This is to certify that the

dissertation entitled

A STUDY OF JUDGMENTAL STANDARD SETTING METHODS

presented by

Ira J. Washington, III

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

Irwin J. Lehmann
Major professor

Date 6/15/92

A STUDY OF JUDGMENTAL STANDARD SETTING METHODS

By

Ira J. Washington, III

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

1992

ABSTRACT

A STUDY OF JUDGMENTAL STANDARD SETTING METHODS

By

Ira J. Washington, III

Judgmental standard setting techniques are used to set examination cut scores based upon item content. Past research has shown the three most often used techniques set different standards when applied to the same examination material. No attempt has been made to determine which standard setting procedure produces the most valid result, when compared to an external criterion. The problem of determining which procedure produces the most valid cut score standard for a particular situation will be the focus of this study.

There were 3 specific purposes in the study: 1) to replicate the results of previous standard setting research, 2) to determine whether the procedures differed in consistency of scores produced by the judges as individuals, and 3) to establish which procedure agreed most in pass/fail classification decisions beyond chance, when compared to an external criterion.

The following measures were used to study the research objectives. Four introductory statistics course instructors were asked to serve as expert subject matter judges for the study. The instructors had all taught the same course with similar materials and content. The course included three achievement examinations each paired with a minimum competency examination.

The judges set standards applying each of the three techniques over the three achievement examinations. Each judge, therefore set nine cut scores. With a minimum competency examination as the external criterion, the procedure(s) least discrepant with the external criterion was identified.

Under the conditions outlined for the study, the following conclusions were reached:

1) The standard setting procedures produce different mean number right cut scores when applied to common subject matter.

2) The standard setting procedures do not differ in the ranked inter-rater reliability of judges' ratings.

3) When the pass/fail decisions of the procedures were compared to those of an external criterion at both the 70% and 80% proficiency criterion levels the ranked Kappa coefficients did not differ.

ACKNOWLEDGMENTS

The author wishes to thank all the people who persisted with and encouraged my completion of this study. In particular thanks go out to my dissertation committee; and specifically to Dr. Irvin J. Lehmann who was the committee chair.

My parents Ira and Ruby who put me here, Brenda my wife, Lynne and Laura my genius kids made this all possible. Finally, thanks to Jack MacMahon and company for conducting our Sunday morn excursions into the wilds of Michigan.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

Introduction

According to Ebel (1979), tests are used in education to measure a student's achievement and evaluate his/her educational progress within the classroom. This paper is concerned with classroom achievement tests that are used to obtain a measurement of student learning. An item format often used in this type of test is the multiple choice question. Multiple-choice items are typically scored dichotomously with 1 for a correct answer and 0 for an incorrect answer. The correct answers are summed to obtain a total test score. This total test score may be expressed as a percentage of correct items out of the total of all test items and this percentage may be used to infer a quantitative measure of the student's educational progress.

The percentage of items answered correctly can be used to place examinees along a continuum of inferred knowledge about a well-defined knowledge domain. On the continuum, there is a point that divides examinees with sufficient knowledge, as indicated by their total test score, from those with insufficient information regarding the knowledge domain. This point is known as the cut or minimal cut score (Allen & Yen, 1979).

Theoretically, the minimal cut score is suitable for inferring whether the individual has obtained sufficient subject matter knowledge to be considered minimally competent in the content area of interest. However, there are differences of opinion as to the most appropriate

procedure for determining the cut score. The standard setting procedures examined in the following text assume that minimal competency is a continuously distributed ability and are classified as continuum models (Berk, 1986).

Background of the Problem

Several methods exist for setting the cut score (Millman, 1973; Glass, 1978; Berk, 1986). According to Berk, the standard setting procedures can be separated into three categories: 1) judgmental, 2) judgmental-empirical, and 3) empirical-judgmental. The judgmental techniques set standards based primarily upon decisions made regarding test-item content. Judgmental-empirical techniques are based primarily on judgments of item content and secondarily on student performance data. Empirical-judgmental procedures are based primarily on student test-performance data and secondarily on judgments of test-item content.

Berk (1986) has identified 23 standard setting techniques. Judgmental and judgmental-empirical methods accounted for 11 and 7 of the procedures respectively. The balance of the techniques were classified as empirical-judgmental. Of the 18 judgmental and judgmental-empirical techniques 11 were derived from the procedures of Nedelsky (1954), Angoff (1971) or Ebel (1972). The 11 judgmental methods with the exception of two were based either on the methods of Nedelsky (1954), Angoff (1971), or Ebel (1972). Four of the 7 judgmental-empirical procedures were based on Angoff's (1971) method. The procedures of Nedelsky (1954), Angoff (1971), and Ebel (1972), require judges or subject matter experts to set the minimal competence level based on their judgment of the minimally competent student's probability of successfully responding to individual test items.

Therefore the methods are categorized as "judgmental" and based upon the "continuum" model(Berk,1986, p. 147). Since these three procedures are used most often for standard setting, they were the research focus.

Problem Statement

Berk also reviewed studies of the different standard setting procedures. Twenty-two studies were reviewed and in each case the studies were comparative in nature. It was found that the procedures produced different mean standards when applied to the same test. It is important to know that the procedures produce different standards when applied to the same test. However, no attempt has been made to determine which standard setting procedure produces the most valid result. The problem of determining which procedure produces the most valid cut score standard for a particular situation will be the focus of this study.

Importance of the Problem

Knowledge of which procedure produces the most valid cut score is important in choosing a methodology in a particular situation. Since evidently the procedures produce different mean cut scores, it follows that one or more of the procedures should produce a standard that gives the optimal cut score. Arbitrary selection of a standard setting procedure can be of grave consequence when issues such as competency certification and employment opportunity are concerned. Therefore, a defensible process for selection of a standard setting procedure is necessary.

Differences in performance standards set for the minimally competent student should, in theory, be small when cut scores set by different procedures are compared. If the procedures are based upon the

same definition of what constitutes acceptable performance, a basis exists for comparing the outcomes produced by each method. It is imperative, considering the differences reported in the literature, to determine which specific procedure(s) is (are) best applied in a specific setting. Doing so would in part address the concern expressed by Glass (1978) over the inequities shown by minimal standards set through the different techniques. To determine the more appropriate method for a particular situation, one must examine the validity of each method subject to the usual limitations of such studies (Mehrens and Lehmann, 1984).

Statement of Purpose

Three purposes were the focus of the present study. They were:

1) Establish whether the procedures of Angoff, Ebel, and Nedelsky produce different cut scores when applied to the same test(s), a replication of previous results.

2) Determine whether the procedures differ in terms of the consistency of individual judge's ratings of the same item-test content (a test of inter-rater reliability).

3) Establish which procedure makes pass/fail classifications that agree most beyond chance with classifications made by an external criterion. This hopefully will be a contribution to new knowledge.

The three purposes constitute objectives that will lead to the goal of establishing which standard setting procedure functions best in a particular situation.

Overview

Comparative studies of the procedures appear in Chapter II, the literature review chapter. Contained within Chapter III are statements of the hypotheses that were tested, descriptions of the standard setting

techniques of Nedelsky, Ebel and Angoff, and the study participants. An outline of instructions to the participants followed by the experimental procedures are included. The experimental components, and descriptions of both the achievement and criterion examinations also appear in the chapter. Chapter IV includes the data analyses. The results of the study, a discussion and conclusions, in addition to implications for future research appear in Chapter V.

CHAPTER II

Literature Review

Glasnapp et al.(1983)

Glasnapp, Poggio, and Eros (1983) conducted a study that examined the validity of the assumptions underlying the Ebel and Angoff standard setting methods. For Angoff's procedure the assumption that judges would be able to approximate the probability of success for minimally competent students for each item was tested. Three assumptions were tested for Ebel's procedure: 1) judges can estimate the difficulty of items for students to be tested, 2) judges can classify items reliably according to item relevance, and 3) judges reliably assign expected proportions of correctly answered items for minimally competent examinees by cells within a relevance by difficulty item matrix.

Examinees for the study were all students in grades 2, 4, 6, 8, and 11 in a large midwestern state. The students were tested for minimal competency in the areas of reading and mathematics. The grade 2 tests had 45 items each, and the tests for the remaining grade levels consisted of 60 items each. Minimum competency standards were set for each grade level during Spring 1980 and Spring 1982. Approximately 100 teachers at each grade level were randomly assigned to use either the Angoff or Ebel standard setting technique in each of the two years for both subject areas.

To assess the validity of the Angoff method assumption, the mean probabilities of success predicted for the minimally competent student

were correlated with the item difficulties obtained by students classified by teachers as minimally competent and with the item difficulties for all examinees. Ebel's first assumption was tested by correlating obtained item difficulties with the item difficulty values estimated by the judges when setting standards. The second assumption regarding the judges' ability to rate items according to their relevance was tested by taking the average sums of each set of three items intended to measure a specific educational objective and correlating it with the mean teacher rating of curricular fit for the objective. (There were 15 objectives at grade 2 and 20 for each of the other grades.) The third assumption of the Ebel method was evaluated through comparisons of average percentage descriptive statistics.

The data of the Glasnapp et al. (1983) study related to Angoff's method and the first assumption of Ebel's procedure produced average correlations between the average difficulty estimates of the judges and the actual item difficulties for the minimally competent group that ranged from .40 to .50 for grades 8 and 11, and .60 to .70 for grades 2, 4, and 6. According to Glasnapp et al. the correlations were neither high enough nor consistent enough to warrant confidence in the judges' abilities to estimate item difficulty values for minimally competent examinees for either the Angoff or the Ebel procedures. When the standard set using Angoff's method was compared to the raw score mean for students judged to be minimally competent, differences ranged from -3.6 items for eighth grade reading to 5.52 for sixth grade reading. This inconsistency provided additional evidence of the lack of validity of the judgments produced by the Angoff method.

The evaluation of correlations between judges average item relevance ratings and teacher ratings of curricular fit by objective provided consistent

results across grade levels for mathematics but not for reading. For reading, the grade 2 correlation was very high (r= .934). The correlation for grade 8 was negative and moderately high (r= -.60), while the correlation for grade 11 (r= -.02) indicated no relationship. The authors suggested that a change in content of the reading test to emphasize life skills in grades 8 and 11 may have accounted for the observed correlations. The correlations for grades 4 (r= .37) and 6 (r= .40) were low and not significant.

The correlations for mathematics were high for grades 2 and 6 (r= .924 and .846, respectively). For grades 4 and 8 the correlations were r= .724 and r= .637, respectively. The correlation for the grade 11 test was low (r= .183), but again the grade 11 test emphasized life skills. For both subjects, the authors questioned the use of relevance ratings when "life skills" was part of the examination content. The third assumption involved in Ebel's standard setting technique was that after classifying all items into a matrix defined by levels of relevance and difficulty, judges could reliably assign expected item proportions correct for the minimally competent examinee. If so, the authors expected the judges' cell proportions to be consistent over time and perhaps over grade levels within the content areas. The cell proportions for 1980 and 1982 were compared by computing the absolute values of the differences between those percentages. The absolute differences were tabulated by content area and grade level. Overall, the absolute differences indicated the presence of variability in the cell percentages over time. The authors noted that larger absolute differences were found in the cell defined by difficulty "hard" and "questionable" relevance, and that in practice judges do not tend to place many minimum competency test items in this category. When this category was removed from the matrix, less variation was evident in the absolute differences.

Ninety-seven percent of the absolute differences were below 10 percent, in contrast to 78 percent when the difficulty "hard", "questionable" relevance cell was not removed.

Andrew & Hecht (1976)

In a study that compared the standard setting procedures of Nedelsky (1954) and Ebel (1972), Andrew and Hecht (1976) found that the two methods yielded significantly different cut scores. In their study, standards were set simultaneously for both methods by two groups of judges for a nationally administered examination.

Ebel's procedure was used on the even numbered examination items and Nedelsky's procedure applied to the odd numbered items. Initially the judges set standards as individuals and then the individual standards were averaged to obtain a cut score. Subsequently, the judges used a group consensus approach to set standards. After the test had been administered to several hundred examinees, the odd-even halves of the examination were tested for equivalence in terms of mean scores and standard deviations. The mean individual cut scores for Nedelsky's procedure were 50.3% and 53.7%, respectively for the two groups of judges. Consensus judgments for the two groups of judges were 46.3% and 57.3% respectively. Individual cut score means for the Ebel procedure were 68.8% for Group 1, and 68.0% for Group 2, and the consensus judgments were 68.4% and 67.6%, respectively. No statistically significant difference was found between the individual and group standards for the two groups for either of the two methods.

To determine whether the overall standards set with the two methods differed, a 2 x 2 analysis of variance (methods by groups) was conducted. Consensus standards were the dependent variables for the

analysis. A significant main effect was found for method but not for groups. The overall percentage correct cut scores for the Nedelsky and Ebel methods were 49 and 68 respectively. From the analyses, three conclusions were drawn: 1) the two standard setting procedures yielded different cut scores for equivalent examination material; 2) different groups of judges who used the same procedure set similar cut scores for equivalent examination material; and 3) for both methods, mean individually set cut scores did not differ from those set by a group consensus approach.

Skakun & Kling (1980)

Skakun and Kling (1980) compared the Nedelsky standard setting procedure and two modified versions of Ebel's method. The examination used in the study was taken from the national General Surgery Certification Test item library.

Test items in the library were classified according to relevance and taxonomy by the item writer and a test committee. Three categories of relevance (essential, important, and acceptable) and three categories of taxonomy (factual, comprehension, and problem solving) were used. Item difficulty values available from previous test administrations were used to classify the items as easy, medium or hard. Items with p values greater than .80 were categorized as easy, those with p values between .30 and .80 as medium, and while those with p values of less than .30 were classified as hard.

The first of the modified matrices was defined by the descriptions difficulty and taxonomy, and was referred to as Ebel I. Relevance and taxonomy were the marginal variables for the second matrix that was

referred to as the Ebel II modification. For both modified Ebel methods, judges were required to examine the items within each cell in a matrix and decide what proportion of the items the "barely qualifiable" examinee would be expected to answer correctly.

Eight judges who had participated in the General Surgery Test Committee set standards for the study. The judges first used Nedelsky's method and then six months later used the Ebel I and Ebel II procedures. The resulting cut scores were 66.7% for the Nedelsky method, and 69.7% and 71.7% for the Ebel I and II techniques, respectively. Nedelsky's method exhibited more variability than the modified Ebel procedures. Reliabilities for the mean ratings on the items were .98 for both modified Ebel procedures and .61 for Nedelsky's method.

As a possible rationale for the differences found between the standards, the authors cited Shepard's (1980) suggestion that the Nedelsky task of eliminating the wrong response was less difficult than selecting the correct answer to an item and therefore lower standards would result. Also the authors suggested that for the Nedelsky method, the fixed probability correct response scale increments may have contributed to the greater variability in that procedure.

Brennan & Lockwood (1980)

Brennan and Lockwood (1980) compared the Nedelsky (1954) and Angoff (1972) standard setting methods through the use of generalizability theory to examine multiple sources of error variance within each cut score procedure. Three random sources of variation were associated with each procedure: raters, items, and the interaction between items and raters.

Statistics were also computed for the variability of mean scores for each procedure assuming five raters and 126 items.

Generalizability results for both methods were characterized by large interaction variance components. All three variance components were larger for the Nedelsky procedure. When it was assumed there were five raters and 126 items, estimated variance components for generalizing over both raters and items and raters given an infinite number of items were larger for the Nedelsky method.

Both methods were also analyzed within a single design. A three factor analysis of variance (2 procedures by 5 raters by 126 items) was used to compare the procedures, there were 7 possible variance components for the analysis of variance table. The variance components that included procedure were large relative to the other components that again suggested more variability due to procedures than to raters. More variability was also evident in Nedelsky's procedure than in Angoff's method when the average inter-rater covariances were compared (.013 vs. .006). A possible explanation for the greater variability in the Nedelsky method was attributed to the fixed probabilities of success (.25, .33, .50, and 1.00) for the Nedelsky method. In contrast, the Angoff method used a continuous probability scale.

Harasym (1981)

Harasym (1981) also evaluated the standard setting procedures of Angoff (1972) and Nedelsky (1954). In a three year study, Harasym collected data from 212 medical students at the University of Calgary. Two types of items were used in the examinations for this study. The first, type A, was the multiple choice item in which the student, given the item stem, would select the best response from X alternative choices. The

second, type B items, consisted of a stem with K response alternatives. Examinees were given the task of selecting the alternative(s) that would make the item stem true.

Standards set for the study were specific to item type. For item type A, the Nedelsky procedure was utilized, and the criterion score was calculated as previously described. A modification of Angoff's method was used to set standards for the type B items.

Angoff's procedure was modified in part due to the nature of the items. Since the items were to be scored on a discrete scale the authors changed the procedure to make it appropriate to the scale. Judges were instructed to "Identify those alternatives that the minimally competent student must know to be either true or false." Each item was scored 5, 4, 3, 2, or 1, according to the sum of the number of response alternatives that the minimally competent student should recognize as correct or incorrect. The scoring method for the items differed from Angoff's original method in that a discrete scale was used to assign p (s) to items. The cut score for the minimally competent student was computed as the sum of the item probabilities of success.

Three exams were given over the period of three years from 1979 to 1981. The item types A and B were determined to be statistically equivalent over the three examinations. Standards set in terms of percentage correct items for 1979, 1980, and 1981 were 60%, 63%, and 60% for Nedelsky's procedure. The corresponding cut scores for Angoff's technique were 69%, 77%, and 78%. Over the three year period the average difference in standards was 13.5%. When the percentages of students classified as minimally competent were compared, 99% were categorized as minimally competent by the Nedelsky cut score versus 82% with Angoff's standard.

Harasym concluded that the Nedelsky method set lower cut score standards, and as a result, more students were classified as minimally competent than under the Angoff standard.

Behuniak et al. (1982)

Behuniak, Archambault, and Gable (1982) studied the Angoff and Nedelsky standard setting methods. Standards were set for a mathematics and a reading examination that were administered to 460 ninth grade students in a Connecticut school district. The mathematics exam had 90 items and covered 30 curriculum objectives; the reading exam had 80 items and covered 11 objectives. The Kuder-Richardson reliability estimates were KR20= .93 and .91 for the mathematics and reading exams, respectively. Thirty items from each test were rated by the judges in the study.

Judges for the study were divided into two groups, with two sub-groups of three judges assigned to Angoff's procedure and two sub-groups of four judges assigned to Nedelsky's procedure. Comparisons of the cut scores were made between the two sub-groups within each method and between groups by method for both the reading and mathematics examinations. From the results, it was evident that separate groups of judges who employed the same method did not arrive at the same standards.

A significant difference between methods was found for the reading test but not for the mathematics test. The authors concluded that the standard setting methods did produce different cut scores and that judges in the same group set closer standards to one another than to those standards produced by judges who used the same method in a different group.

Saunders et al. (1981)

Saunders, Ryan, and Huynh (1981) compared two versions of Nedelsky's standard setting procedure. For version I, judges were asked to categorize response options into two groups, those the minimally competent student would recognize as correct and those the student would not recognize as correct. Version II had a third response category of "undecided" for response alternatives that judges could not classify as in version I. Cut scores were calculated for version I with the procedure as originally presented by Nedelsky (1954). Version II used one half of the response options categorized as "undecided" in calculating the probability of success for each item.

Saunders et. al. used 118 graduate students in an introductory educational research course as judges. First, the students took the 40 item course midterm examination. Then, after the examinations were returned and discussed, the students were asked to set standards for the exam. Students were randomly assigned to two groups, Group A or B, to set standards for the examination. Group A (n=59) used Nedelsky I and Group B (n=59) used Nedelsky II to set standards for the midterm. To reduce the possible confounding effect based on the students' prior knowledge of the course standards, the students were not required to compute the resultant cut scores. No consensus cut scores were calculated; all standards were set individually.

The Kolomogorov-Smirnov test for medians indicated no difference in cut score distributions for Groups A and B. The median cut scores, when rounded to the nearest whole number, were identical for both versions of Nedelsky's method. The percentage of agreement with decisions made based on the course instructor's cut score were 94% and

96% for the Nedelsky I and Nedelsky II methods, respectively. In instances of disagreement for both versions of the Nedelsky procedure, the Nedelsky cut score would have passed the students whereas the instructor's standard would have failed the students. It was concluded that both versions of Nedelsky's method yielded similar results. The authors also suggested use of the group cut score medians due to their substantial agreement with the criterion of the instructor's established score. The median score also effectively reduced the variation in group cut scores because it was not influenced by extreme scores. In terms of the variance in the two procedures, version II was recommended even though no statistically significant difference in variance was found. The authors also suggested caution in the interpretation of the results because students who did not construct the examination, and who possibly did not have a broad knowledge of the content area, were used to set standards.

Discussion

Collectively, the reviewed studies suggest that standard setting procedures based upon "judging minimal competence" produce standards consistent across judges for a single method but highly variable across methods for groups of judges. With respect to assumptions, the Glasnapp et al. (1983) study of the assumptions in Ebel's and Angoff's methods revealed the questionable nature of the assumption that judges could predict item difficulty values for either method. The second assumption of Ebel's method, that judges could reliably classify items into relevance categories, was viewed as plausible. The third assumption of Ebel's method was that judges could reliably assign expected proportions of correctly answered items for minimally competent examinees by cells within a relevance by difficulty item matrix. This assumption was seen as reasonably valid under

the circumstances in which they were tested. Further consideration of validation of the assumptions inherent in other cut score setting methods and situations is warranted in view of these results.

Both studies of the standard setting procedures of Nedelsky and Ebel had cut scores set by Nedelsky's procedure that were higher than those for Ebel's method (Andrew and Hecht, 1976; Skakun and Kling, 1980). Nedelsky's method exhibited more variability than did Ebel's in the Skakun and Kling (1980) study. No comparison of variability was made for the second study.

The comparisons of Angoff's and Nedelsky's methods showed Nedelsky's method to be more variable, as it also was in comparisons with Ebel's method (Brennan and Lockwood, 1980). In the Harasym (1981) study cut scores were higher for Angoff's standard in every occurrence when compared to those set with the Nedelsky method. The final study found the two procedures set different cut scores for one test of two for which standards were set (Behuniak et al., 1982). The fact that Nedelsky's procedure yielded more variable cut scores when compared to the methods of Ebel and Angoff suggests that its greater variability may adversely affect the validity of the decisions made.

Some researchers have suggested that the difference in cut scores obtained across methods can be attributed to the specifics of the techniques involved in their calculation (Harasym, 1981; Scriven, 1978; Andrew and Hecht, 1976). More specifically, Harasym pointed out that the Nedelsky procedure uses a discrete probability scale to assign probabilities of the minimally competent student successfully responding to individual test items. The Angoff and Ebel methods assign probabilities on a continuous scale from 0 to 1.0. When the three scales were contrasted, it was evident

that probabilities for the Ebel and Angoff standard setting procedures had more uniform intervals than the Nedelsky procedure that had large initial changes. Thus, the Nedelsky procedure appeared to be the most different from the other standard setting methods. Harasym suggested that the Ebel and Angoff methods be compared to explore the possibility that their closer scale increments would result in standards less different than those noted in other pairwise comparisons that involved the Nedelsky technique.

Commenting on the Andrew and Hecht study, Hambleton (1978) suggested that since both the directions to the judges and the techniques differed, unequal standards should be expected. Hambleton also quoted Ebel (1972):

> "It is clear that a variety of approaches can be used to solve the problem of defining the passing score. Unfortunately, different approaches are likely to give different results." p. (496).

Both statements suggest important considerations when comparing cut score methodologies. Gross (1982) recommended the use of a group consensus approach to narrow the discrepancies reported in different standard setting procedures. Glass (1978), however, viewed the inequities "as virtually damning the technical work from which it arose." (p.249). Glass suggested that proponents of the "different method" rationale should show prior reasons for a preferred technique, and if none exist, admit the arbitrary nature of their choice.

Summary

The literature on cut score methodology reports inconsistent results by the various procedures (Andrew and Hecht, 1976; Harasym, 1981; Brennan and Lockwood, 1980). Such studies were comparative and had

focused only on describing the magnitude of differences between the cut scores set by the Nedelsky, Ebel, Angoff and other techniques. A need exists to address the question of which method provides the most valid cut score based on minimal competence. Knowledge that the methods provide different outcomes is important, but practitioners need evidence of which method is most consistent with an independently determined minimal criterion. With such evidence, decision makers can defend the choice of one technique versus another.

CHAPTER III

Methodology

Preview

Angoff, Ebel and Nedelsky have each suggested techniques that are used to determine absolute standards for achievement tests. The three techniques are based upon subject matter experts' (judges) decisions about the probabilities of students successfully responding to individual examination items. Past research (Glasnapp, 1983; Andrew & Hecht, 1976; Skakun & Kling, 1980; Brennan & Lockwood, 1980; Harasym, 1981; etc.) has demonstrated that these techniques provide different mean number right cut score standards when applied to the same examination.

While previous research has indicated that the procedures produce different cut score standards, research has not established which procedure(s) produces the optimal cut score standard. The objective of the present study was to provide a model that can be used to select the procedure(s) most appropriate for a particular situation. Comparison to an external criterion was used as the method to determine which of the three methods generated the best cut score standard.

The following measures were used to study the research objectives. Four introductory statistics course instructors were asked to serve as expert subject matter judges for the study. The instructors had all taught the same

course with similar materials and content. The course included three achievement examinations each paired with a minimum competency examination. The judges set standards applying each of the three techniques over the three achievement examinations. Each judge, therefore set nine cut scores. With the minimum competency examinations as the external criterion, the procedure(s) least discrepant with the external criterion was identified.

There were 3 purposes in the study: 1) to replicate the results of previous standard setting research, 2) to determine whether the procedures differed in consistency of scores produced by the judges as individuals, and 3) to establish which procedure agreed most in pass/fail classification decisions beyond chance, when compared to an external criterion.

The study served a secondary purpose with reference to the third purpose of the study. When the pass/fail classifications made by the procedures were compared to classifications made by the external criterion a possibility of two types of error existed. An individual classified as pass by a procedure cut score standard could have been placed into the fail category by the criterion examination decision or a student classified as fail by the procedure could be classified as pass by the criterion examination. The two types of error are referred to respectively as "false acceptance" and "false rejection" (Mehrens and Lehmann, 1984). There are situations in which the gravity of the two error types would differ. For example, the error of false acceptance is more serious than false rejection when a cardiologist is being certified.

Conversely, there are circumstances where false rejection is more serious than false acceptance. For example, assume a program exists at a

university that is designed to assist disadvantaged students in graduating from the university. If a yearly program evaluation falsely rejected the program as effective, the loss to society would arguably outweigh the consequences of false acceptance of the program's effectiveness. For the purposes of this study it was assumed the error types were equally serious. For descriptive purposes the frequencies of both error types will be discussed within the context of one error being more serious than the other.

Within this chapter are 1) statements of the hypotheses tested, 2) descriptions of the standard setting techniques of Nedelsky, Ebel and Angoff, 3) information about the study participants, and 4) general participant instructions. Specific participant instructions appear in the appendix. The course of the study, and descriptions of both the criterion and achievement examination are also included. Finally, a description of the analyses performed and statistics used complete the chapter.

Statement of Hypotheses

Concerning the first purpose of the study it was predicted that significant differences would be found in the mean number right cut scores produced by the Nedelsky, Ebel and Angoff techniques for each of three classroom achievement examinations. Regarding the second aim of the study, it was expected that one procedure of the three would exhibit a greater degree of inter-rater reliability for the cut score set on each of the three achievement examinations. Third, it was hypothesized one of the procedures would classify a higher proportion of students across the three examinations, beyond chance occurrence, into the same pass/fail categories as did the external criterion examination.

The proposed hypotheses stated in research form were as follows:

Hypothesis 1: H_1: The mean number right cut scores produced by the methods of Angoff, Ebel, and Nedelsky will be different.

Hypothesis 2: H_1: There will be a difference in the ranked inter-rater reliabilities of the mean number right cut scores produced by the procedures of Angoff, Ebel and Nedelsky.

Hypothesis 3: H_1: There will be a difference in the ranked proportion of students correctly classified, beyond chance occurrence, into pass/fail categories by the procedures of Angoff, Ebel and Nedelsky when compared to an external criterion.

Standard Setting Procedures

Nedelsky (1954)

When the multiple choice items are used with Nedelsky's (1954) method, the subject matter experts identify the alternative answer choices that they judge the minimally competent student should recognize as incorrect. The reciprocal of the number of remaining responses is the predicted probability of success (p (s)) on a particular item for the minimally competent student. The probabilities of success are summed across items to obtain a number right cut score for the examination (see appendix, PROCEDURE B INSTRUCTIONS).

Ebel (1972)

Ebel's method (1972) requires the subject matter experts to judge both the relevance and the difficulty of each test item. Ebel identified four levels of relevance, 1) essential, 2) important, 3) acceptable, and 4) questionable.

Difficulty was divided into three levels, 1) easy, 2) medium, and 3) hard. A 4 x 3 matrix formed by crossing the variables of relevance and difficulty was used in the procedure. Subject matter experts were asked to identify the cell in the matrix that best described the relevance and difficulty of each item for a minimally competent student. After all items have been classified, the judges then assigned a predicted probability of success for a minimally competent student (weight) to each cell of the matrix. The total number of items assigned to each cell was summed, and the sum multiplied by the weight associated with that cell. These products were summed across cells to obtain the number right cut score (see appendix, PROCEDURE C INSTRUCTIONS).

Angoff (1972)

For Angoff's (1972) method, each subject matter expert was asked to assign a probability of success, p (s), to each examination item based on his/her judgment of the ability of a minimally competent student to answer the question correctly. The probability was assigned according to the subject matter expert's own judgment of what constituted a minimally competent student. The assigned probabilities were summed across items to provide a number right cut score for the examination. The number right score was also converted to a percentage correct score (see appendix, PROCEDURE A INSTRUCTIONS).

Experiment Components

Participants. The judges in the study were four statistics instructors from a large midwestern university. Each instructor taught the same introductory statistics course covering similar text and course content.

Standards were set by the judges independently. Saunders, Ryan, and Huynh (1981) in commenting on the Andrew and Hecht (1979) study stated that cut scores obtained by averaging individual judgments did not differ significantly from standards based upon group consensus. Therefore, for the present study, the responses of the four instructors were averaged to determine the minimal passing score assigned by each method.

Standard setting handouts. Materials for the study included a set of instructions for each standard setting method. The procedures for each standard setting method were detailed in a step by step fashion, including illustrative examples. Instructions to the judges also included a definition of a minimally competent student for the purposes of the study. This was done so that each judge would start with the same concept of what constituted a minimally competent student. Each judge had each conducted the CEP 904 (introductory statistics) class in the recent past. A minimally competent student was defined as a student who would have barely earned the letter grade C (2.0) in the class.

Procedures. A set of general instructional materials was prepared for the standard setting techniques. The general instructions included: 1) a cover letter, 2) instructions specific to each procedure, 3) procedure response forms, and 4) a comment sheet. The general instructions for each week listed the sequence to be followed to ensure counterbalancing for method and order of presentation when participants utilized Procedures A (Angoff), B (Nedelsky), and C (Ebel). The materials were prepared in a similar paradigm for each of the three classroom examinations.

Counterbalancing for method was used to randomize effects due to order of presentation. These effects might include interference or learning

effects. The ordered assignments of procedures to judges for the study appear in Table 1.

Table 1

Order of procedures by exam and instructor*

Instructor	Exam I	Exam II	Exam III
1	C B A	A B C	B C A
2	A C B	B C A	C B A
3	B C A	C A B	A B C
4	B A C	C B A	A C B

* A = Angoff
B = Nedelsky
C = Ebel

Examination materials were distributed and collected as units, and approximately two weeks were allowed to set standards for each classroom examination.

The four judges were contacted individually to arrange a time for delivery of the standard setting handouts for the first examination (Exam I). Procedures were explained verbally and judges were given an opportunity to ask questions.

Examinations

Three pairs of examinations (each pair covering about 1/3 of the course) were administered during one term to approximately 75 students enrolled in introductory statistics. One test of each pair was a norm referenced classroom achievement examination; the other test was a minimum competency examination.

Criterion Examinations

Each minimum competency examination was made up of approximately 25 free response, short answer items and was designed to assess the minimum knowledge required to pass that third of the course. From past administrations and empirical data on student performance in subsequent course work, a 70% proficiency level was established on the minimum competency exam for a student to be classified as passing (equivalent to a 2.0 or C grade). On the basis of the continuum model assumption that minimum competency is a continuously distributed ability, an 80% proficiency level was also used for the study. The use of a single criterion level would imply the existence of minimum competency at a discrete level. Inclusion of the 80% proficiency level allowed for study of the range of minimum competency. This step was taken because the relationship of the standards set by the procedures to the criterion could be influenced by the level of the criterion.

The criterion proficiency levels were set and the assumption made that they provided an estimate of the range of minimum competency. In addition, the minimum competency examinations were considered quasi-independent criteria for the following reasons: 1) the minimum competency examination content was limited to essential course knowledge, and was easier than the achievement examination, 2) administration took place outside of the classroom environment, 3) the free response format allowed partial credit scoring, 4) examination content was specifically related to instruction and homework assignments, 5) the minimum competency examination content included questions similar to those on the achievement

examination, 6) parallel forms with answers were given to the students for review, 7) the items had been reviewed by an instructor from another university who had taught a similar course for many years, and 8) teaching assistants who had received training in test construction and scoring and had taught discussion (help) sessions for two years constructed the minimum competency examination. The minimum competency examinations had also been approved by the classroom instructors. As shown in Table 2 it was also the case that the criterion examination scores were positively correlated to the achievement examination test scores at both criterion levels.

Table 2

Achievement examination test score
and minimum competency examination decision correlations

Examination	Criterion level	
	70%	80%
1	.48	.51
2	.63	.53
3	.34	.39

Achievement Examinations

The classroom achievement exams were 30 item multiple-choice tests with five response choices for each item. Kuder-Richardson 20 reliabilities were .77, .78, and .79 for examinations one, two, and three respectively. One of the participating instructors constructed the achievement examinations.

Analyses

Design

Two within-subjects factors were included in the study. The first factor was procedure which had three levels. Examination, the second factor also had three levels. A two within-subjects factors repeated measures design was employed to determine whether differences existed between the average cut scores set by the procedures. Cut score standards were set with the three procedures for each of three examinations. A total of nine observations were taken on each of four expert subject matter judges. The mean number right cut scores produced by the instructors' use of each procedure were the dependent variables for the study. The two within subjects factors; procedure and examination, were the independent variables for the study.

Statistics

All statistics for the study were tested with α=.05. The first hypothesis was tested with a two within-subjects factors repeated measures analysis of variance. If H_0 was rejected for a factor, univariate F-tests were computed for that factor. After the univariate F-tests, contrasts were conducted to ascertain which specific within-subjects factor levels differed.

To determine if the procedures differed in consistency of ratings the inter-rater reliabilities for each procedure were computed. The Kruskal-Wallis test was then used to determine whether the reliabilities differed across the achievement examinations. A more reliable procedure would have a higher mean ranking across the three examinations.

Cohen's Kappa (1968) was computed between the criterion examination pass/fail categorizations and the cut score procedure pass/fail categorizations to determine which procedure exhibited the highest degree of agreement with the criterion examination. When a 2 X 2 table is defined by the crosstabulation of criterion examination pass/fail classifications and the pass/fail classifications of an achievement examination, both observed and expected cell values occur. Expected cell proportions are based upon the joint probabilities of the marginal values as shown in Table 3, part of the observed proportions within each table cell would be expected.

Table 3

Cohen's Kappa Example

		Procedure Decision: Pass	Procedure Decision: Fail	
Criterion Decision	Pass	.25 (.30)	.35 (.30)	.60
	Fail	.25 (.20)	.15 (.20)	.40
		.50	.50	1.00

Within the cell defined by the intersection of *Pass, Pass* the observed proportion has a component part, which is the expected proportion. The expected proportion is computed from the row and column total proportions

that intersect the cell. In this example the expected cell proportion would be: $E_{P,P} = .60\ (.50)$

$$E_{P,P} = .30$$

The values in each cell within parentheses are the expected cell proportions. Therefore the observed proportion for the cell under consideration may have been accounted for entirely by chance agreement as the expected cell proportion is greater than the observed proportion.

Kappa was chosen because the statistic accounts for chance agreement within the classification table. Cohen (1960) states "**k** is the proportion of agreement *after* chance agreement is removed from consideration:" Due to the small number of observations the ranked Kappa coefficients were compared through the Kruskal-Wallis test.

Next, the tabled values for the proportions of false acceptances, false rejections and classifications congruent between the procedures and the criterion examination are presented. Descriptive data concerning the occurrence of false rejection and false acceptance errors and the amount of agreement between the criterion decisions and the procedure decisions provides the opportunity to examine which of the procedure(s) exhibits the fewest of both error types along with which procedure functions best when agreement with an external criterion is the only consideration.

When the Kappa coefficient is used, the assumption is made that a component of the percentage agreement between the variables under consideration is due to chance agreement. Computation of Kappa is more difficult than calculation of percentage agreement and is more subject to misinterpretation than the percentage agreement. If the percentage agreement offered a reasonable approximation of Kappa, the less complex easier understood procedure would be preferred. A comparison between the

average Kappa coefficients and the proportions of overall correct decisions was conducted to determine whether any of the standard setting procedures produced an approximation close to Kappa.

Summary

The proposed study was undertaken to provide a standard setting procedure selection model. Three steps were taken to establish the paradigm. First, it was determined if the standard setting procedures of Angoff, Nedelsky, and Ebel generated different mean right cut scores when applied to the same examination. Four introductory statistics course instructors who had instructed the course previously utilized each procedure. Standards were set for each of three examinations given during the course,

Second, the question whether the techniques exhibited different levels of inter-rater reliability was addressed. Inter-rater reliabilities were calculated for each procedure across the three examinations. The Kruskal-Wallis test was then used to determine whether the reliabilities differed across the achievement examinations.

Subsequently, it was determined which procedure agreed most in pass/fail classifications with decisions made by an established external minimal competency criterion. The Kappa coefficient for each procedure when compared to an external criterion was established for each of the three examinations. The ranked Kappa coefficients were then contrasted by the Kruskal-Wallis test to ascertain whether a difference existed in the order of Kappa rankings.

When the three procedures investigated here are used to classify examinees into pass/fail categories a possibility of two errors exists. On the

one hand, an examinee who is in reality minimally competent could be categorized into the fail category. On the other hand, an examinee not minimally competent could be classified into the pass category. Therefore, a discussion of these errors of false acceptance and false rejection is included.

To explore the suitability of the procedures when chance agreement is considered, the pass/fail decisions of the criterion and the procedures are compared. The percentages of agreement for each procedure are compared to their respective Kappa coefficients to assess which provides a tenable approximation of the coefficient. A discussion of these estimates appears to conclude the study.

CHAPTER IV

Results

During the course of the study three achievement examinations were administered to students in an introductory statistics course. The first exam was taken by 86 students, the second by 76 students and the third by 63 students. The students also were given a minimum competency examination concurrently with each achievement examination. This examinations' pass/fail decision provided the external criterion for the study. For the purposes of this study two proficiency levels of 70% and 80% were utilized for the minimum competency examination.

Four introductory statistics course instructors each used the procedures of Angoff, Ebel and Nedelsky to set standards for the three achievement examinations. Therefore, nine cut score standards were set by each instructor. The mean number right cut scores produced by the standards were compared through a two within-subjects factor repeated measures analysis of variance. Coefficient alpha was computed for each method to compare the ranked inter-rater consistency among cut scores produced by the techniques. Also, the Kappa coefficient was used to compare the pass/fail classifications of the cut scores produced by the three procedures with those classifications made by the criterion examination. This was done at both the 70% and 80% proficiency levels.

Cut Score Standards

Mean number right cut score standards set by the judges for the three achievement examinations utilizing the procedures of Angoff and Nedelsky were within three items (points). Ebel's procedure set cut scores that were greater by at least seven items than the other procedures. In fact the difference increased by one item with each examination subsequent to the first. The measures of variability for the cut scores did not exhibit any pattern of order. The lowest ranked standard deviation (smallest) for the first examination was produced by Ebel's method. For the second examination Angoff's method was ranked lowest. Nedelsky's and Ebel's techniques were virtually tied for the lowest amount of variability on the third examination. In contrast Brennan and Lockwood (1980) and Skakun and Kling (1980) reported Nedelsky's procedure exhibited more variability than Angoff and Ebel's methods. Mean number right cut score standards and standard deviations are listed in Table 4, the cut scores were rounded to the nearest whole number

The cut scores set by the instructor who constructed the achievement examination varied little from those set by the group. The greatest variation placed the instructor's cut score 1.23 items lower than the group mean cut score. For 7 out of 9 average cut scores established, the instructor's cut score differed by 1 item or less from the group mean.

Table 4

Number right minimum standards for each procedure by exam and instructor

Exam I			
		Procedure	
Instructor	Angoff	Nedelsky	Ebel
1	8.05	7.73	18.50
2	13.08	11.14	19.89
3	15.60	13.12	21.10
4	11.00	10.06	16.50
Mean	11.93	10.51	19.00
Standard deviation	3.20	2.25	1.98

Exam II			
		Procedure	
Instructor	Angoff	Nedelsky	Ebel
1	7.75	7.19	19.20
2	14.50	7.97	21.10
3	11.20	7.66	14.50
4	9.10	15.65	23.10
Mean	10.64	9.62	19.48
Standard deviation	2.94	4.03	3.68

Exam III			
		Procedure	
Instructor	Angoff	Nedelsky	Ebel
1	7.80	7.48	20.70
2	13.25	10.68	21.30
3	13.00	7.32	17.50
4	8.60	7.69	19.20
Mean	10.66	8.29	19.68
Standard deviation	2.86	1.60	1.70

Cut Score Standard Comparisons. Two within-subjects factors, examination and procedure, each with three levels were the study's independent variables. The multivariate test for no examination effect resulted in a Wilk's Criterion L statistic that suggested no statistically significant effect on the dependent variable. Wilk's Criterion for the test of no procedure effect was L= .024, the exact F(2,2)= 40.24, $p< .05$; this indicated the factor procedure was significant regarding differences in the dependent variable.

The univariate tests of the within-subjects effects had results parallel to those for the multivariate tests. No significant examination effect was observed whereas a significant procedure effect was noted. The first research hypothesis was supported as the Wilk's Criterion L and the univariate F statistics indicated different mean number right cut scores were produced by the procedures.

Contrasts were calculated to determine where the specific differences, based on the factor procedure, existed between the mean cut scores. The first contrast between the procedures of Angoff and Nedelsky yielded an F statistic indicating no difference between the procedures. The studies conducted by Harasym (1981) and Behuniak et al. (1982) produced cut scores higher for the Angoff method in two of three cases. In the third case there was no difference in the cut scores as in the present study. The second contrast between the procedures of Nedelsky and Ebel was significant with $F(1,3) = 112.65$, $p < .05$. The cut scores produced by Ebel's method were higher than those produced by Nedelsky's method. This in contrast to the results of the Andrew and Hecht (1976) and Skakun and Kling (1980) studies that revealed that Nedelsky's method produced the higher cut score. The

final contrast between the procedures of Angoff and Ebel was also significant $F(1,3) = 27.46, p < .05$. The cut scores set by Ebel's method were larger than those produced by Angoff's procedure.

Harasym (1981) observed that the Nedelsky method assigns item success probabilities (p(s)) for the minimally competent student on a discrete probability scale. A test item with 4 response choices would have probabilities of success of .33, .50 and 1.0. The Angoff and Ebel methods assign success probabilities on a continuous scale from 0 to 1.0. Based on this difference it was expected that cut scores set by Nedelsky's method would differ from those set by the other two procedures. In the present study no difference existed between cut scores produced by the procedures of Nedelsky and Angoff. There was a difference, however, between the mean cut scores produced by the Nedelsky and Ebel methods. In addition there was also a difference in the cut scores produced by the Angoff and Ebel methods.

A number of the cut scores set by the individual instructors with the procedures of Angoff and Nedelsky were close to scores that would be expected by chance. Therefore, the possibility existed that a student could be classified as minimally competent based upon a test score barely in excess of one obtained by random guessing. A cut score that could be earned by guessing would supply no information about the examinee's level of subject matter knowledge.

Each examination item on the three thirty item achievement examinations had five response choices. Assuming an equal probability of .20 for selecting each response choice at random, the chance score for each examination would be 6 correct items. Ebel's method allowed a minimum probability of success (p(s)) of .30 for each examination item. This p(s)

suggests that were an examinee guessing at each item, the examinee's expected score would be 9 items. Therefore, to compare Ebel's method to the other methods the p(s) of .30 was used as the minimum success probability.

The frequency of those items to which the instructors assigned item success probabilities for the minimally competent student of less than or equal to chance were compiled. Table 5 contains the data for the average number, and standard deviation of items with assigned success probabilities less than or equal to the .20 chance level for the methods of Angoff and Nedelsky. Comparable data for Ebel's technique is also presented for items with p(s) of .30 or less.

Table 5

Number of items with assigned success probabilities ≤ chance.

Instructor	Exam I Angoff	Procedure Nedelsky	Ebel
1	9.00	17.00	7.00
2	3.00	1.00	3.00
3	13.00	5.00	4.00
4	5.00	2.00	2.00
Mean number	7.50	6.25	4.00
Standard deviation	4.44	7.37	2.16

Instructor	Exam II Angoff	Procedure Nedelsky	Ebel
1	0.00	16.00	9.00
2	2.00	2.00	0.00
3	14.00	11.00	3.00
4	9.00	0.00	0.00
Mean number	6.25	7.25	3.00
Standard deviation	6.45	7.54	4.24

Instructor	Exam III Angoff	Procedure Nedelsky	Ebel
1	6.00	20.00	9.00
2	5.00	0.00	2.00
3	14.00	10.00	0.00
4	17.00	21.00	2.00
Mean number	10.50	12.75	3.25
Standard deviation	5.92	9.85	3.95

The mean number of items assigned success probabilities for the minimally competent student of less than or equal to chance ranged from a low of 3 for Ebel's procedure to a high of 13 for Nedelsky's technique. A two within-subjects factor multivariate analysis of variance for repeated

measures was used to determine whether the procedures differed in the number of items assigned success probabilities less than or equal to chance. The factors for the analysis were examination and procedure, each with three levels. The multivariate tests for examination and procedure indicated no effect for either factor. Therefore, the procedures were equal in the number of items assigned success probabilities less than or equal to the chance success probability.

Inter-rater Reliability Comparisons

The nine procedures were ranked in terms of reliability, with the lowest reliability coefficient receiving a rank of 1, and the highest reliability coefficient receiving rank of 9. Therefore a higher mean rank for a procedure would indicate a greater degree of reliability. Assuming a procedure received the ranks of 9, 8, and 7, the maximum mean rank possible under the current circumstances would be eight. Ebel's method produced the highest mean item rank. Nedelsky's method elicited the lowest mean item rank. The inter-rater reliabilities and their respective mean ranks appear in Table 6.

Table 6

Inter-rater reliability and rank for each procedure by examination

	Procedure		
Examination	Angoff	Nedelsky	Ebel
1	.58 (7)	.24 (4)	.59 (8)
2	.49 (6)	.20 (3)	.65 (9)
3	.16 (2)	-.51 (1)	.28 (5)
Mean rank	5.00	2.67	7.33

Overall, the reliabilities ranged from moderate to very low. The Kruskal-Wallis test was used to determine whether the procedures resulted in different ranked reliability coefficients. The comparison indicated the procedures did not differ in terms of ranked reliabilities.

Criterion Comparisons

For the purpose of classifying students as minimally competent or not, the average number right cut score for each procedure was rounded to the nearest whole number. Cohen's Kappa was computed to determine which standard setting procedure agreed most with the pass/fail decisions made by the criterion examination for each examination at both the 70% and 80% proficiency levels. Due to the small number of observations, the ranked Kappa statistics for each procedure across examinations were compared through the Kruskal-Wallis test.

Criterion-related comparisons for the procedures were based upon the Kappa statistic between the independent criterion classifications and the classifications made by each procedure. The Kappa statistics and their ranks for decisions at the 70% criterion level appear in Table 7.

Table 7
Kappa and rank for each procedure by examination
70% criterion level

	Procedure		
Examination	Angoff	Nedelsky	Ebel
1	.44 (7)	.35 (6)	.25 (4)
2	.58 (8)	.59 (9)	.33 (5)
3	.13 (2)	.16 (3)	.01 (1)
Mean rank	5.67	6.00	3.33

For the 70% criterion level Nedelsky's procedure produced a mean rank slightly higher than did Angoff's. Ebel's procedure produced the

lowest mean rank for the three examinations. The average Kappas over the three examinations were .38, .37, and .20 for the procedures of Angoff, Nedelsky and Ebel respectively. Both the Angoff and Nedelsky procedures generated average Kappas nearly twice the magnitude of Ebel's technique. Nevertheless, the Kruskal-Wallis test indicated no difference between the procedures in terms of the ranked Kappa statistics.

Table 8
Kappa and rank for each procedure by examination
80% criterion level

	Procedure		
Examination	Angoff	Nedelsky	Ebel
1	.30 (7)	.26 (6)	.40 (8)
2	.19 (1)	.21 (2.5)	.45 (9)
3	.24 (4.5)	.24 (4.5)	.21 (2.5)
Mean rank	4.17	4.33	6.50

At the 80% criterion level, as detailed in Table 8 above, Ebel's method produced the highest mean rank over the three examinations. However, as in the comparisons at the 70% criterion level, the Kruskal-Wallis test indicated no difference in the ranked Kappa statistics. The mean ranks of the Kappa statistics were close for the procedures of Angoff and Nedelsky. Ebel's method at this criterion level produced the highest mean rank. The average Kappa statistic for Ebel's procedure was greater than the average Kappas of Angoff and Nedelsky by .11 over the three examinations.

A question remained how well the procedures functioned when correct decisions regardless of whether based on chance or not were considered. Therefore, the amount of agreement and disagreement between the criterion and procedure pass/fail decisions was examined. Also the

nature of the pass/fail decision errors associated with each procedure were determined at both the 70% and 80% criterion levels.

70% Criterion. Table 9 contains a summary of correct and incorrect classification proportions at the 70% proficiency level for the first achievement examination. The correct decision sums appear in the corner of the bottom right quadrant of the 2 X 2 tables. The incorrect proportion sums appear in the corner of the bottom left quadrant of the tables.

Table 9

70% Proficiency level criterion and procedure decisions

Procedure

Exam 1 (n= 86)

Criterion Decision	Angoff Pass	Angoff Fail	Nedelsky Pass	Nedelsky Fail	Ebel Pass	Ebel Fail
Pass	64 / .74	2 / .04	65 / .76	3 / .03	38 / .44	30 / .35
Fail	10 / .12	10 / .12	11 / .13	7 / .08	4 / .04	15 / .17
	.16	.84	.16	.84	.39	.61

Exam 2 (n= 76)

Criterion Decision	Angoff Pass	Angoff Fail	Nedelsky Pass	Nedelsky Fail	Ebel Pass	Ebel Fail
Pass	61 / .80	2 / .03	63 / .83	0 / .00	37 / .49	26 / .34
Fail	6 / .08	7 / .09	7 / .09	6 / .08	0 / .00	13 / .17
	.11	.89	.09	.91	.34	.66

Exam 3 (n= 63)

Criterion Decision	Angoff Pass	Angoff Fail	Nedelsky Pass	Nedelsky Fail	Ebel Pass	Ebel Fail
Pass	51 / .81	11 / .17	60 / .95	2 / .03	12 / .19	50 / .79
Fail	0 / .00	1 / .02	0 / .00	1 / .02	0 / .00	1 / .02
	.17	.83	.03	.97	.79	.21

For the three examinations, Angoff's method had an average proportion correct classification of .85, Nedelsky's, .91, and Ebel's, .49. Standard deviations for the methods were .03, .07 and .25 respectively. At this proficiency level the average proportion correct for the Nedelsky method was ranked highest. Angoff's method which ranked second was within .06 points of Nedelsky's method. Ebel's method which ranked third with a correct proportion of .49 was much lower than the second ranked procedure of Angoff. The magnitude of the difference in standard deviations among the three procedures was also immense. The standard deviation for Ebel's method was four times greater than the second largest standard deviation. This was partially based upon the large decrease in the proportion correct for Ebel's method on the third examination. The least amount of variation was found in Angoff's method. Compared to Angoff's procedure, Nedelsky's method was less than .04 points greater in variability.

80% Criterion. On the average, at the 80% proficiency level correct decisions were .68 for Angoff's method .70 for Nedelsky's and .65 for Ebel's. Angoff's procedure produced a standard deviation of .038 and Nedelsky's .032. In contrast Ebel's procedure produced a .114 standard deviation. Table 10 below includes the classification cross-tabulations for the 80% criterion level.

Table 10

80% Proficiency level criterion and procedure decisions

Procedure

Criterion Decision

		Angoff		Nedelsky		Ebel		
		Pass	Fail	Pass	Fail	Pass	Fail	
Pass		51 .59	2 .04	52 .60	1 .01	34 .40	19 .22	Exam 1
Fail		23 .27	10 .12	24 .28	9 .11	7 .08	26 .30	n= 86
		.29	.71	.29	.71	.30	.70	

		Angoff		Nedelsky		Ebel		
		Pass	Fail	Pass	Fail	Pass	Fail	
Pass		42 .55	2 .03	44 .58	0 .00	30 .40	14 .18	Exam 2
Fail		25 .33	7 .09	26 .34	6 .08	7 .09	25 .33	n= 76
		.36	.64	.34	.66	.27	.73	

		Angoff		Nedelsky		Ebel		
		Pass	Fail	Pass	Fail	Pass	Fail	
Pass		37 .59	5 .08	42 .67	0 .00	12 .19	30 .48	Exam 3
Fail		14 .22	7 .11	18 .28	3 .05	0 .00	21 .33	n= 63
		.30	.70	.28	.72	.48	.52	

With the 80% proficiency level the procedures were close in the average proportions of correct classification and differed by no more than .05 points. When the parallel decisions made at the 70% level were compared to those at the 80% level it was found that Ebel's method increased by .16 points, whereas Angoff's and Nedelsky's procedures decreased by .17 and .21 respectively. The standard deviations for the Angoff and Nedelsky methods were very close; in contrast greater variability was produced by

Ebel's technique. Ebel's technique was almost four times more variable than the other techniques.

When the minimum competency criterion was at 80%, the standard setting procedures performed equally well in terms of overall correct decisions. At the 70% level the procedures of Angoff and Nedelsky performed better than the procedure of Ebel. The procedures seem to converge in terms of effectiveness when the proficiency level of the criterion examination increases. At both proficiency levels Ebel's method appeared to be more variable.

False Acceptances and Rejections The classification errors for each procedure by examination at both the 70 % and 80 % criterion levels appear above in Tables 8 and 9 respectively. As shown in Table 11 below, for each technique the cell located at the intersection of the criterion decision of *pass* and the procedure decision of *fail* contains the proportion of false rejection classification errors. In this example 35% of the classifications were errors of false rejection. Conversely, errors of false acceptance (.25) are located in the cell defined by a criterion decision of *fail* and a procedure classification of *pass*.

Table 11

False Acceptance False Rejection Example

Criterion Decision	Procedure Decision: Pass	Procedure Decision: Fail	
Pass	.25	.35	.60
Fail	.25	.15	.40
	.50	.50	1.00

Examination of the errors will be based first on the assumption that false acceptances are the more serious error. Then examination of the data assuming false rejections are the more serious error will be presented. Average proportions of decision errors for the procedures at the two proficiency levels are listed in Table 12.

Table 12

Average classification errors 70% and 80% criterion levels

Criterion level		Procedure		
		Angoff	Nedelsky	Ebel
70%	False acceptance	.07	.07	.01
	False rejection	.08	.02	.49
80%	False acceptance	.27	.30	.06
	False rejection	.05	.02	.29

When false acceptances are considered to be more serious than false rejections the procedures of Angoff and Nedelsky do not appear to perform as well as Ebel's method at either criterion level. The average errors of false acceptance were approximately five times as large for the methods of Angoff and Nedelsky. Thus, Ebel's method would be preferred to set the cut score standard in this situation.

In situations where rejecting a minimally competent individual would be the more serious error, the procedures of Angoff and Nedelsky would be preferred over Ebel's technique. The proportions of false rejections were lower at both criterion levels for the techniques of Angoff and Nedelsky. Nedelsky's method exhibited the least amount of error at the two criterion levels.

Kappa Estimation When the Kappa coefficient is used, the assumption is made that a component of the agreement between the variables under consideration is due to chance agreement. Computation of Kappa while not complex, is more difficult than calculation of percentage agreement. In addition, the coefficient is more subject to misinterpretation than the percentage agreement. Therefore, were it the case that the percentage agreement offered a reasonable approximation of Kappa, the less complex easier understood procedure would be preferred. A comparison between the average Kappa coefficients and the proportions of overall correct decisions at both criterion levels was conducted to determine whether any of the standard setting procedures produced an approximation close to Kappa.

To judge the how well the Kappa coefficient was approximated, the ratio of the average Kappa coefficient to the average correct classification proportion was computed. The size of the computed ratio does not serve to

indicate whether the Kappa coefficient or the correct classification percentage was high or low. For example the ratio .20/.20 would be 1, whereas the ratio .70/1.00 would be .70 and lower compared to the first ratio. In this example the Kappa for the second ratio is much higher than the first, but the first ratio shows the approximation of Kappa to be more accurate. The average Kappas, correct classification proportions and the Kappa/Correct classification ratios appear in Table 13 that follows.

Table 13

Average Kappa and correct classifications 70% and 80% criterion levels

Criterion level		Procedure		
		Angoff	Nedelsky	Ebel
70%	Average Kappa	.38	.37	.20
	Avg. correct clas.	.85	.91	.49
	Kappa/Correct ratio	.66	.41	.41
80%	Average Kappa	.24	.24	.35
	Avg. correct clas.	.68	.70	.65
	Kappa/Correct ratio	.35	.34	.54

The ratios exhibit a differential effect based upon criterion level. At the 70% criterion level, Angoff's method generates the largest ratio. The largest ratio at the 80% criterion level was produced by Ebel's procedure. These results would imply that under a stringent criterion the correct agreement proportion could be used in lieu of Kappa to evaluate the effectiveness of Ebel's method when compared to an external criterion.

Conversely, when the criterion proficiency level is low, the most reasonable approximation of Kappa was generated by Angoff's procedure. Overall, the Kappa/Correct ratios suggest the average correct classification percentages do not furnish an accurate estimate of the average Kappa coefficients.

CHAPTER V

Summary

The first purpose of the study was to determine whether the standard setting procedures of Angoff, Nedelsky and Ebel would produce different mean cut scores when applied to the same subject matter. The second purpose was to ascertain if the procedures differed in rater reliability. The final goal was to determine which procedure made the most correct pass/fail classifications beyond chance agreement when compared to an external criterion.

The steps taken to accomplish the stated purposes of this study were as follows. Three achievement examinations were administered to students in an introductory statistics course. The students also were given a minimum competency examination along with each achievement examination. For the purposes of this study proficiency levels of 70% and 80% were set for the minimum competency examination.

Four introductory statistics course instructors each used the procedures of Angoff, Ebel and Nedelsky to set standards for the three achievement examinations. Therefore, nine cut score standards were set by each instructor. The mean number right cut scores produced by the standards were compared with a two within-subjects factor repeated measures analysis of variance. Coefficient alpha was computed as a measure of the inter-rater reliability of the cut score procedures. Also, Kappa statistics were computed as a measure of proportion of agreement

beyond chance between the pass/fail decisions of the procedures and the criterion examinations at both the 70% and 80% proficiency levels. Finally, the pass/fail classifications of the cut scores produced by the three procedures were compared to those classifications made by the criterion examination at both proficiency levels.

When the mean number right cut scores were compared, Ebel's standard setting procedure yielded higher cut scores than did the procedures of Angoff and Nedelsky. There was no statistically significant difference in the mean number right cut scores produced by the approaches of Angoff and Nedelsky. Cut scores produced by the Nedelsky procedure were slightly higher than those produced by Angoff's method.

Regarding the inter-rater reliabilities, the Kruskal-Wallis test indicated no difference between the cut score setting procedure reliabilities across examinations. Within examinations, Ebel's method was ranked highest in each case. On the first examination, Ebel's procedure had the highest reliability coefficient followed closely by Angoff's technique. For the second examination, the reliability of Ebel's procedure increased while a decrease was noted for the Angoff and Nedelsky techniques. A decrease was observed in the reliabilities of all three techniques on the third examination; all reliabilities were low.

The Kappa statistic was calculated between the pass/fail decisions of the procedures and the external criterion for each of the three achievement examinations. The external criterion was used at proficiency levels of both 70% and 80%. The differences between the Kappa statistics and their respective ranks were studied at the level of the individual achievement examinations. Kappa statistics were also examined as averages over the three achievement examinations.

At the 70% proficiency level the overall Kappa coefficients were moderate to low. They ranged from a high of .59 to a low of .01. Angoff's method was ranked highest for the first examination. Nedelsky's procedure had the highest Kappa statistic for the second and third examinations. Although ranked higher for the second examination, Nedelsky's method produced a Kappa statistic that only differed by .01 compared to that obtained with Angoff's procedure. For the third examination, the difference between the first and second ranked Kappas of Nedelsky and Angoff was .03. However, the Kappas produced by Ebel's method ranked lowest for each examination. Nedelsky's method, ranked second for the first examination and was .10 higher than Ebel's. For the second and third examinations Angoff's method was ranked second and larger by .25 and .12 for the respective examinations.

The procedures were also ranked from highest to lowest without regard to examination. The lowest Kappa was assigned the rank 1 while 9 was the rank assigned to the highest Kappa coefficient. The mean ranks for Angoff and Nedelsky's techniques were very similar and both were larger than Ebel's. The mean ranks for the techniques of Angoff and Nedelsky were almost twice the magnitude of Ebel's.

When the proficiency level was increased to 80%, Ebel's method resulted in the highest Kappa statistic on two out of three achievement examinations. For the first examination at the 80% proficiency level, Ebel's procedure was greater than Angoff's second ranked procedure of by a .10 margin. For the second examination, the difference between Ebel's procedure and the second ranked method of Nedelsky was greater a margin of .24 separated Ebel's procedure from the second ranked procedure of Nedelsky. For the third examination in contrast, the Kappa statistics of

Angoff and Nedelsky were both .24 whereas Kappa for Ebel's procedure was .21.

In the 70% criterion situation when the Kappa coefficients were ranked over all examinations, the mean ranks of Angoff and Nedelsky's procedures were almost equal. At the 80% criterion, Ebel's method produced a mean rank higher than the other procedures. However the Kruskal-Wallis test suggested no difference between the mean ranks for the three procedures at either criterion level.

With respect to pass/fail decisions in agreement beyond chance with an external criterion examination, the procedures were incongruous at the different criterion levels. For the 70% criterion level the Kappa coefficient of Nedelsky's procedure was highest for two of the three examinations. However for the second and third examinations the amount of agreement was less than .05 higher than the next highest procedure. Over the three examinations the average Kappas were moderate to low. The average Kappas were .38, .37, and .20 for the procedures of Angoff, Nedelsky and Ebel respectively.

At the 80% criterion level Ebel's method produced the largest Kappa for examinations one and two. The three procedures did not differ much for the third examination with .03 being the largest difference between the procedures. As was the case with the 70% criterion level, the average Kappas were low to moderate. Average Kappa coefficients for the three examinations were .24 for the procedures of both Angoff and Nedelsky. Ebel's method at this criterion level produced an average Kappa of .35.

The change in the Kappa statistics for the first examination when the proficiency level of the criterion was increased indicated a moderate decrease for Angoff and Nedelsky's methods, in contrast to an increase of

similar proportions for Ebel's technique. Nevertheless, there was a large decrease for the second examination in the Kappas produced by Angoff and Nedelsky's techniques compared to an increase of magnitude similar to the first examination for Ebel's procedure. The third examination 80% proficiency level Kappa coefficients all exhibited increases over the 70% proficiency level Kappas for examination three. The increment for Ebel's process was twice that of the second ranked procedure of Angoff. The procedures of Angoff and Nedelsky exhibited agreement beyond chance rates at the 70% proficiency level that were high when compared to Ebel's method. At the 80% proficiency level the differences were less.

When the amount of agreement between the pass/fail decisions of the standard setting procedures and the criterion examination with chance agreement *included* were compared the results were similar to those for the Kappa comparisons. The average agreement over the three examinations for the 70% proficiency level was computed for each procedure. The procedures of Angoff and Nedelsky were in agreement with the decisions made by the external criterion test at the respective rates of .85 and .91 versus .49 for Ebel's method. When the criterion level was set at 80%, a decline was noted in the correct average classification decisions for the Angoff and Nedelsky procedures and an increase in the correct average proportion for Ebel's method. The averages over three examinations at the 80% proficiency level were .70, .68 and .65 for the respective procedures of Nedelsky, Angoff and Ebel.

Conclusions

Results of the present study partially paralleled those of previous research. Prior research has shown that the procedures produced different mean number right cut scores when applied to the same subject matter.

The studies conducted by Andrew & Hecht, (1976) and Skakun & Kling (1980) that involved the procedures of Nedelsky and Ebel resulted in higher cut scores set by Ebel's method. This was also the case for the present study. Research conducted by Harasym (1981) compared a modified version of Angoff's procedure with Nedelsky's method and found Nedelsky's technique set a greater cut score standard. Behuniak et al (1982) compared the methods of Nedelsky and Angoff and found that Angoff's method set a higher cut score standard. In contrast to both studies, the present study showed no statistically significant differences between the cut score standards established by the procedures of Angoff and Nedelsky.

Under the conditions outlined for the study, the following conclusions were reached:

1) The standard setting procedures of Angoff, Ebel and Nedelsky produce different mean number right cut scores when applied to common subject matter.

2) The standard setting procedures of Angoff, Ebel and Nedelsky do not differ in the ranked inter-rater reliability of judges' ratings.

3) When the pass/fail decisions of the procedures were compared to those of an external criterion at both the 70% and 80% proficiency criterion levels the ranked Kappa coefficients did not differ.

Limitations

The results of the study should be viewed with regard to the limitations that follow. First, with such a small group of judges, more variability in the mean cut score averages was expected. Second, the

minimum competency criterion examinations were not entirely independent. The results of this study apply to a graduate level statistics course. However, different results could occur with different subject matter areas and or subjects. One of the instructors constructed the achievement examination independent of the other instructors. This could have biased the selection of the achievement examination item content. Students who only wished to pass the course may have only performed well enough to pass the minimum competency examination and not as well as possible on the achievement examination.

Recommendations

A viable paradigm has been established that will facilitate the selection of an absolute standard setting procedure for achievement examinations. The paradigm consists of measuring the degree of association between the correct pass/fail decisions of the three standard setting procedures and the pass/fail decisions of an independent criterion. In addition, it was determined that the cut score procedures did produce mean number right cut score standards that had statistically significant differences. Also, the cut score standards did exhibit differences in inter-rater reliability.

There are a number of issues that should be addressed in future research. Research should focus upon a larger scale application of the standard setting selection process. Less variable estimates of the number right cut scores set by each procedure would be obtained with a larger number of instructors. The cut score estimates also may be more precise if the judges received more structured training in the standard setting procedures. Also, a group consensus approach to constructing the

achievement examination should be conducted to assure that examination item content not contain instructor bias.

Results of this study suggest that standard setting procedures do not function differently at the two established levels of minimum competency. Consideration should be given to studying how well the procedures function over a wider range of criterion proficiency levels. It could be determined at what point in the range of minimum competency the standard setting procedures diverge in term of effectiveness. Changes in effectiveness based on movement of the criterion level suggests the selection of the standard setting procedure best for a particular situation.

It would be informative to use a modified version of Ebel's procedure in which only the relevance/difficulty categories of essential/easy, important/easy and important/medium are used. Those categories encompass items that a minimally competent student would be expected to answer correctly. The remaining categories should include items that would measure more than minimal competence, therefore the inclusion of the other categories may inflate the number right cut score produced by this procedure.

Further consideration should be given to other characteristics of the standard setting procedures important in procedure selection. Such characteristics include time to learn each procedure, relative ease of use, and time necessary to implement each method. The influence of different types of subject matter on the effectiveness of the standard setting procedures also warrants investigation.

GENERAL INSTRUCTIONS

For test 1 you should have the following materials:

1. A scored copy of the test
2. An instruction sheet for each procedure
3. A response form for each procedure
4. Scratch paper
5. A pencil
6. An envelope to return the materials.

If any of these materials are not in your packet, please contact Ira Washington immediately.

Please complete the following steps for this packet of materials:

1. Read and follow carefully the instructions for each procedure.
2. Record your judgments on the response form provided.
3. PLEASE COMPLETE THE PROCEDURES IN THE FOLLOWING ORDER:
   1.
   2.
   3.
4. Place the examination copy and response forms into the envelope provided.
5. Please leave the envelope with your secretary to be picked up by Ira Washington.
6. Retain the instruction sheets for each procedure for use with the package of examination materials you will receive in approximately one week.

Please note:

1. If at all possible, the judgments for this packet need to be completed within one week.

2. If there are any questions, please call.

3. Next week you will receive a different examination and will be asked to complete the same procedures but in a different prespecified order.

PROCEDURE A

INSTRUCTIONS

1. Think of several students from your past CEP 904 (ED 869) Quantitative Methods in Educational research class(es) who barely received a 2.0 grade. Use these students as models when judging each test item.

2. For each test item, estimate the probability (a decimal between 0 and 1 inclusive) that the lowest 2.0 student would answer that item correctly. When making the estimate, consider the item quality, the number of skills necessary to answer correctly, and the homogeneity of the distracters.

3. Write the probability you estimated in step 2 next to the corresponding item number on the Response Form for Procedure A.

4. Repeat steps 2 and 3 for each of the 30 items on the test.

Sample Item and Response Form

The sample item and response form shown below illustrate Procedure A.

Item

1. Which of the following is a measure of central tendency?
   *A. Mean
   B. Maximum
   C. Medium
   D. Most

Response Form

1. .50

2. ____

3. ____

To answer the question correctly, the lowest 2.0 student would have to know the exact spelling of the three central tendency measures. Students possessing partial knowledge are likely to be confused by the "sound alike" distracters. However, the item requires simple recall and therefore should be easy to answer correctly. You can see the instructor marking this response form estimated the lowest 2.0 student would have a probability of .50 of answering the item correctly.

Note:

If a student were guessing randomly, the student would have a 25% (1/4) chance of selecting the correct answer. However, if you felt one or more of the distracters would be especially attractive to the lowest 2.0 student, you may estimate the probability of correct response at less than .25.

Due to the straight forward nature of writing the probabilities on the Response Form, a practice item has been omitted. However, if the procedure seems at all unclear, please do not hesitate to ask questions for clarification.

APPENDIX

PARTICIPANT INSTRUCTIONS

PROCEDURE A

RESPONSE FORM

Directions

In the space next to the item number, write your estimate of the probability that the lowest 2.0 student will answer the item correctly.

1. _____

2. _____

3. _____

4. _____

5. _____

6. _____

7. _____

8. _____

9. _____

10. _____

11. _____

12. _____

13. _____

14. _____

15. _____

16. _____

17. _____

18. _____

19. _____

20. _____

21. _____

22. _____

23. _____

24. _____

25. _____

26. _____

27. _____

28. _____

29. _____

30. _____

PROCEDURE B

INSTRUCTIONS

1. Think of several students from your past CEP 904 (ED 869) Quantitative Methods in Educational research class(es) who barely received a 2.0 grade. Use these students as models when judging each test item.

2. For each test item, identify the alternative(s) the lowest 2.0 students would reject as incorrect. The responses you identify are the one to which students receiving grades below 2.0 would be attracted.

3. On the Procedure B Response Form, find the item number corresponding to the one on which you are working. Letters to the right of that item represent response alternatives for the item. For each response alternative identified in step 2, cross out the corresponding letter next to that item on the response form.

4. Repeat steps 2 and 3 for each of the thirty items on the test.

Sample Item and Response Form

The sample item and response form below illustrate Procedure B.

Item

1. Which of the following is a measure of central tendency?
   *A. Mean
   B. Maximum
   C. Medium
   D. Most

Response Form

1. A ~~B~~ C ~~D~~

2. A B C D

3. A B C D

For the example above the instructor felt the lowest 2.0 student would know responses "B" and "D" were incorrect so the "B" and "D" next to item 1 on the response form were crossed out.

If the instructor estimated the lowest 2.0 student would identify all three incorrect alternatives, the letters **B, C, and D** would have all been crossed out. Had the instructor felt the lowest 2.0 student would only recognize alternative **D** as incorrect, only the letter **D** would have been crossed out on the response form next to the appropriate item number. If the instructor felt such a student could not identify any of the alternatives as definitely incorrect, no letters would have been crossed out for that item.

Practice Item

2. What is the maximum percentage of wolf ancestry permitted for a pet in the state of Michigan?

*A. 93

B. 83

C. 73

D. 63

Directions

Mark the Practice Item Response Form below assuming that the lowest 2.0 student would have known any alternative less than 73% was incorrect.

Practice Item Response Form

1. A B C D

2 A B C D

3. A B C D

4. A B C D

Go to the next page for the correct answer.

Practice Item Answer

Your response form should be marked as follows:

1. A B C D

2 A B ~~C~~ ~~D~~

3. A B C D

4. A B C D

Response choices C and D next to item number 2 were crossed out because both were less than 75%.

PROCEDURE B

RESPONSE FORM

Directions

For each test item, cross out the alternative(s) you believe the lowest 2.0 student would recognize as incorrect.

1.	A	B	C	D	E
2.	A	B	C	D	E
3.	A	B	C	D	E
4.	A	B	C	D	E
5.	A	B	C	D	E
6.	A	B	C	D	E
7.	A	B	C	D	E
8.	A	B	C	D	E
9.	A	B	C	D	E
10.	A	B	C	D	E
11.	A	B	C	D	E
12.	A	B	C	D	E
13.	A	B	C	D	E
14.	A	B	C	D	E
15.	A	B	C	D	E
16.	A	B	C	D	E
17.	A	B	C	D	E
18.	A	B	C	D	E
19.	A	B	C	D	E
20.	A	B	C	D	E
21.	A	B	C	D	E
22.	A	B	C	D	E
23.	A	B	C	D	E
24.	A	B	C	D	E
25.	A	B	C	D	E
26.	A	B	C	D	E

27.	A	B	C	D	E
28.	A	B	C	D	E
29.	A	B	C	D	E
30.	A	B	C	D	E

PROCEDURE C

INSTRUCTIONS

1. Think of several students from your past CEP 904 (ED 869) Quantitative Methods in Educational research class(es) who barely received a 2.0 grade. Use these students as models when judging each test item.

2. Classify each item as easy, medium, or hard relative to the ability of the lowest 2.0 student to answer the item.

3. Then classify each item as essential, important, acceptable or questionable, relative to the importance of mastery of the item for the lowest 2.0 student.

4. On the matrix response form for Procedure C, write the item number in the cell located at the intersection of the difficulty and relevance categories which you selected for the item in steps 2 and 3.

   Note: The cells medium essential, hard essential, and hard important are not to be used.

5. Repeat steps 2 through 4 for each of the 30 test items.

Sample Item and Response Form Matrix

The sample item and response form matrix shown below illustrate Procedure C.

Practice Item

1. What is the maximum percentage of wolf ancestry permitted for a pet in the state of Michigan?

   * A. 93

   B. 83

   C. 73

   D. 63

Response Form Matrix

Relevance	Difficulty: Easy	Medium	Hard
Essential			
Important			
Acceptable			
Questionable	1		

In the above example, the instructor for a hypothetical biology class classified the item as easy questionable. Simple recall was required to answer the item correctly, and the instructor felt the item was trivial and lacked relevance. The instructor's classification is shown by the item number **1** written in the cell located at the intersection of easy difficulty and questionable relevance on the matrix response form.

Practice Item

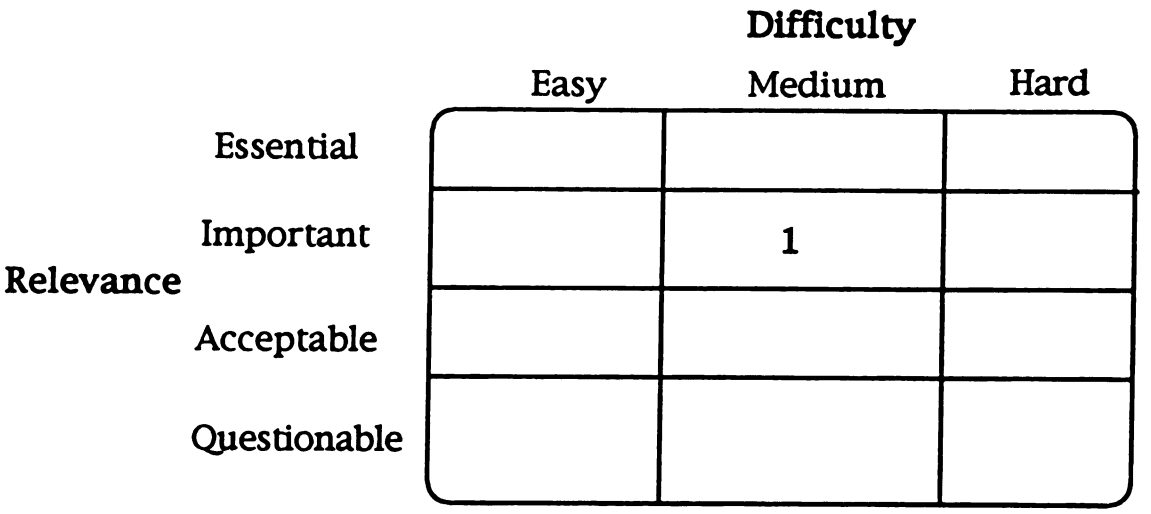
1. Which of the following is a measure of central tendency?
   *A. Mean
   B. Maximum
   C. Medium
   D. Most

Directions.
Classify the item assuming the content is important for the lowest 2.0 statistics student, and that it is of medium difficulty due to the similarity of the distracters to the correct response alternative.

Practice Response Form Matrix

Relevance	Difficulty: Easy	Medium	Hard
Essential			
Important			
Acceptable			
Questionable			

Go to the next page for the correct answer.


Difficulty
Easy
Medium
Hard
Essential
Important
Acceptable
Questionable
Relevance
1

LIST OF REFERENCES

LIST OF REFERENCES

Allen, M. J. and Yen, W. M. (1979). Introduction to measurement theory. Monterey, Calif: Brooks/Cole Publishing Co.

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

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

Behuniak, P., Archambault, F., and Gable, G. (1982) Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational and Psychological Measurement, 42, 247-255.

Berk, R. A., (1986) A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

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

Ebel, R. L. (1972) Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice Hall.

Glasnapp, D. R., Poggio, J. P. & Eros, D. S. (1983). An analysis of the validity of judgmental methods used to set test standards. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Glass, G. V. Standards and criteria. (1978). Journal of Educational Measurement, 15, 237-261.

Gross, L. J. (1982). Standards and criteria: A response to Glass' criticism of the Nedelsky technique. Journal of Educational Measurement, 19, 59-162.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in the instructional setting. Journal of Educational Measurement, 15, 277-289.

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard setting procedures on evaluation outcome. Educational and Psychological Measurement, 41, 725-734.

Mehrens, W. and Lehmann, I. (1984). Measurement and Evaluation in Education and Psychology. (3 rd. ed.) New York: Holt Rinehart and Winston Inc.

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

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

Saunders, J. C., Ryan, J. P. and Huynh, H. A. (1980). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 5, 209-217.

Scriven, M. (1978). How to anchor standards. Journal of Educational Measurement, 15, 273-275.

Skakun, E. N., and Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.