DOCUMENT RESUME

ED 198 156 TM 810 089

AUTHOR	Rock, D. A.; And Others
TITLE	An Empirical Comparison of Judgmental Approaches to Standard Setting Procedures.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-80-7
PUB DATE	Jun 80
NOTE	25p.
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Competence; *Content Analysis; Criterion Referenced Tests; *Cutting Scores; *Judges; Probability; Real Estate Occupations; *Scoring; State Licensing Boards; *Test Items
IDENTIFIERS	*Angoff Methods; Interrater Reliability; *Nedelsky Method

ABSTRACT

An experiment was designed that varied cutting score procedures, instructions, and types of judges in order to address the following questions concerning the Real Estate Licensing Examination: (1) Will the cutting score levels produced by groups of judges from differing backgrounds (academicians vs. practitioners vs. lawyers) using the same method and instructions be different? (2) Will the agreement between item rating profiles vary across these different groups of judges? (3) Does either agreement across items and/or levels vary systematically by instruction/method? It was found that three out of four groups of judges arrived at significantly higher cutting score levels using the Angoff method than when using the Nedelsky procedure. The Angoff procedure was more effective in setting standards that distinguished the minimally qualified practitioner from the individual with average qualifications. Although the Angoff method demonstrated somewhat higher interjudge agreement with respect to the patterns of item responses the average correlation between item profiles was generally low for both procedures. (Author/GK)

RR-80-7

RESEARCH REPORT

An Empirical Comparison of Judgmental Approaches to Standard Setting Procedures

D. A. Rock, E. L. Davis and C. Werts

ETS

Educational Testing Service
Princeton, New Jersey
June 1980

An Empirical Comparison of Judgmental Approaches to Standard Setting Procedures

D. A. Rock, E. L. Davis and C. Werts

Educational Testing Service
Princeton, New Jersey
June 1980

ABSTRACT

An experiment was designed that varied cutting score procedures, instructions, and types of judges in order to address the following questions: (1) Will the cutting score levels produced by groups of judges from differing backgrounds using the same method and instructions be different? (2) Will the agreement between item rating profiles vary across these different groups of judges? (3) Does either agreement across items and/or levels vary systematically by instruction or method?

It was found that three out of four groups of judges arrived at significantly higher cutting score levels using the Angoff method than when using the Nedelsky procedure. The Angoff procedure was more effective in setting standards that distinguished the minimally qualified practitioner from the individual with average qualifications. Although the Angoff method demonstrated somewhat higher interjudge agreement with respect to the patterns of item responses the average correlation between item profiles was generally low for both procedures.

An Empirical Comparison of Judgmental Approaches to Standard Setting Procedures

INTRODUCTION

The interpretation of test scores with respect to absolute standards rather than to the performance of others has become an increasi.gly important practice with the advent of evaluation concepts such as minimal competence and mastery-non-mastery. The literature refers to testing decisions which relate an individual's performance to that of others in the same population of test takers as norm-referenced testing while testing decisions which relate test performance to absolute standards are freauently referred to as criterion-referenced testing.

While there is much discussion of criterion referenced testing per se in the literature (Anastasi, 1976; Millman, 1974; Popham and Husek, 1969), there is very little information on how to set cutting scores in criterion-referenced situations. As occupational licensing and certification become more widespread, the develooment of systematic and professionally defensible methods of setting cutting scores becomes a necessity.

Ebel (1972) discusses a compensatory item orobability method that leads to a single oassing score. The items of a test are classified into a two-way grid with judged item importance and item difficulty as the dimensions. A further judgment is made of the proportion of items in each cell of the grid that must be passed by a "minimally qualified barely passing" examinee. For each cell, this proportion and the number of items on the test placed into that cell are multiplied together. The sum of these products (accumulated over all cells) is the number of items that must be answered correctly if the test is to be passed.

Angoff (1971) gives the following item probability method: "...ask each judge to state the probability that the minimally acceptable person would answer each item correctly. In effect, the judge would think of a number of minimally acceptable persons instead of only one such person who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score." (p. 515).

A variant to this probabilistic approach was described more than 20 years ago by Nedelsky (1954). A passing score for multiple-choice items is constructed as follows: For each item, judges identify those distractors that the barely passing individual should be able to eliminate. The reciprocal of the number of remaining options (including the keyed choice) is calculated for that item. Thus, for a five-choice item in which two distractors were judged to be the ones that even a barely passing student would not choose, the reciprocal is 1/3 or .33. Assuming that the test is scored one point for each correct answer a "guessing score" is the sum of these reciprocals computed for all of the items in the test. This "guessing score" can be considered the cutting score that discriminates the minimally qualified from the non-qualified. Obviously the so-called "guess score" is not purely a guess score since its estimation is based on partial knowledge and in general will lead to a cutting score substantially above what one would arrive at using the traditional "guessing" formula.

Andrew and Hecht (1976) in an empicial study compared Ebel's procedure with Nedelsky's. Specifically, the study was designed to determine (a) whether the cutting score levels for comparable samples of items would vary depending upon the standard setting procedure used to establish this level and (b) whether for each of the two standard setting procedures the cutting score for a sample of test items would vary depending upon the group of judges used. They found that within each of the methods there was relatively high agreement among the groups of judges with respect to cutting score levels. That is, both methods lead to consistent estimates of a cutting score. There were, however, considerable differences between the methods on the absolute value of the cutting score. They found that the Nedelsky method led to a significantly lower cutting score than did the Ebel method.

The principles of generalizability theory were used by Brennan and Lockwood (1979) in their study comparing the Angoff and Nedelsky methods of setting cutting scores. They discovered greater variability over items in the probabilities generated by the Nedelsky proceiure. Also the intrarater variation was somewhat higher for the Nedelsky method, while the average cutting scores produced were lower. Brennan and Lockwood also examined the specific alternatives chosen by raters for the Nedelsky method and discovered that while raters might agree on the number of item distractors to eliminate, they might not agree on the specific distractors.

None of the above studies systematically evaluated the invariance of the cutting score levels across groups of judges whose backgrounds vary. That is, which if any of the methods yields a consistent cutting score level across populations of judges who are characterized by diverse sources of knowledge (e.g. academicians vs. practitioners vs. lawyers). If one or more of the methods is relatively invariant with respect to generalizability of results across different populations of experts, then the "knotty" question of who are the most appropriate groups to make cutting score judgments becomes less critical.

Another important question on which there is little or no research information is the relative sensitivity or discriminability of the methods to variations in instructions. That is, if judges were asked to evaluate items with respect to both minimally acceptable persons and persons possessing average qualifications, the resulting two cutting scores should be well-defined with minimum overlap. That is, other things being equal a preferred method would lead to cutting scores which would distinguish the minimally qualified from individuals with average qualifications.

In an effort to answer sote of these questions an experiment was designed that varied methods, instructions, and types of judges. Angoff and Nedelsky methods were compared with respect to cutting score levels obtained under instructions having to do with minimally competent individuals as well as persons with average competence. Four groups of judges characterized by three different types of backgrounds were employed. More specifically, the research addressed the following ouestions.

1. Does either method produce systematically higher cutting scores than the other?
2. Do any groups of judges systematically set higher cutting scores than the others?
3. How do score judgement for the minimally competent examinee differ from judgement for an examinee with average qualifications?
4. How do score judgments for the examinee with average qualifications compare to empirical estimates of mean scores based on pre-test item data?

PROCEDURE

SUBJECTS:

The sixteen judges in the present study were members of four standing committees used by Educational Testing Service as test question reviewers for the Real Estate Licensing Examination. The Minority and Sex Bias Review Committee, four judges, consiscs of real estate commissioners and administrative officers of real estate commissions. They serve as licensirg officers and in some cases are also practicing brokers. The Practicing Broker Review Committee, three judges, includes practicing brokers who also are state real estate commissioners. The Legal Review Committee's, four participants are either assistant attorneys general involved with real estate commissions or act as legal counsel to a real estate commission. The remaining group of five judges are from the Consultant Review Committee which is composed of professors of real estate courses from major universities and who also have served as item writers. It was thought that these four panels of judges represented a broad knowledge of the competencies of a real estate salesperson and yet represented both the academic viewpoint as well as that of the practicing brokers.

DESCRIPTION OF TASKS

In order to examine both the Angoff and the Nedelsky method for determing cutting scores for salespersons with minimal and average competence, four sets of instruction summarized below, were developed:

INSTRUCTION #1

> Under this task your judgments about the test questions are to be made with reference to your conception of a minimally knowledgeable salesperson. You will judge what percentage of the salespersons in this minimally knowledgeable group would know the answer to each question and then mark on an accompanying coded answer sheet the percentage that comes closest to your judgment.

INSTRUCTION #2

> Under this task your judgments about the test questions are to be made with reference to your conception of a practicing salesperson of average knowledge. You will judge what percentage of the salespersons in this average knowledge group would know the answers to each question and mark on an accompanying coded answer sheet the percentage that comes closest to your judgment.

INSTRUCTION #3

For this task, you will inspect each item distractor and identify those distractors which the minimally knowledgeable salesperson should be able to eliminate. That is, you will identify those distractors which a minimally knowledgeable salesperson would recognize as being obviously wrong. On an extremely easy item, this might be all the distractor options (leaving only the keyed option). On a very difficult item, you may feel that a minimally knowledgeable individual may not be able to eliminate any distractor option. On your coded answer sheet, you will circle the distractor(s) which would be eliminated by a minimally knowledgeable salesperson.

INSTRUCTION #4

For this task, you will inspect each item distractor and identify those distractors which the typical or average knowledgeable salesperson would be able to eliminate. That is, you will identify those distractors which a salesperson possessing average knowledge would recognize as being obviously wrong. On an extremely easy item, this might be all distractors except the keyed response. On a very hard item, a salesperson with average knowledge may not be able to eliminate any distractors. On your coded answer sheet, you will circle the distractor(s) which would be eliminated by a typical salesperson possessing average knowledge.

Instruction #1 and Instruction #2 were the minimal and average qualification instructions for the Angoff method and Instructions #3 and #4 were the corresponding qualification instructions for the Nedelsky method. The presentation of four sets of instructions, along with four specially developed parallel forms of the Real Estate Examination, was counterbalanced over the four groups of judges. This was done in the following manner:

GROUP 1 (Minority and Sex Bias Review Committee)

Instruction #4:	Form 3
Instruction #3:	Form 1
Instruction #2:	Form 2
Instruction #1:	Form 4

GROUP 2 (Practicing Broker Review Committee)

Instruction #2:	Form 1
Instruction #1:	Form 3
Instruction #4:	Form 4
Instruction #3:	Form 2

GROUP 3 (Legal Review Committee)

Instruction #3:	Form 4
Instruction #4:	Form 2
Instruction #1:	Form 1
Instruction #2:	Form 3

GROUP 4 (Consultant Panel)

Instruction #1:	Form 2
Instruction #2:	Form 4
Instruction #3:	Form 3
Instruction #4:	Form 1

In this way, the experiment partially controlled for both practice and form effects. Each of the four parallel forms consisted of 64 four-choice items.

The structural model for the experimental design was:

$$Y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + \pi_{m(i)} + \alpha\beta_{ij} + \alpha\gamma_{ik}$$
$$+ \beta\gamma_{jk} + \beta\pi_{jm(i)} + \gamma\pi_{km(i)} + \alpha\beta\gamma_{ijk}$$
$$+ \beta\gamma\pi_{jkm(i)}$$

Where: Y_{ijkm} = cutting score for the mth judge in the ith group under the jth instruction and kth method.

α_i = group (i = 1,4)

β_j = instruction (j = 1, 2)

γ_k = method (k = 1, 2)

$\pi_{m(i)}$ = judge nested within the ith group.

RESULTS

The experimental effects in the repeated measures design were estimated using least squares procedures. Table 1 gives the means for the four groups of judges, two methods, and the skill level instructions. Table 2 presents the analysis of variance of the cutting score levels. The reader will note that while a significant group main effect was observed, there was also a significant group X method interaction. As can be seen by looking at the mean cutting scores presented in Table 1, the Angoff method consistently resulted in the setting of higher cutting scores. Also the cutting score for the average competence instruction was higher than the minimally competent instruction for both methods across all groups. Regardless of the group of judges, the Angoff method produced the smallest variation between cutting score levels across both method and instruction. Figure 1 presents a plot of the means for each method for each of the four groups. The interaction is disordinal, that is, all groups with the exception of Group 3 (the Legal Review Committee) obtained considerably higher cutting scores for the Angoff procedure. It is also interesting to note that the judges with an academic background (Group 4) had the largest method effect.

In Table 2 the significant main effect for instruction and the lack of a statistically significant interaction between instruction and method suggests that both methods are capable of yielding cutting scores that discriminate the minimally qualified from individuals of average qualifications. However, a closer inspection of the data indicates that the interaction between methods and instructions fell just short of significance ($p \cong .06$). A comparison of the spread between the mean cutting scores for minimally qualified and those who have average qualifications for the two methods indicate that the Angoff method

Table 1

Group Mean Cutting Scores by Instructions and Methods

	Angoff				Nedelsky			
	Minimal		Average		Minimal		Average	
Group*	Score	Percent of Total	Score	Percent of Total	Score	Percent of Total	Score	Percent of Total
1	40.085	62.6	44.972	70.3	28.312	44.2	32.250	50.4
2	35.013	54.7	41.936	65.5	27.639	43.2	31.833	49.7
3	38.155	59.6	48.445	75.7	40.542	63.3	45.854	71.6
4	43.484	67.9	53.035	82.9	27.517	43.0	28.304	44.2

*Group 1 consisted of four members of the Minority and Sex Bias Review Committee.
Group 2 consisted of three members of the Practicing Broker Review Committee.
Group 3 consisted of four members of the Legal Review Committee.
Group 4 consisted of five members of the Consultant Review Committee.

Table 2

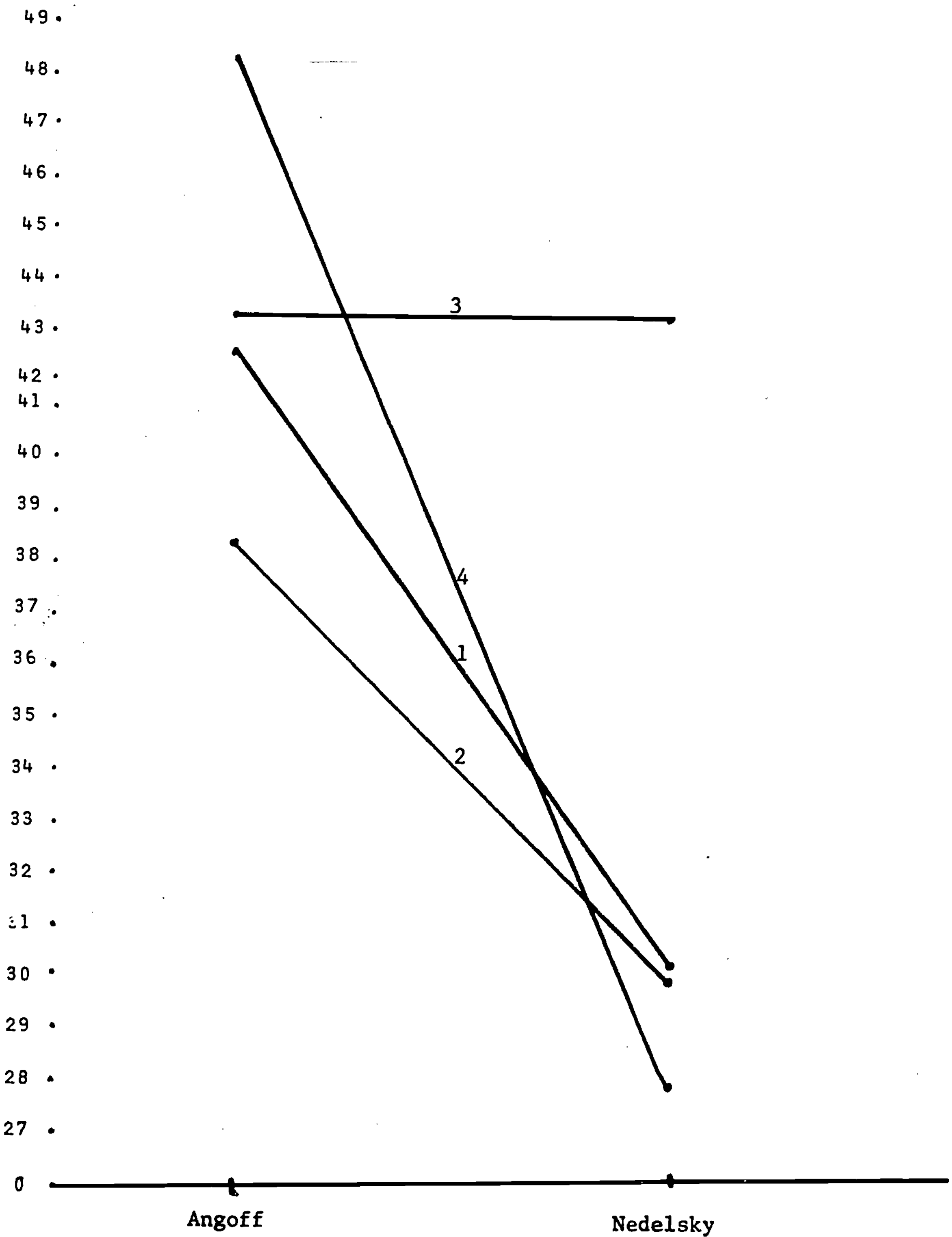
Analysis of Variance Table for the Cutting Scores

Source	Degrees of Freedom	F Ratio
Between	16	
Grand Mean	1	1114.1147*
Error	15	
Group	3	4.4268**
Error	12	
Within	48	
Instruction	1	41.1110*
Error	15	
Instruction X Group	3	0.6257
Error	12	
Method	1	18.1910*
Error	15	
Method X Group	3	5.4313*
Error	12	
Instruction X Method	1	4.1533
Error	15	
Instruction X Method X Group	3	0.5276
Error	12	

* $p < .01$
** $p < .05$

Figure 1

Mean Cutting Scores for the Four Groups of Judges For the Angoff and Nedelsky Methods



appears to be somewhat more discriminating. That is, the Angoff method yielded cutting score means of 39.18 and 47.10 for the minimally qualified individuals and individuals possessing average qualifications, respectively. The comparable figures for the Nedelsky method were 31.00 and 34.56.

With respect to cutting score level the results suggest that item probability judgments from most populations of experts will give significantly higher cutting score levels when using the Angoff method than when using the Nedelsky method. Although both methods appear to be able to yield cutting scores which discriminate the "idealized" individual having minimal qualifications, the Angoff method seems to be somewhat more discriminating than the Nedelsky method.

In order to investigate levels of agreement inter-judge correlations across their item judgments were computed, transformed using Fisher's r to z, and then averaged within the cells of the original design. High correlations between pairs of judges within the same cell indicate that the vector profiles generated by their respective judgments in the same set of items are similar. It would seem that preferred methods would demonstrate both a higher inter-judge agreement with respect to item judgments as well as greater consistency with respect to cutting score level both within and across populations.

Although it is tempting to use the transformed correlations as dependent variables in the previous analysis of variance design, this would leave the unsolved problem of how to determine the appropriate degrees of freedom for error terms as well as an acceptable method for correcting the varying dependencies among the within cell correlations. However, a simple comparison of the Angoff and

Nedelsky methods with respect to their average intercorrelations indicate that there was somewhat greater inter-judge agreement in item profiles for the Angoff method ($\bar{r}$ = .28) than for the Nedelsky method ($\bar{r}$ = .13).

Although the average intercorrelation is virtually equal for the minimum and average instructions ($\bar{r}$ = .214 and .215 respectively) there appears to be an interaction with method. That is, the Angoff method yielded average intercorrelations of .32 and .24 for minimum and average qualifications while the corresponding figures for Nedelsky were .11 and .19. Although the agreement was generally low for both methods, it appears that there was somewhat more agreement under the Angoff procedure. The differences in correlations do not appear to be the results of systematically smaller within-judge variance across items for either method. That is, there was no systematic difference in the range of item ratings the judges gave items under the different methods.

Group membership and inter-judge agreement also showed some interesting relationships. Group 3 (the Legal Committee) and Group 4 (professors of real estate) demonstrated higher with in group agreement regardless of method and instruction ($\bar{r}$ = .33 and .29 respectively) than did either the Sex and Minority Group ($\bar{r}$ = .09) and the Practicing Brokers Group ($\bar{r}$ = .15). The lawyers appear to be more consistent with respect to both cutting score level and inter-judge agreement across methods.

The academicians (Group 4) were characterized by the least stability in cutting score levels across methods yet they demonstrated almost as much inter-judge agreement within method as did the lawyers.

The Angoff method judgements for individuals with average qualifications yielded a cutting score that was somewhat less than the estimated average score for the applicant population (47.10 versus 51.99). The parallel estimate using the Nedelsky method was considerably lower than the estimated applicant mean score (34.56 versus 51.99). This estimate of the applicant population mean score was based on item pretest data.

The Nedelsky derived cutting score levels are sufficiently low that one must question their usefulness in practical situations except as a prescreening device rather than a final or sole criterion for licensing. Knowledge of less than half of the information judged as relevant to performing an occupation does not seem to be sufficiently rigorous criteria for licensing. The Angoff cutting score seems to be somewhat closer to the "mark" in that when considering an individual with average knowledge the judges arrived at a cutting score much closer to the mean score for the applicant population.

DISCUSSION

Certain of the results confirm the findings of the Andrew and Hecht (1976) and the Brennan and Lockwood (1979) studies. In particular, it was found that item probability methods based on the Angoff procedure tended to yield significantly higher cutting scores than the procedure outlined by Nedelsky. These findings applied to both the minimal and average qualification instructions. In addition, this study indicated that the Angoff method showed somewhat higher inter-judge agreement and was better able to define cutting scores with less overlap when judging on the basis of minimally qualified individuals vs. those with average qualifications.

The question arises: Why or how did the group of lawyers manage to arrive at the same cutting score estimate for both the Angoff and Nedelsky methods? One possibility is an experimenter effect. That is, in any field experiment with human subjects there is a possibility that the subjects or some class of subjects may consciously or unconsciously perceive that a positive goal of their task would be to orient their behavior to bring about what they see as consistent results. In fact, in the case of lawyers, their training and experience may encourage this sort of need for consistent answers regardless of the path taken to arrive at the answer.

Observations made during the experiment suggest that the short training session with examples which were offered before the experiment began may not have been sufficiently comprehensive for a complete understanding of the tasks by all group members. Questions from participants indicated that they found the Nedelsky task far more difficult to carry out. It is felt that this possibly incomplete and differential understanding of the Nedelsky tasks by some participants contributed to the lower level of agreement than was found in the Angoff tasks. The difficulty of the Nedelsky task for some participants was underscored by the fact that on the average it took twice as long to complete as the Angoff method.

CONCLUSIONS

Three out of four groups of judges arrived at significantly higher cutting score levels using the Angoff method than when using the Nedelsky procedure. The Angoff procedure was more effective in setting cutting score levels that distinguished the minimally qualified practitioner from the individual with average qualifications. Although inter-judge agreement with respect to the pattern of item responses was generally low for both procedures, the Angoff method demonstrated somewhat higher agreement.

REFERENCES

Anastasi, A. Psychological Testing (Fourth Edition). New York: MacMillan, 1976.

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

Angoff, W. H. Scales, Norms, and Equivalent Scores. In R. L. Thorndike (editor) Educational Measurement (Second Edition). Washington, D. C.: American Council on Education, 1971, Ch. 15(a).

Brennan, R. L. and Lockwood, R. E. A comparison of two cutting scores procedures using generalizability theory. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1979.

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

Millman, J. Criterion-Referenced Measurement. In W. J. Popham (Ed.), Evaluation in Education: Current Applications. Berkeley, California: McCutchan Publishing Co., 1974.

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

Popham, W. and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.