A Comparison of Two Approaches to Setting Passing Scores Based on the Nedelsky Procedure

Joseph C. Saunders, Joseph P. Ryan, and Huynh Huynh
University of South Carolina

Two versions of the Nedelsky procedure for setting minimum passing scores are compared. Two groups of judges, one using each version, set passing scores for a classroom test. Comparisons of the resulting sets of passing scores are made on the basis of (1) the raw distributions of passing scores, (2) the consistency of pass-fail decisions between the two versions, (3) the consistency of pass-fail decisions between each version and the passing score established by the test designer, and (4) the mean pairwise agreement between judges across groups. The two versions of the procedure are found to produce essentially equivalent results. In addition, a significant relationship is observed between the passing score set by a judge and that judge's level of achievement in the content area of the test.

Passing scores are needed in a broad variety of situations, including (1) entrance examinations, (2) tests for advancement of students from unit to unit in individually prescribed instructional programs, (3) minimum competency testing, and (4) certification or licensing examinations. Though writers such as Glass (1978) charge that passing scores for minimum competency testing are usually selected arbitrarily and frequently used unwisely, others (Hambleton, 1978; Shepard, 1976) have documented the need for cutoff scores in such areas as objectives-based programs and individualized instruction. This paper presumes the practical necessity of passing scores and explores ways in which they can be established more objectively.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 2, Spring 1981, pp. 209–217
© Copyright 1981 Applied Psychological Measurement Inc.
0146-6216/81/020209-09$1.45

Procedures for Setting Passing Scores

Various procedures for setting passing scores or "standards" have been developed (see Meskauskas, 1976). Most can be placed into one of three broad categories: (1) comparisons with the performance of others, (2) considerations of the consequences of misclassification, and (3) examinations of item content. Standard-setting procedures in the first two categories generally require actual student response data or assume a theoretical, statistical distribution of such data; content-based methods use judgments of content experts. Content-based methods frequently are used with tests when student performance data are not available.

Methods for determining passing scores by analyzing test content require a judge or group of judges to estimate the probable score of a hypothetical examinee responding at the level of minimum acceptable performance. Three of the best-known content-based procedures are those proposed by Angoff (1971), Ebel (1972), and Nedelsky (1954). In using the Angoff method, each judge estimates the probability that the "minimally acceptable person" would respond correctly to each item; the passing score is de-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

termined by summing the estimated item probabilities (Angoff, 1971; Zieky & Livingston, 1977). In the Ebel procedure, judges sort items into categories of "relevance" and "difficulty." Each judge then estimates the proportion of correct answers in each category expected of a "minimally qualified" examinee. The passing score is the weighted sum of these proportions, with the weight for each category being the number of items it contains (Ebel, 1972).

The Nedelsky method is restricted to multiple-choice tests. Every response option is considered by each judge, who decides which options could be rejected as incorrect by an examinee performing at the minimum passing level. The probability that someone at this level would respond correctly to the item is taken to be the reciprocal of the number of remaining options (i.e., one divided by the number of options that the minimally performing examinee should not be able to reject). The passing score is the sum of these reciprocals for all items.[1] In all cases, the passing score can be expressed as a fraction or percentage of the total number of items.

Comparisons of the Application of the Methods

The methods discussed above, though operationally quite different, have logical similarities. All involve a judge or group of judges with expertise in the content area, who consider the items on a test and estimate the probable score of a "minimally competent" student on the basis of a value judgment. It might seem that they could be expected to produce equivalent passing scores. Research reported in the literature indicates that this equivalence is not always observed. In a study comparing the Ebel and Nedelsky procedures, Andrew and Hecht (1976) found that the two standard-setting methods produced significantly different passing scores. Perhaps an even more important consideration was that 45% of the examinees being tested were classified differently by the two passing scores (Glass, 1978). In research utilizing the Nedelsky and Angoff procedures, Brennan and Lockwood (1979) also reported a substantial difference in the resulting passing scores.

When several judges are used, the variation among judges' individual passing scores can also become an issue. A certain degree of variation might be expected. It is usually suggested that the different passing scores be reconciled either by averaging the scores or by requiring judges to reach a consensus passing score. Andrew and Hecht (1976) found that passing scores obtained by consensus and by averaging did not differ significantly. In at least one reported case, however, the amount of variation among passing scores set by a group of judges using the Nedelsky procedure was substantial, and the procedure was rejected as unfeasible (Meskauskas & Webster, 1975). The averaging process treats the variation in passing scores as random or "error" variation. It might be, however, that differences in passing scores are related systematically to characteristics of the judges. If passing scores are to be useful, they should not depend too much on the characteristics of a particular judge or group of judges. Such characteristics, once identified, could possibly be controlled to prevent them from exerting an undue influence on the standard-setting process. One characteristic that intuitively might be expected to show such a relationship is the judge's own level of achievement in the relevant area.

Focus of This Paper

This paper deals only with the Nedelsky procedure. Two versions of the procedure appear to be in use. In the first version, judges must classify response options into two categories: (1) those that should be rejected as incorrect by the minimally performing examinee and (2) those that should not. In the alternative version, a third category, "undecided," is also used when the

[1] In the original formulation, Nedelsky (1954) offers further refinements, such as estimating the standard deviation of the chance distribution of scores and using it in conjunction with setting the passing score. These refinements are not considered in this paper.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

judge is unable to classify the response option as one that either should or should not be rejected. In calculating passing scores, an option classified as undecided counts as one-half a remaining option. That is, the probability that a minimally qualified student would respond correctly to the item is considered to be the reciprocal of the sum of the number of options which could not be rejected and one-half the number of undecided options. Decisions between the two versions seem to be based on the preferences of the judges, rather than on any theoretical consideration (e.g., Amitrano-Paiva & Vu, 1979; Smilansky & Guerin, 1976). Nedelsky (1954) discussed the use of the alternative procedure; he apparently felt the two versions were equivalent.

The purpose of this paper is twofold. First, a comparison is made between the two versions of the Nedelsky procedure. Second, the relationship between the achievement levels of judges and the passing scores they set was assessed.

Method

Subjects

In order to compare the two versions of the Nedelsky procedure, subjects acting as judges were divided into two groups. Group A used the two-category version of the procedure to set passing scores on an achievement test, and Group B used the three-category version. The results were compared using the distributions of passing scores, as well as the consistency of decisions based upon the scores. Also, to determine the relationship between judges' achievement and passing score, the correlation between measures of the two variables was calculated.

Data for the study were obtained from students in an introductory course in educational research and measurement. The course was conducted via videotape at a number of regional campuses of a large state university. All subjects were graduate students; many were experienced teachers.

Instrument

The instrument for which passing scores were set, and by which judges' achievement levels were determined, was the course midterm examination, a 40-item, four-option, multiple-choice test, constructed by the course instructor (the second author). The test covered such topics as the nature of the research process, observation and measurement, sampling, and item analysis. The exam has been revised over several years to reach a high degree of content validity; in its most recent administration, it showed an internal consistency (KR20) reliability index of .82. Thus, scores on the test are considered to be valid and reliable measures of achievement.

Treatment Groups

All students enrolled in the course wrote the midterm examination as a regular course requirement. The exams routinely were graded and returned to the students for discussion in class. The students then were asked to participate in an exercise involving the use of the Nedelsky procedure to determine a passing score for the test. Although participation in the exercise was voluntary, more than 95% of the students chose to participate. Of the 148 students agreeing to participate, 30 were deleted from the study due to failure to follow instructions, missing identification codes, or missing achievement data, leaving 118 students as the sample used in the experiment. Students were assigned randomly to groups, stratified by course section to control for possible differences among regional campuses. Then they were given copies of the test, along with detailed instructions on the Nedelsky procedure. Instructions for the two groups differed only with respect to the version of the procedure used.

Definition of Minimum Competence

Minimum acceptable performance was defined for the students as the lowest level of per-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

formance on the test for which a grade of "B" would be awarded. This level was chosen as appropriate, since one of the requirements of the students' degree programs is that a "B" average be maintained. For each incorrect response option on the test, the students were instructed to respond to the question, "Should the student performing at the minimum acceptable level (as defined above) be able to reject this option as incorrect?" Spaces were provided for that purpose beside each option. For the two-category version (Group A) of the procedure, the possible responses were "yes" and "no." The three-category version (Group B) also allowed "undecided" as a possible choice. In order to minimize any possible confounding effect produced by the students' knowledge of previously existing course standards, they were not required to calculate their resulting Nedelsky passing scores; this was done by the authors. Each student responded individually; no attempt was made to determine consensus passing scores.

Comparison Procedures

The frequency distributions of passing scores produced by the two groups were compared using the Kolmogorov-Smirnov two-sample test, a broad test sensitive to any difference in the two distributions (Hollander & Wolfe, 1973). The distributions of passing scores are given in Table 1. All passing scores were rounded upward to the nearest whole number, that is, the number of correctly answered items necessary for an examinee to be classified as passing. Decision consistency was assessed via comparisons of the proportions of students writing the exam who were classified similarly by the two versions. The comparisons involve both the raw proportion of agreement and coefficient kappa, (κ), which corrects for chance agreement. To assess overall agreement between the two groups of judges, both the mean and median passing scores for each group were used. Also, decisions based on the groups' passing scores were compared with those based on the standard established by the course instructor. In addition, to assess the consistency of decisions between individual judges, the average proportion of agreement for all possible pairs of judges (across groups) was calculated. Finally, to assess the relationship between judges' achievement and passing score, the Pearson product-moment correlation coefficient was determined for the students' examination grades and their Nedelsky passing scores. For this calculation, the two groups were combined.

Results

Although the overall passing score distributions for the two groups, displayed in Table 1, were not identical, no significant difference (p = .36) was found. As can be seen in Table 2, the two forms also produced highly consistent classification decisions. If the mean passing score for each group is used as a standard, only 7 of 185 students taking the test would have been classified differently, a proportion of agreement of .96 (κ = .94). The exact median passing scores from the two groups were 31.17 and 31.37, respectively. Rounding upward, both these values became 32. Thus, use of the median passing score produced the surprising result of complete agreement in classification.

That the two versions produced passing scores yielding consistent decisions does not, in itself, mean that the scores are useful in practice. However, further comparisons of decisions based on the Nedelsky passing scores with those based on standards previously established by the course instructor (32 correct answers for a grade of "B") also showed a high degree of agreement (Table 3). Using the group mean passing score as the standard, 11 of 185 students were classified differently by Group A (the two-category version) and the course instructor's preset standard (proportion of agreement = .94, κ = .94). For Group B (the three-category versions), the proportion was .98 (7 students classified differently), with κ = .96. The group medians, rounded up to 32, coincided exactly with the course instructor's standard. Here again, use of

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Distributions of Passing Scores from Two Versions of the Nedelsky Procedure

Passing Score	Frequency Group A	Group B	Passing Score	Frequency Group A	Group B
13	0	1	26	2	4
14	1	0	27	1	0
15	0	0	28	5	2
16	2	1	29	4	4
17	0	1	30	0	1
18	1	0	31	3	5
19	0	0	32	5	3
20	3	1	33	2	3
21	1	0	34	6	10
22	1	0	35	6	5
23	2	2	36	3	2
24	2	4	37	3	5
25	1	2	38	5	3

	N	MEAN	MEDIAN	S.D.
Group A	59	29.88	31.17	6.38
Group B	59	30.51	31.37	5.79

Kolmogorov-Smirnov D = .170 (p = .36)

the group medians produced complete agreement. Considering the judges individually, the average pairwise proportion of agreement was .66.

As was noted previously, students in both groups were combined to consider the relationship between judges' achievement and passing score. Such a relationship, if it exists, might be expected to hold across methods; in any event, the demonstrated equivalence of the two forms suggests the reasonableness of combining the two groups. The linear correlation between achievement and passing score for the students in the study was .30 with $p = .001$. (For the groups considered separately, the correlations were .37 and .25, with $p = .004$ and $p = .05$, respectively.) Thus, achievement in the subject matter area accounted for 9% of the observed variation in passing scores.

Discussion

From the results of this study, the two- and three-category versions of the Nedelsky procedure yield equivalent results. The finding holds both in terms of the empirical distributions of passing scores and of consistency in classification decisions. Additionally, there was a close correspondence both in distributions of passing scores and in classification decisions between passing scores set by the students and those set by the preset standard established by the course instructor. The substantially higher level of agreement of grouped passing scores over the

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Decision Consistency of Passing Scores from Two Versions of the Nedelsky Procedure

Case I: Using the mean of several judges

		Group A: fail	Group A: pass	
Group B	fail	44	7	51
	pass	0	134	134
		44	141	185

($p = .96$, $\kappa = .94$)

Case II: Using the median of several judges

		Group A: fail	Group A: pass	
Group B	fail	55	0	55
	pass	0	134	134
		55	134	185

($p = 1.00$, $\kappa = 1.00$)

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 3
Decision Consistency of Course Instructor's Standard with Passing Scores from Two Versions of the Nedelsky Procedure

Case I: Using the mean of several judges

		Group A			Group B		
		fail	pass		fail	pass	
Instructor's Pre-set Standard	fail	44	11	55	51	4	55
	pass	0	130	130	0	130	130
		44	141	185	51	134	185
		(p = .94, κ = .94)			(p = .98, κ = .96)		

Case II: Using the median of several judges

		Group A			Group B		
		fail	pass		fail	pass	
Instructor's Pre-set Standard	fail	55	0	55	55	0	55
	pass	0	130	130	0	130	130
		55	130	185	55	130	185
		(p = 1.00, κ = 1.00)			(p = 1.00, κ = 1.00)		

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

average pairwise agreement suggests the utility of using more than one judge when feasible. Although either the mean or the median of several judges' passing scores could be used to set the final passing standards, the median, rather than the mean, might be more appropriate. The median's resistance to the influence of extreme scores would seem to reduce some of the effect of variability in passing scores from a group of judges.

Some variation was observed in the scores from both groups of judges. The slightly smaller standard deviation of passing scores from Group B, using the three-category version of the procedure, might be a point in favor of the use of that version. The significant positive correlation between judges' achievement and passing score indicates that at least a small portion of the observed variation in passing scores was related systematically to a characteristic of the judges. Other relevant characteristics that also relate systematically to judges' passing scores might be identified. Knowledge of these characteristics and their relationship to passing scores could lead to their elimination, control, or utilization in the standard-setting process.

Consider the case where several judges are selected from a large pool. This might occur, for example, when local standards are established for a minimum competency of high school graduation examination. The variation in different judges' passing scores might be reduced by selecting judges of similar ability. Alternatively, passing scores from different judges might be adjusted, based on their differing ability levels. Such a procedure might make it easier to reach a consensus passing score. Admittedly, this might not be practical, or even desirable, in many circumstances. In the last analysis, passing scores must still be based on value judgments. Nevertheless, by removing or controlling some of the effects of selecting particular individuals as judges, the setting of passing scores on the basis of expert judgment might be made a more objective process.

In conclusion, this study has shown that the two versions of the Nedelsky procedure considered here produced equivalent passing scores. Also, it was shown that the passing scores set by different judges were related positively to the judges' own achievement. It should be noted that the study involved the setting of passing scores for a single test, using as judges students who took the test but who were not responsible for constructing it. Such judges may not have the broad knowledge of other students, of how such tested content fits into the total curriculum, and of the subject matter itself that, say, faculty members might have. Thus, the observed results must be seen as suggestive rather than conclusive. However, given the results of this study, a choice between the two versions justifiably could be made on practical grounds, such as the preference of the judges.

References

Amitrano-Paiva, R. E., & Vu, N. V. *Standards for acceptable level of performance in an objectives-based medical curriculum: A case study.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement,* 1976, *45,* 4-9.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education, 1971.

Brennan, R. L., & Lockwood, R. E. *A comparison of two cutting scores 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 NJ: Prentice-Hall, 1972.

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

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement,* 1978, *15,* 277-290.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Hollander, M., & Wolfe, D. A. *Nonparametric statistical methods.* New York: Wiley, 1973.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research,* 1976, *46,* 133-158.

Meskauskas, J. A., & Webster, G. W. The American Board of Internal Medicine recertification examination process and results. *Annals of Internal Medicine,* 1975, *82,* 577-581.

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

Shepard, L. A. Setting standards and living with them. *Florida Journal of Educational Research,* 1976, *18,* 28-32.

Smilansky, J., & Guerin, R. O. *Minimal acceptable performance levels for criterion-referenced multiple-choice examinations and their validation.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

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

Acknowledgments

This work was performed pursuant to Grant No. NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not necessarily reflect NIE positions or policy and no official endorsement should be inferred. The comments of Anthony J. Nitko and Elizabeth M. Haran are gratefully acknowledged.

Author's Address

Send requests for reprints or further information to Huynh Huynh, College of Education, University of South Carolina, Columbia SC 29208.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/