Comparison of Direct and Indirect Methods For Setting Minimum Passing Scores

Richard R. Reilly
Stevens Institute of Technology

Donald L. Zink and Edmond W. Israelski
American Telephone & Telegraph Company

Several studies have compared different judgmental methods of setting passing scores by estimating item difficulties for the minimally competent examinee. Usually, a direct method of estimating item difficulties has been compared with an indirect method suggested by Nedelsky (1954). Nedelsky's method has usually resulted in a substantially lower cutoff score than that arrived at with a direct method. Two studies were carried out for the purpose of comparing a direct method of setting passing scores with an indirect method that allowed judges to estimate the probability of the minimally competent examinee eliminating each incorrect alternative. In Study 1 a sample of 52 first-level supervisors used both methods to estimate passing scores on a content-oriented selection test for building maintenance specialists. In Study 2 a sample of 62 first-level supervisors used both methods to estimate passing scores on an entry level auto mechanics test. Results of both studies showed that the variance component for method was relatively small and that for raters was relatively large. Reliability estimates of judgments and correlations between judged difficulties and empirical difficulties showed the Angoff (1971) approach to be slightly superior. Results showed no particular advantage to using an indirect approach for estimating minimal competence.

Recently, the problem of setting passing scores has received considerable attention from researchers. Most of this research has focused on the use of systematic judgment to set minimum cutoffs. Methods proposed by Angoff (1971), Ebel (1972), and Nedelsky (1954) involve the judgment of experts regarding the probability of minimally competent examinees passing specific items. The Angoff and Ebel methods might be called direct approaches in that they ask the judge to estimate the probability for each item (Angoff) or subset of items (Ebel) directly, whereas the approach suggested by Nedelsky is indirect. Nedelsky's procedure requires judges to indicate how many incorrect alternatives a minimally competent examinee could eliminate. The item probability is then estimated as the reciprocal of the number of remaining options.

Several studies have compared the direct and indirect approaches to setting cutoffs. Studies by Andrew and Hecht (1976), Brennan and Lockwood (1980), Skakun and Kling (1980), and Harasym (1980) found the Nedelsky procedure to result in lower cutoffs. In most cases the difference was substantial. In only one study (Skakun & Kling) was the difference in passing scores *less* than 10 percentage points of the total number of items. Glass (1978) cited the Andrew and Hecht study as evidence that the method chosen has serious consequences and argued that this casts doubt on the premise that judgment is an acceptable method for establishing minimal competence.

Both direct methods involve relatively straightforward reasoning. An item, or subset of items, is presented and each judge is asked to make a probability estimate. In the Nedelsky approach, how-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 4, Fall 1984, pp. 421–429

0146-6216/84/040421-09$1.70

ever, the judges are constrained to estimating probabilities of 1.0 or 0 for elimination of the incorrect options. That is, the implied probability is 1.0 if the judge thinks that the minimally competent candidate could eliminate the option and 0 if the judge thinks that the minimally competent candidate could not eliminate the option. It is then assumed that the candidate will make a random guess among the remaining alternatives. Viewed in this light, the Nedelsky approach seems unreasonable on two counts. First, the constraints on judges' probability estimates are unrealistic; and secondly, the assumption that examinees operate randomly is not supported by research on examinee guessing behavior (e.g., Slakter, 1967).

A more reasonable indirect procedure would have judges specify the *probability* of a minimally competent examinee eliminating each incorrect alternative. The probability of answering the item correctly could then be estimated with rules of basic probability. On a test with four choices for each item, for example, given probabilities for elimination of incorrect options A, B, and C, the probability of choosing the correct answer, D, can be found using the multiplication theorem:

$$P(D) = \sum_{k=1}^{2}\sum_{j=1}^{2}\sum_{i=1}^{2} P(D|A_i, B_j, C_k)\, P(A_i, B_j, C_k). \quad (1)$$

$P(D|A_i, B_j, C_k)$ represents the conditional probability of choosing D, given the joint occurrence of A_i, B_j, C_k, where the subscript 1 represents elimination of an option and the subscript 2 represents nonelimination. $P(A_i, B_j, C_k)$ is the probability of the joint occurrence of a particular combination of eliminated and noneliminated options, assuming that the probabilities of elimination of each alternative are independent. If, for example, the probabilities of eliminating options A, B, and C were .6, .7, and .9, respectively, the probability of choosing D would be .66.

Estimating Minimum Competence with the Indirect Method

For purposes of the indirect method, assumptions about examinee behavior are similar to those of Nedelsky (1954). It is assumed that the final outcome of a multiple-choice question is the result of a two-step procedure on the part of the examinee. In the first step, alternatives thought to be incorrect are eliminated; the second step involves choosing randomly among the remaining alternatives. Because Nedelsky allows only probabilities of 1.0 or 0 to be attached to the elimination of the incorrect alternatives, the probability of the correct answer is calculated simply by taking the reciprocal of the m remaining alternatives. This paper's method allows the probability for elimination of incorrect alternatives to vary from 0 to 1, so that the probability of a correct answer must be calculated as the sum of all possible joint occurrences of eliminated and noneliminated options leading to a correct answer.

In a four-choice item, for instance, there are eight possible combinations of elimination or nonelimination of the incorrect alternatives. Table 1 shows (1) the probability of each occurrence, $P(A_i, B_j, C_k)$; (2) the conditional probability of a correct answer, D, given that occurrence, $P(D|A_i, B_j, C_k)$; and (3) the probability of an incorrect answer given the same occurrence, $P(\bar{D}|A_i, B_j, C_k)$. Each row shows the probabilities of a correct and incorrect answer when certain options are eliminated. In the seventh row, for example, the probability of eliminating option C only *and* obtaining a correct answer is equal to the probability of eliminating option C only, which is $(.4) \times (.3) \times (.9) = .108$ multiplied by the conditional probability of a correct answer given the elimination of C, $\frac{1}{3} \times (.108) = .036$.

If the above approach is reasonable, then passing scores based on probability estimates for the elimination of incorrect alternatives should yield results comparable to those obtained with a more direct approach, such as Angoff's (1971). More importantly, such a finding would answer the criticism of Glass (1978) that the method of judgment produces differences large enough to preclude the use of judgment in standard setting.

At a more practical level, the two studies described herein examined the feasibility of applying judgmental techniques to set minimum scores on tests of job knowledge in two skilled craft areas,

Table 1
Probability of Joint Occurrences and Correct, and Incorrect Answers Given Different Possible Eliminated Options

Options Eliminated	$P(A_i, B_j, C_k)$	P(Correct)	P(Incorrect)
A, B, C	.378	.378	.0
A, B	.042	.021	.021
A, C	.162	.081	.081
A	.018	.006	.012
B, C	.252	.126	.126
B	.028	.009	.019
C	.108	.036	.072
None	.012	.003	.009
Item Probability		.660	.330

where the judges were first-level supervisors in an actual work setting. Two separate experiments were carried out using test items designed to measure job knowledge for building maintenance specialists (BMS; Study 1) and automotive mechanics (Study 2). In both investigations, the Angoff approach was compared with an indirect approach, but the mode of item presentation differed in the two studies. It was hypothesized (1) that the difference between passing scores using the two methods would be smaller than differences found in studies comparing direct estimation methods with Nedelsky's (1954) procedure, and (2) that both methods would produce similar results with respect to interjudge consistency.

Study 1

In Study 1, the Angoff (1971) procedure was compared with an indirect approach, in which judges were asked to consider an item and estimate the probability that a minimally competent BMS could eliminate each incorrect alternative. Study 2 differed from Study 1 in that the *correct* option was not shown to judges.

Method

Procedure. The Building Maintenance Qualification Test (BMQT), consisting of 139 items, was prepared for selecting BMS personnel. All of the items involved specific knowledge related to the BMS job as indicated by a job analysis. Each item was prepared by a panel of experts in the field and then reviewed for relevance and accuracy by another independent panel of experts. For practical reasons (i.e., the time required of experts), the 139 items were divided into two parts. Part I consisted of the first 70 items and Part II consisted of the last 69 items.

First-level BMS supervisors, none of whom were involved in test construction, were randomly assigned to one of two conditions. In the first condition the direct method of estimation proposed by Angoff (1971) was used in Part I to estimate probabilities for the minimally competent person. The indirect method proposed by the present authors was then used to estimate probabilities for the items in Part II. In the second condition the indirect method was used for Part I and the direct method was used for Part II.

Under the direct estimation condition, judges were given the following instructions:

> This method asks you to estimate the probability that a minimally qualified incumbent would answer the question correctly. In making your estimates it may be useful to think in terms of 100 minimally qualified incumbents: how many, or what percentage of those 100 would be able to answer the questions correctly?

Under the indirect estimation condition, judges were given the following instructions:

> One way that can be used to take a multiple-choice test is to reject or eliminate from consideration those alternatives that are recognized to be *incorrect*. We ask you to *estimate the probability* that a minimally qualified incumbent would *not choose* each incorrect alternative as the answer to the question. It will be useful to think of 100 minimally qualified incumbents: for each incorrect alternative, how many, or what percentage of those 100 would reject that alternative as the answer to the question?

Under each condition the concept of minimum competence was discussed, examples were used to clarify the procedure, and any questions judges might have about either procedure were answered.

Sample. The sample of judges consisted of 52 first-level supervisors in the building maintenance area. Because of some unusable data, the number of judges in the first condition was 27, versus 25 for the second condition.

Analyses. The judgments for Parts I and II of the BMQT were analyzed separately with a three-factor nested effects ANOVA, with raters nested within method. Variance components for each factor were estimated using methods suggested by Brennan (1977). The consistency with which judges assigned p values within the direct and indirect methods was examined using a method described by Winer (1971, pp. 289–290). The term "reliability coefficient" is used in discussing the obtained statistic, though an important distinction between intrajudge and interjudge reliability should be made. A given judge may be quite reliable in the sense proposed by van der Linden (1982) but quite inconsistent with respect to other judges' notions of mastery.

The correlation between the item probabilities yielded by both estimation methods was calculated for the 139 items. As a final step, the correlation between actual item p values (i.e., proportion passing) for a sample of 462 job applicants and p values estimated using each judgmental technique was calculated. (These correlations were based on a subset of 100 items that comprised Form A of the BMQT.)

In terms of a latent trait model it would be ideal to have the ability parameter estimate for the minimally competent examinee as well as individual item parameters. Then, probabilities yielded by the latent trait model could be compared with the probability estimates of a judge, or group of judges, for a minimally competent examinee. In the absence of such information, however, it is logical to assume that regardless of the ability parameter value, the probability of a minimally competent examinee passing an item will be positively correlated with the p value for the item, based on a random sample of examinees. Thus, the correlation between the empirical p values and the average p values across judges may be viewed as one type of evidence that the judgments have some validity.

Results

The ANOVA results and variance component estimates for the two parts of the BMQT are shown in Table 2 (for ease of interpretation all probabilities have been multiplied by 100). For both parts of the test, the estimated variance component for method was relatively small. (For Part I, the estimate was negative, suggesting an interpretation of zero for the method component.) The variance component for raters was the largest of the four experimental sources of variance in both parts.

Table 3 shows the interjudge reliability coefficients and the mean estimated probabilities for both methods and for both test parts. Because the number of judges in the two conditions differed, the reliability coefficients were stepped down to estimate the reliability for one judge. For both parts of the test the reliability for the direct method was higher. The means yielded by the direct method were also higher for both parts. For Part I, the difference was approximately 2.6 percentage points; for Part II, the difference was about 6.8 percentage points.

Table 4 presents the intercorrelations between the item probabilities estimated by three methods. The correlation between the direct and indirect probabilities was .8, suggesting reasonable agreement between the two methods. The correlations with the empirical item p values were .71 for the

Table 2
ANOVA and Variance Components for Probability Estimates for Parts I and II of BMQT

Part and Effect	df	MS	$\hat{\sigma}^2_\alpha$
Part I			
Method (M)	1	576.29	-14.80
Raters [R(M)]	46	24947.55	352.07
Items (I)	69	2988.47	45.88
MxI	69	786.35	20.14
Error	3,174	302.96	302.96
Part II			
Method (M)	1	26147.91	8.00
Raters [R(M)]	48	12040.46	121.29
Items (I)	68	5925.02	107.76
MxI	68	537.03	12.61
Error	3,264	221.68	221.68

direct method and .60 for the indirect method. Polynomial regression failed to support any nonlinearity in either relationship. Results for the first study indicated that the two methods produced less of a difference, on the average, than most previous studies; the data also suggest reasonable agreement between the two methods in terms of estimated item probabilities, and between each method and empirical probabilities.

Study 2

Method

Procedure. The second study involved determining a minimum passing score on a test for automotive maintenance mechanics. The Automotive Maintenance Qualification Test (AMQT) consisted of 175 items. For study purposes the test was randomly divided into two parts; Part I consisted of the first 88 items and Part II consisted of the remaining 87 items. The study procedures and instructions were the same as those used in Study 1, with one major exception. Judges using the indirect method were not shown the correct option when asked to indicate the probability of a minimally competent candidate being able to eliminate each of the three incorrect options.

Sample. The Study 2 sample consisted of 64 first-level supervisors in automotive mechanics randomly divided into two groups. One group used

Table 3
Mean Probabilities and Inter-Rater Reliabilities for Two Methods of Judging Minimal Competence

	Reliability		Means	
Method	Part I	Part II	Part I	Part II
Direct	.88(.21)	.93(.36)	65.32	77.96
Indirect	.80(.14)	.93(.33)	62.71	71.13

Note. Reliabilities for one rater are shown in parentheses.

Table 4
Intercorrelations Between Item Probability Estimates Using Direct, Indirect and Empirical Methods

Method	Indirect	Empirical
Direct	.80	.71
Indirect		.60

the direct method for Part I and the indirect method for Part II. The second group used the indirect method for Part I and the direct method for Part II.

Analyses. The analyses were the same as those performed in Study 1. The judgments for Parts I and II of the AMQT test were analyzed with the same ANOVA model, and variance components were estimated. Interjudge reliability coefficients were estimated for both direct and indirect estimation procedures. The correlation between the item probabilities yielded by both estimation methods was calculated for the 175 items. Empirical data on 190 recently trained auto mechanics were collected for two different 100-item forms (95 subjects each). Correlations between the estimated probabilities and the actual item *p* values were then calculated. (The two forms consisted of overlapping [39 common items] 100-item subsets.)

Results

Table 5 shows the results of the ANOVA and estimated variance components for Part I of the AMQT. It can be seen that the component related to method is relatively small. Consistent with the results of Study 1, the component for raters is the largest of the four experimental sources of variance. Table 5 also shows the results for Part II of the 175-item test. The component for method is larger than in the three other analyses, but is still relatively small compared to the rater component.

Table 6 shows the reliabilities and means for both methods and both parts of the test. As in Study 1, the reliabilities are higher for the direct method. The pattern of means, however, is not consistent. For Part I, the indirect method yielded a slightly higher mean; for Part II, the direct method

Table 5
ANOVA and Variance Components for Probability Estimates for Parts I and II of AMQT

Part and Effect	df	MS	$\hat{\sigma}_\alpha$
Part I			
Method (M)	1	11851.08	7.19
Raters [R(M)]	62	31156.58	350.96
Items (I)	87	3665.78	38.42
MxI	87	1206.95	29.21
Error	5394	272.30	272.30
Part II			
Method (M)	1	135181.53	40.64
Raters [R(M)]	62	20891.32	237.31
Items (I)	86	3221.19	28.42
MxI	86	1402.02	36.15
Error	5332	245.32	245.32

Table 6
Mean Probabilities and Single-Rater Reliabilities for Two Methods of Judging Minimal Competence

	Reliability		Means	
Method	Part I	Part II	Part I	Part II
Direct	.92(.25)	.92(.26)	64.45	74.41
Indirect	.83(.13)	.81(.12)	67.34	64.43

Note. Reliabilities for one rater are shown in parentheses.

yielded a mean that was 10 percentage points higher. Table 7 shows the item correlational data. The correlation between direct and indirect probability estimates was .51 across the sample of 175 items. For the judgmental versus empirical estimates, the direct estimation procedure resulted in higher correlations for both Forms A and B.

Discussion

The results of both studies suggest that the large differences in method obtained when the Nedelsky (1954) estimation procedure is compared with direct estimation procedures may be a result of unreasonable assumptions underlying the method. With the exception of Part II of the AMQT, the method components were all relatively small. The mean differences yielded by direct and indirect methods in the present study were smaller than those obtained in three of the previous four studies (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Harasym, 1981), and in one instance (Part I of the AMQT), the indirect method actually yielded a slightly higher mean. Only the Skakun and Kling (1980) study found differences comparable to the present study.

Apart from the advantage of focusing the judges' attention on the distractors, the indirect method used in the present investigation, like the Nedelsky approach, offers no particular advantage in efficiency. Indeed, the indirect method calls for more separate judgments and is more time consuming than the Angoff (1971) procedure. The direct estimation procedures, especially the Angoff procedure, are simpler to use and easier to explain to judges. Thus, in order for a case to be made for the use of an indirect approach to setting passing scores, superiority for the indirect method should be demonstrated.

It should also be noted that the indirect method may be based on some overly simplistic assumptions. The probabilities of elimination of options may not be independent as assumed here; examinees may eliminate pairs or triples of options in answering an item, based on some common characteristic. Obviously, this would call for a different and more complex model.

Table 7
Intercorrelations Between Item Probability Estimates Using Direct, Indirect and Empirical Methods

Method	Indirect	Form A Empirical	Form B Empirical
Direct	.51	.57	.45
Indirect		.47	.32

None of the results in the present study supported the superiority of the indirect method. The reliabilities were higher for the direct method (though, in one case, the difference was trivial). In addition, the correlations between the judged probabilities and empirical p values were consistently higher for the direct estimates. In using either the direct or the indirect method, the judgment task is ultimately one of estimating item difficulty for a hypothetical "minimally competent candidate." If it is assumed that the item difficulties for minimally competent examinees are monotonically related to the average item difficulties for a random sample of examinees, then the extent to which the judges' perceptions of difficulty correlate with the actual difficulty of items provides one type of validity evidence for the judgments. The correlations between estimated item difficulties and actual item difficulties in the two present studies were consistently higher for the direct method, suggesting that the direct method results in more reasonable difficulty estimates. Thus, the results suggest that judges can use the direct method more reliably and with better validity.

An important finding of the present study was the large variance component found for raters. Most previous studies have employed as judges small groups of individuals who have been involved in test development or who have some background in measurement. In the present study, the judges were first-level supervisors, had no background in test development or measurement, and were not involved in the test development process. Although the actual reliabilities were reasonably high, ranging from the low .80s to low .90s, the number of raters was quite large relative to other studies. The large rater component suggests that employers wishing to set passing scores with a judgmental approach should either use a large representative sample of raters, or require that raters undergo more extensive training in the judgmental process.

Shepard (1980) noted that variability among judges may be a more serious problem than differences due to approaches. She suggested three steps that might be taken to cope " . . . with the threat to validity implied by extreme ranges in judges' standards" (p. 454). First, representativeness of expertise should be ensured by carefully choosing the sample of judges. In the present studies an effort was made to select samples of judges who were representative in terms of expertise and geographic location of the larger population of supervisors. A second suggestion was that differences among judges' passing scores might be correlated with other individual differences. A recent study (Saunders, Ryan, & Huynh, 1981), for example, found the level of achievement of judges to be positively correlated with the passing score set. Although such systematic data were not analyzed in the present two studies, research of this type can be helpful in providing greater insight into the reasons for interjudge variability. The third suggestion made by Shepard was that validity evidence be collected. One type of evidence, the extent to which judges' estimates of difficulty match empirical difficulties, can be helpful in assessing the degree to which different methods or different judges are making reasonable discriminations among items.

It should be noted that for test items from more homogeneous domains, an alternative approach to the present method has been suggested by van der Linden (1982); the approach utilizes latent trait theory to examine intrajudge consistency and can yield useful descriptive data to help assess alternative methods of setting passing scores. In the present study, unfortunately, the test domain could not be considered homogeneous; in fact, the test was constructed to be content relevant for the jobs of auto mechanic and building maintenance specialist. Thus, the emphasis was on matching items to job tasks performed, rather than on sampling items from some homogeneous domain.

Conclusions

The present paper described two studies comparing direct and indirect methods of setting passing scores. The present results suggest that given more reasonable assumptions about the indirect judgment task, the differences between indirect and direct approaches will be smaller than found in most previous studies. However, the results of both studies are consistent in producing no evidence to support the superiority of an indirect estimation procedure. Quite to the contrary, the data suggest

that judgments made directly are made more reliably and are more consistent with empirical difficulties. Thus, if a judgmental approach is to be used, there is no reason to prefer an indirect method of setting passing scores.

References

Andrew, B. J., & Hecht, J. T. (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., pp. 508–600). Washington DC: American Council on Education.

Brennan, R. L. (1977). *Generalizability analyses: Principles and procedures* (ACT Technical Bulletin No. 26). Iowa City IA: The American College Testing Program.

Brennan, R. L., & 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 NJ: Prentice-Hall.

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

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. *Educational and Psychological Measurement, 4,* 725–734.

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

Saunders, J. C., Ryan, J. P., & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement, 5,* 209–217.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447–467.

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

Slakter, M. H. (1967). Risk taking on objective examinations. *American Educational Research Journal, 4,* 31–42.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295–308.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Author's Address

Send requests for reprints or further information to Richard R. Reilly, Department of Management, Stevens Institute of Technology, Hoboken NJ 07030, U.S.A.