels		Difficulty Levels	
Easy	= 10.75	Easy	= 10.00
Medium	= 0	Medium	= 0
Difficult	= -6.75	Difficult	= -11.25

Derived values for relevance of items were correlated to determine the strength of the relationship between assignments of items to relevance categories. A Pearson correlation coefficient of +0.63 was produced.

(2) Assignment to Difficulty Levels: Table 12 summarizes the agreement between committees in assigning the 97 test items to levels of difficulty. Cell entries represent number of items.

Values derived for item difficulty were correlated to investigate the relationship between item assignments to levels of difficulty. A coefficient of +0.41 was produced.

Table 12

Committee Agreement on Assignment of Items to Difficulty Levels

	Committee 2		
Committee 1	Easy	Medium	Difficult
Easy	21	7	13
Medium	4	6	9
Difficult	5	4	28

(3) Assignment of Percentages to Blocks: Table 13 summarizes the agreement between committees in assigning percentages to cross categories of item relevance x item difficulty. Cell entries represent assigned percentages.

An overall comparison of agreement between committees in applying Ebel's procedure is presented in Table 14.

Table 14

Comparison of Decisions of Two Committees Applying Ebel's Method to 97 Test Items

Decisions of Committees	Number of Test Items
Committees agreed on both relevance and difficulty categories	32*
Committees agreed on either relevance or difficulty category	39
Committees disagreed on both relevance and difficulty categories	26
	97 test items

*Committees also agreed on the percentages assigned to cells for 14 of these 32 items.

Table 13

Committee Agreement on Assignment of Percentages to Cells

Relevance Categories	Difficulty Levels					
	Easy		Medium		Difficult	
	Committee 1	Committee 2	Committee 1	Committee 2	Committee 1	Committee 2
Essential	90 %	95 %	85 %	85 %	80 %	75 %
Important	75 %	90 %	50 %	80 %	40 %	65 %
Acceptable	25 %	60 %	15 %	40 %	5 %	20 %
Questionable	5 %	0 %	2 %	0 %	0 %	0 %

To investigate the overall stability of Ebel's method, test items were correlated using the percentage assigned to the cell in which the item was classified as the value of the item. A correlation coefficient of +0.67 was produced.

Nedelsky's Method: One scaling component was analyzed to determine stability of the method: (1) eliminated distractors.

(1) Eliminated Distractors: Table 15 summarizes the agreement between the two committees in eliminating (or retaining) distractors. The 97 test items included 344 distractors.

Table 15

Committee Agreement on Eliminating Distractors

Decisions of Committees	Number of Distractors
Committees agreed to eliminate	72
Committees agreed to retain	192
Committees disagreed on whether to eliminate or to retain	80
	344 distractors

By assigning a value of +1 to each distractor eliminated and a value of 0 to each distractor retained, distractors were correlated to determine the relationship between decisions made on distractors. A Pearson correlation coefficient of +0.48 was produced.

An overall comparison of agreement between committees in applying Nedelsky's procedure to 97 test items is presented in Table 16.

Table 16

Comparison of Decisions of Two Committees Applying Nedelsky's Method to 97 Test Items

Decisions of Committees	Number of Test Items
Committees agreed to eliminate identical distractors (or no distractors)	50
Committees agreed to eliminate at least one but not all identical distractors	25
Committees agreed to eliminate no identical distractors	22
	97 test items

Using the probability of success assigned to a test item as its value, test items were correlated to analyze the stability of results obtained through Nedelsky's procedure. A correlation coefficient of +0.56 was produced.

Stability of Scaling Components Across Methods

Minimum passing scores established by the two committees using different criterion-referencing methods on the operative dentistry examination are presented in Table 17. Reported scores are based on the 97 test items in the examination.

Table 17

Minimum Passing Raw Scores Established by Two Committees Using Two Criterion-Referencing Methods on the Operative Dentistry Examination

Ebel's Method Committee 1	Nedelsky's Method Committee 2
51	41

A one-way analysis of variance was performed to test for significant differences between methods/committees. Item values were adjusted to provide homogeneity of variances. Because each method is, in practice, purported to produce stable scoring standards across committees of judges, the analysis of variance was actually performed to further investigate the consistency between methods. Cell means and a summary of the analysis of variance are presented in Table 18.

Table 18

One-Way ANOVA: Methods

Ebel's Method Committee 1	Nedelsky's Method Committee 2
Mean = 0.9906	Mean = 0.9534

Source of Variance	d.f.	SS	MS	F
Between Cells	1	0.06733	0.06733	2.50
Within Cells	192	5.17026	0.02693	
Total	193	5.23759		

Table 18 indicates no significant difference between methods.

Values assigned to test items were correlated to determine the overall consistency with which the two methods elicited similar item weightings. In Ebel's method, the percentage assigned to the cross-category in which a test item was classified defined its value. In Nedelsky's method, the probability of success assigned to an item defined its value. Correlating these values produced a coefficient of +0.20.

Item difficulty was identified as the scaling component common to both procedures. By solving the 12 equations for Relevance Category + Difficulty Level = Percentage Assigned To Block using the procedure outlined in Table 6, derived values for difficulty of items reviewed through Ebel's method were obtained. Values derived for levels of item difficulty follow.

Difficulty Levels		
Easy	=	+8.25
Medium	=	0
Difficult	=	-21.75

Probability of success assigned to an item through Nedelsky's method was interpreted as an indicator of item difficulty. Derived values for item difficulty and assigned probabilities were correlated and produced a Pearson correlation coefficient of +0.32.

Relationship Between Criterion-Referenced Measures and Actual Performance Data

To determine whether weightings assigned to items through criterion-referencing procedures are consistent with performance by candidates for licensure, item values assigned through methods and actual performance data collected after administration of the operative dentistry examination were compared. The relationship between each method and National Board item analysis data for the 97 operative dentistry test items was analyzed separately.

Ebel's Method: With Ebel's method, the percentage assigned to the cell into which an item was categorized defined the item value. The average performance of candidates who selected the correct response identified the value of actual performance on the item. Ebel item weightings and item performance values were correlated to determine the strength of the relationship between the two measures. A correlation coefficient of +0.12 was produced.

Because average performance on an item is an index of the actual difficulty of the item, performance values were compared with difficulty levels assigned to items through

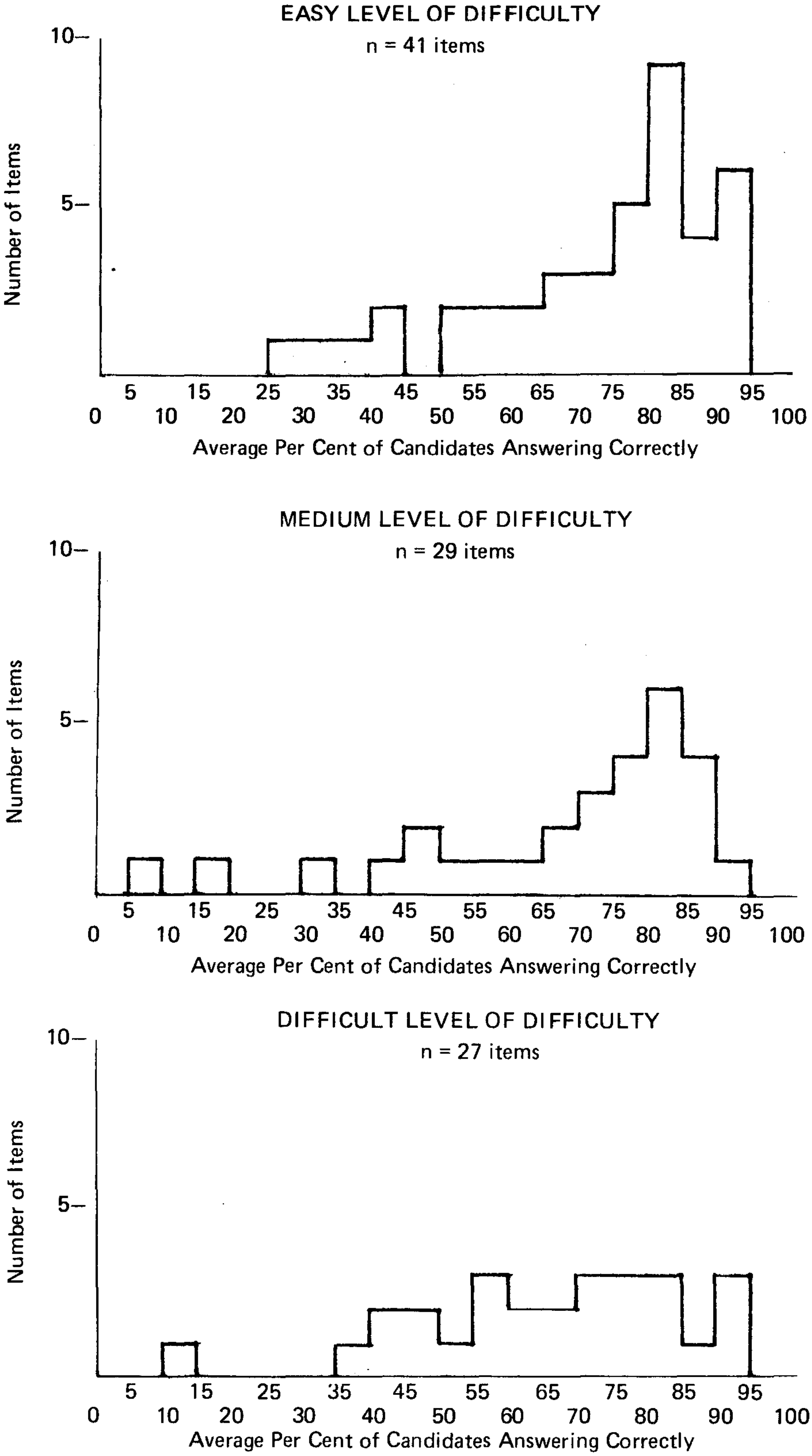
Ebel's procedure. If the level of difficulty assigned to an item in Ebel's method is consistent with how candidates for licensure performed on the item, one could expect a relatively large percentage of candidates to have answered correctly those items labeled "easy," a lesser percentage of candidates to have answered correctly those items labeled "medium" and a relatively small percentage of candidates to have answered correctly those items labeled "difficult." Histograms showing actual performance on 97 items assigned to each of Ebel's three levels of difficulty are presented as Figure 3.

<u>Nedelsky's Method</u>: With Nedelsky's method, the probability of success assigned to an item defined the item value. Item probabilities were correlated with actual average performance on the items to determine the strength of the relationship between these two measures. A correlation coefficient of +0.20 was produced.

Using Nedelsky's method, whether a distractor is eliminated or retained indicates the level of difficulty of the distractor. Actual performance values for distractors were compared with the committee's decisions to eliminate or retain distractors. If the decision to eliminate or retain

Figure 3

Histograms of Performance Data by Ebel Levels of Difficulty



EASY LEVEL OF DIFFICULTY
n = 41 items

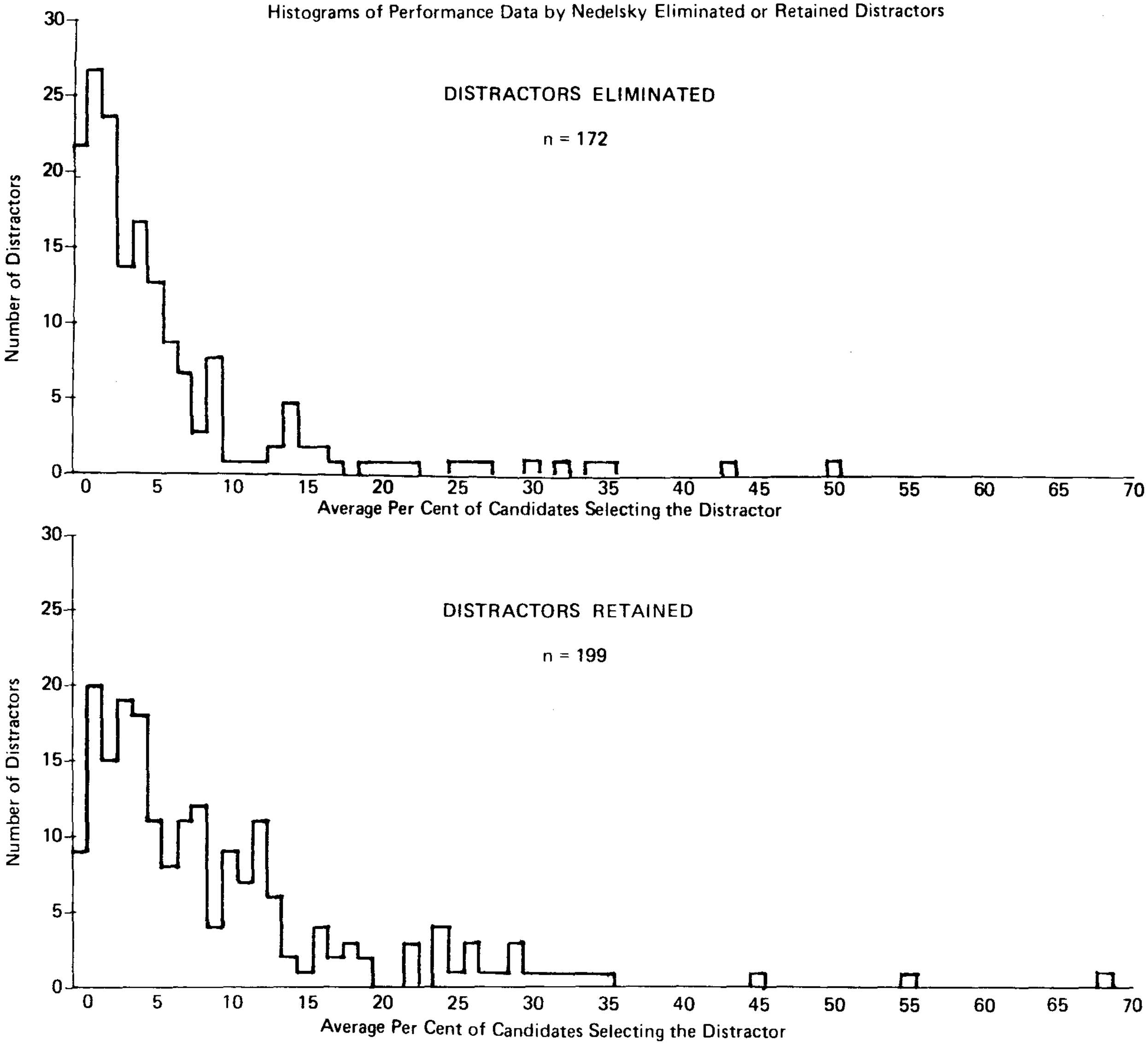
MEDIUM LEVEL OF DIFFICULTY
n = 29 items

DIFFICULT LEVEL OF DIFFICULTY
n = 27 items

a distractor is consistent with how candidates for licensure performed on the distractor, one could expect a relatively small percentage of candidates to have selected an "eliminated" distractor as the correct response and a relatively large percentage of candidates to have selected a "retained" distractor as the correct response. Histograms showing actual performance on 371 distractors determined, through Nedelsky's procedure, to be eliminated or retained are presented as Figure 4.

Figure 4

Histograms of Performance Data by Nedelsky Eliminated or Retained Distractors



CHAPTER V

DISCUSSION

Differences Between Methods and Differences Between Committees

At the outset of the study, it was believed that if Ebel's and Nedelsky's criterion-referencing procedures were stable, different committees of experts using the same method would establish similar scoring standards for the same examination content, and different methods used by the same committee of experts on the same examination content would elicit similar standards for scoring. In fact, the results of an analysis of variance of methods x committees indicate a statistically significant difference between committees and a significant interaction effect. Because interaction is evidenced, identifying what caused the interaction is of prime importance.

Graphs of cell means and results of the tests for simple main effects and "honestly" significant differences indicate that significant interaction occurred at the level of Committee 1 using Ebel's method. Further study of the graph allows for speculation on the relationship between minimum passing scores produced by the two methods. Committee 2 using Ebel's method produced the highest cell mean. Because this mean did not differ significantly from the results of Nedelsky's procedure, it is suggested that Ebel's method elicits higher minimum passing levels than does Nedelsky's method.

The significant interaction effect raises questions regarding the similarity between committees and the stability of methods. It could be argued that because committee members represented diverse geographic regions, the way in which individual members conceptualized relevance and difficulty of test items may have differed. Each committee's results were reported as a consensus of the judgments made by the committee to minimize the effects of dissimilarity among members in their approaches to evaluating test items. Using committees with more than four members may provide better control of this variable.

Stability of Each Method

It appears that the stability of a criterion-referencing method is a function of the stability of scaling components inherent in the method. The extent of agreement between committees on scaling components of a method suggests indices by which the reliability of the overall method can be assessed.

Ebel's Method: The consistency with which committees assigned relevance categories to test items using Ebel's method (Table 11) exceeds randomness. If relevance categories had been randomly assigned to items, results would show committee agreement that approximately six items test knowledge of each essential, acceptable, important and questionable information. The extent of committee agreement in assigning levels of relevance to items suggests that the concepts of "important" and "acceptable" are difficult for committees to define operationally. Categories of "essential" and "questionable" relevance are easier to define. The correlation of committee assignments of relevance categories (+0.63) suggests noteworthy stability in the relevance component of Ebel's method.

The consistency with which committees assigned levels of difficulty to test items in Ebel's method (Table 12) exceeds randomness in only two of the three levels. Random assignment of difficulty levels to items would have resulted in committee agreement that approximately 11 items were "easy," 11 items were "medium" and 11 items were "difficult" to answer. Because so few items were assigned by both committees to the "medium" level, it appears that committees are unable to define operationally a "medium" level of difficulty. The correlation of committee assignments of difficulty levels (+0.41) suggests that the difficulty component in Ebel's method contributes less to overall stability of the method than does the relevance component.

A comparison of percentages assigned to cells by the two committees (Table 13) suggests little stability in this scaling component; a different pattern for assigning percentages is implied in the results of each committee.

Committee 1 appears to have assigned percentages by viewing the total matrix and the interrelationships of relevance categories and difficulty levels. No two cells were assigned the same percentage. Also, a descending order of percentages is noted beginning with the "essential/easy" cross-category, moving across difficulty levels and ending with the "questionable/difficult" cross-category. The pattern implies committee judgment that: knowledge of "essential/difficult" information is more important than knowledge of "important/easy" information, knowledge of "important/difficult" information is more important than knowledge of "acceptable/easy" information and knowledge of "acceptable/difficult" information is more important than knowledge of "questionable/easy" information.

Committee 2 appears to have assigned percentages to cells by viewing each relevance category as a separate component and difficulty levels as ordered steps within the component. A descending order of percentages is noted, but assignments to cells are not unique. This pattern implies committee judgment that item relevance and item difficulty are somewhat independent. The implied scale for relevance places "essential" at the top and "questionable" at the bottom; the implied scale for difficulty places "easy" at the top and "difficult" at the bottom.

A comparison of overall judgments made by the committees using Ebel's method (Table 14) implies inconsistency within the method. However, the correlation of test items (+0.67) suggests consistency that is most probably a function of the relative stability of the relevance component. It appears that greater stability in standards produced by Ebel's method would result if all items tested highly relevant information. Further study of the stability of Ebel's method might include a modification in the procedure. Committees of experts could be asked to evaluate the relevance of test items first. Items that are judged to test unimportant information could be replaced with more relevant items. Once an examination is constructed, committees would evaluate difficulty of items and assign percentages to these levels. It is suggested that more consistent scoring standards would result.

Another area for further investigation is the scaling component in Ebel's method of assigning percentages to cells. Because committees seemed to differ in how they operationally defined the interrelationships of relevance categories and difficulty levels, it is suggested that greater stability in scoring standards might be obtained if percentages were determined before items were evaluated. In fact, each cell may carry a predetermined percentage that is consistent across examinations.

Nedelsky's Method: The extent of committee agreement on eliminating and retaining distractors (Table 15) suggests that the scaling component in Nedelsky's method is relatively stable. The correlation of distractors (+0.48) lends some support to this supposition. Comparison of overall committee evaluations of items (Table 16) and a correlation of items (+0.56) suggest that Nedelsky's method is just about as consistent in the standards it produces as is Ebel's method.

A comparison of the correlation coefficient for distractors and the correlation coefficient for items suggests a potential source of error in Nedelsky's procedure. If committees eliminate the same number of distractors on a test item, the same probability of success is assigned to the item whether or not identical distractors are eliminated. In this light, perhaps the coefficient obtained by correlating distractors is a more appropriate estimate of the reliability of Nedelsky's method.

Stability of Scaling Components Across Methods

Results of the analysis of variance on identical items subjected to the two criterion-referencing methods indicate no statistically significant difference in methods. In light of the earlier finding that Committee 1 using Ebel's procedure produced significantly different scoring standards, it was thought that the difference in methods would reach statistical significance.

Correlation of Ebel item percentages and Nedelsky item probabilities (+0.20) suggests little similarity between standards produced by the two methods. Correlation of Ebel and Nedelsky difficulty values (+0.32) also supports the supposition that the two methods produce markedly different results. Lack of a statistically significant difference in methods seems to suggest that an uncontrolled variable may have contaminated results.

An assumption of and a purported advantage to using Ebel's or Nedelsky's procedure is that results are consistent across panels of experts. For this reason, committees were not controlled in this portion of the study. But earlier analyses suggest that unreliable scoring standards are produced by different committees using the same criterion-referencing procedure. Before the two procedures can be adequately evaluated, further study should be conducted to control for variability among committees.

Relationship Between Criterion-Referenced Measures and Actual Performance Data

Any method of setting scoring standards should be acceptable to the psychometric community and should contribute to the purpose of the examination. Before a criterion-referenced measure could be adopted for scoring National Board examinations, the validity of the measure must be confirmed.

Ebel's Method: Correlation of Ebel item percentages and average percentages of candidates who answered items correctly (+0.12) suggests almost no relationship between Ebel scoring standards and actual performance. Graphs of item performance data by Ebel level of difficulty (Figure 3) confirm an earlier supposition that committees are unable to define operationally the concept of a "medium" level of difficulty. The graph of items on the "easy" level displays an acceptable shape and position on the scale; it appears that "easy" items are identifiable. The shape of the graph of items on the "difficult" level seems appropriate, but its position is too high on the scale of actual performance. Overall, it appears that the committee misjudged the performance of candidates.

Nedelsky's Method: The correlation of Nedelsky item probabilities and actual candidate performance (+0.20) also suggests a weak relationship between these measures. Graphs of distractor performance data by distractors eliminated and distractors retained (Figure 4) suggest that the committee judged candidates to perform at a lower level than occurred. In general, the distractors eliminated were appropriately identified, but many of the distractors retained attracted low percentages of candidates and, by Nedelsky's method, should have been eliminated.

Conclusions

From the results of this study, it appears that neither Ebel's nor Nedelsky's method of criterion-referencing is well suited for establishing scoring standards for National Board dental examinations. The methods fail to account for factors that affect candidate performance.

In assigning percentages to cells in Ebel's procedure, no consideration is given to candidates' answering some test items correctly by guessing correctly. The method allows for assignment of 0% to a cell; this seems unreasonable. If standards produced by Ebel's method were used with vigor to score examinations, it is reasonable to assume that a small percentage of candidates would pass who are not minimally competent.

In Nedelsky's method, relevance of a test item is not considered. Determining which distractors would be eliminated by a minimally qualified candidate becomes more difficult when evaluating items that test questionable information.

If examinations were ideally constructed to test knowledge of only essential information, it is reasonable to suggest that both Ebel's and Nedelsky's procedures would elicit more stable scoring standards. Because no test is ideal, it appears that the most desirable type of criterion-referenced measure is one that combines the advantages of Ebel's and Nedelsky's methods--one that reliably accounts for item relevance, difficulty of distractors and the effects of guessing correctly. Currently, the most powerful variable in setting scoring standards is the method selected to obtain the measure.

Of note are the reactions of committee members to working with the two criterion-referencing procedures. At the outset of the study, it was thought that Nedelsky's method would be easier to use because it requires judgment on only one factor--elimination of distractors. While Committee 2 favored using Nedelsky's method, Committee 1 found Ebel's method easier to apply.

The ease with which a committee applies a criterion-referencing procedure might be related to how members operationally define "minimally competent." It appears that before a committee uses a method, a form of inter-rater reliability could be established. Committee members might be asked to apply the procedure to sample test items. Discussion of how individuals evaluated items could identify a common denominator for conceptualizing "minimally competent." This common denominator would serve as the baseline for evaluating all test items.

Before a decision can be made as to the value of criterion-referenced measures, it appears that further study is warranted. The results of this study suggest that criterion-referencing methods do not produce stable scoring standards. Too, the assumption of consistency across panels of judges in operationally defining "minimally competent" is questioned. Further investigation into the reliability and the validity of criterion-referenced measures is needed.

CHAPTER VI

SUMMARY

A study was conducted to investigate the stability, equivalence and other characteristics of two criterion-referencing methods for establishing scoring standards. Two panels of experts were asked to superimpose criterion-referenced measures on recently administered National Board dental examinations to test the hypotheses that different committees of experts using the same method on the same examination content establish similar scoring standards, and that two methods used by the same committee on the same examination content elicit similar scoring standards. Results of the initial phase of the study indicated that different committees using the same standard setting procedure on identical test items do not necessarily establish similar overall scoring standards, and that different standard setting procedures used by the same committee on equivalent samples of test content do not necessarily elicit similar scoring standards.

Study of the stability of each criterion-referencing method centered around investigating the internal consistency of measurement components inherent in each method. Correlations of committee decisions resulted in modest coefficients that indicated stability of the relevance component of Ebel's method and the eliminated distractors component of Nedelsky's procedure. The internal consistency of other measurement components was minimal.

Study of the consistency between the two criterion-referencing methods centered around comparing the overall procedures and comparing the difficulty components inherent in both methods. Data indicated a weak relationship between the scoring standards established using the two procedures. Correlation of the difficulty levels assigned to items also produced a weak relationship.

Scoring standards established through the two criterion-referencing procedures were compared with actual performance data collected after administration of an examination to determine the practical significance of using either method. Correlation coefficients indicated that standards established through either criterion-referencing method are unrelated to performance of candidates for licensure.

These results raise questions regarding the reliability and the validity of the criterion-referencing procedures investigated. The results also suggest that even when different methods are based on similar conceptualizations, markedly different scoring standards may result.

BIBLIOGRAPHY

Andrew, Barbara J., & Hecht, James T. An investigation of two criterion-referenced procedures for setting examination standards. Philadelphia: National Board of Medical Examiners, 1976.

Angoff, W. H. Scales with nonmeaningful origins and units of measurement. _Educational and Psychological Measurement_. 1962, _22_, 27-34.

Block, James H. Criterion-referenced measurements: potential. _The University of Chicago School Review_, 1971, _79(2)_, 289-298.

Brennan, Robert L., & Lockwood, Robert E. A comparison of two cutting score procedures using generalizability theory. _ACT Technical Bulletin No. 33_. Iowa City: American College Testing Program, 1979.

Cronbach, James H. Test validation. In R. L. Thorndike (Ed.), _Educational measurement_. Washington, D. C.: American Council on Education, 1971, 443-503.

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), _Problems in criterion-referenced measurement_. Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, 1974.

Ebel, Robert L. Criterion-referenced measurements: limitations. _The University of Chicago School Review_, 1971, _79(2)_, 282-288.

Ebel, Robert L. _Essentials of educational measurement_. Englewood Cliffs: Prentice-Hall, Inc., 1972.

Emrick, J. A. An evaluation model for mastery testing. _Journal of Educational Measurement_, 1971, _8_, 321-326.

Gardner, Eric F. Normative standard scores. Educational and Psychological Measurement, 1962, 22, 7-14.

Glaser, G. R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Hieronymus, A. N. Today's testing: what do we know how to do? Proceedings of the 1971 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1972.

Klein, S. Evaluating tests in terms of the information they provide. UCLA Evaluation Comment. Los Angeles: UCLA Graduate School of Education, 1970, 2, 1-6.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Downers Grove, Ill.: Institute for Educational Research, April, 1972.

Lindquist, E. F. Selecting appropriate score scales for tests. Proceedings of the 1952 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1953.

Meskauskas, John A. Evaluation models for criterion-referenced testing: views regarding mastery and standard setting. Review of Educational Research, 1976, 46(1), 133-158.

Millman, Jason. Reporting student progress: a case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 227-230.

Millman, Jason. Tables for determining number of items needed on domain-referenced tests and number of students to be tested. Instructional Objective Exchange, April, 1972.

Millman, Jason. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

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

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, 1974.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 7(1), 1-9.

Setting standards of competence: the minimum pass level. Chicago: University of Illinois Center for Educational Development, 1973.

Stake, Robert E. Testing hazards in performance contracting. Phi Delta Kappan, 1971, 53, 583-589.

Thorndike, Robert L. (Ed.). Educational measurement. Washington, D. C.: American Council on Education, 1971.

APPENDIX A

ESTABLISHING CRITERION FOR MINIMUM PASSING SCORE
EBEL'S METHOD

Committee Function: Committee members are asked to establish a minimum passing score by analyzing test items for degree of difficulty and relevance in terms of performance expected of a minimally qualified (barely passing) candidate.

Procedure:

1. For each item, determine level of difficulty and level of relevance and assign the item to the appropriate cross-category in the grid.

2. Determine the expected percentage of passing for items in each category. These percentages indicate the passing level expected of a minimally qualified candidate.

3. The minimum passing score is the sum of products of number of test items in each category X percentage assigned to the category.

EXAMINATION: ________________________________

Relevance Categories	Difficulty Levels		
	Easy	Medium	Difficult
Essential	%	%	%
Important	%	%	%
Acceptable	%	%	%
Questionable	%	%	%

# of Items x % = Product

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

________ x ___ = ________

____

Minimum Passing Score = ________

APPENDIX B

ESTABLISHING CRITERION FOR MINIMUM PASSING SCORE NEDELSKY'S METHOD

Committee Function: Committee members are asked to establish a minimum passing score by analyzing test items for probability of a minimally qualified (barely passing) candidate choosing the correct responses.

Procedure:

1. For each test item, determine the responses that could be rejected by a minimally qualified candidate as being incorrect and cross through these responses.

2. For each test item, determine the number of remaining responses and assign the reciprocal of that number to the item. The reciprocal indicates the chance for success for a minimally qualified candidate.

3. The minimum passing score is the sum of all reciprocals.

EXAMINATION: ________________________________

Item No.	Success Rate	Item No.	Success Rate	Item No.	Success Rate	Item No.	Success Rate	Item No.	Success Rate
1 -		21 -		41 -		61 -		81 -	
2 -		22 -		42 -		62 -		82 -	
3 -		23 -		43 -		63 -		83 -	
4 -		24 -		44 -		64 -		84 -	
5 -		25 -		45 -		65 -		85 -	
6 -		26 -		46 -		66 -		86 -	
7 -		27 -		47 -		67 -		87 -	
8 -		28 -		48 -		68 -		88 -	
9 -		29 -		49 -		69 -		89 -	
10 -		30 -		50 -		70 -		90 -	
11 -		31 -		51 -		71 -		91 -	
12 -		32 -		52 -		72 -		92 -	
13 -		33 -		53 -		73 -		93 -	
14 -		34 -		54 -		74 -		94 -	
15 -		35 -		55 -		75 -		95 -	
16 -		36 -		56 -		76 -		96 -	
17 -		37 -		57 -		77 -		97 -	
18 -		38 -		58 -		78 -		98 -	
19 -		39 -		59 -		79 -		99 -	
20 -		40 -		60 -		80 -		100 -	

_____ + _____ + _____ + _____ + _____

Minimum Passing Score = _____

APPROVAL SHEET

The thesis submitted by Maribeth Hladis has been read and approved by the following committee:

Dr. Jack A. Kavanagh, Director
Associate Professor and Chairman, Foundations

Dr. Ronald Morgan
Associate Professor, Foundations

The final copies have been examined by the director of the thesis and the signature that appears below verifies the fact that any necessary changes have been incorporated and that the thesis is now given final approval by the Committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirements for the degree of Master of Arts.

4/17/80
Date

Jack A. Kavanagh
Director's Signature