Journal of Educational Measurement
Spring 1998, Vol. 35, No. 1, pp. 69–81

Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method

James C. Impara and **Barbara S. Plake**
University of Nebraska–Lincoln

The Angoff (1971) standard setting method requires expert panelists to (a) conceptualize candidates who possess the qualifications of interest (e.g., the minimally qualified) and (b) estimate actual item performance for these candidates. Past and current research (Bejar, 1983; Shepard, 1994) suggests that estimating item performance is difficult for panelists. If panelists cannot perform this task, the validity of the standard based on these estimates is in question. This study tested the ability of 26 classroom teachers to estimate item performance for two groups of their students on a locally developed district-wide science test. Teachers were more accurate in estimating the performance of the total group than of the "borderline group," but in neither case was their accuracy level high. Implications of this finding for the validity of item performance estimates by panelists using the Angoff standard setting method are discussed.

In the typical judgmental standard setting study, a panel of subject matter experts (SMEs) is asked to make predictions of how certain examinees (often the minimally competent) will perform on the items on a test. The panelists, or judges, typically are experts in the specialized subject field of the examination.

Jaeger (1991) identified eight qualifications that characterize SMEs for a standard setting study. These qualifications include (a) excelling in their domain of specialization, (b) conceptualizing broad patterns of knowledge in their domain, (c) performing (problem solving) rapidly in their domain, (d) processing at a deeper conceptual level in their domain than novices, (e) analyzing problems in their domain qualitatively, (f) employing strong self-monitoring skills, (g) judging problem difficulty more accurately than novices, and (h) having a more complex semantic memory than novices.

Standard setting methods have been classified in a variety of ways. Kane (1994) classified standard setting methods as being test centered or examinee centered. Although both classifications rely on judgments of SMEs, an essential feature of test-centered methods (e.g., Angoff, 1971; Ebel, 1972) is that they assume the ability of panelists to predict accurately the item performance of certain examinees. In the Angoff method, experts are required to estimate the proportion of examinees in the target population, usually minimally competent candidates, who will answer each of the test questions correctly.

The authors would like to acknowledge Carla Noerrlinger of the Omaha Public Schools Research Division for her assistance in the administration of the study and some data analyses. Requests for reprints should be sent to the first author.

Although Jaeger does specify that the judges should be able to judge problem difficulty more accurately than novices, he does not clarify how their judgment of problem difficulty should be quantified. When using the Angoff standard setting method, judges must be able to quantify their estimates of problem difficulty in a probability metric, specifically estimating the probability that a randomly selected minimally competent candidate will be able to answer the item correctly (often operationalized by having the judges instead estimate the proportion of minimally competent candidates who will correctly answer each item).

Thus, in addition to the eight qualifications articulated by Jaeger (1991), two additional qualifications are assumed. These expert judges are assumed to have the skills necessary to (a) conceptualize the minimally competent candidate group by identifying the skills and achievement levels typical of this group of candidates and (b) predict how well such individuals will perform on each item on a test on a probability (or proportion correct) metric.

Jaeger (1989) indicates that little is known about the ability of judges to conceptualize a minimally competent individual or about the extent to which training will bring about in SMEs the sensitivity necessary for formulating this conceptualization.

Reid's (1991) summary of several standard setting studies demonstrated that panelists are not universally competent in estimating item difficulty. Shepard (1994) further demonstrated that trained judges systematically erred in their estimates of item performance by overestimating examinee performance on difficult items and underestimating examinee performance on easy items. Other studies, outside the area of standard setting, also provide evidence that the task of estimating item difficulty is difficult for judges (Bejar, 1983; Lorge & Kruglov, 1953; Thorndike, 1980). In these studies it was found that, in general, judges are able to rank order items accurately in terms of item difficulty, but they are not particularly accurate in estimating actual levels of examinee performance.

In essence, we believe that Jaeger's (1991) criteria for identifying experts are insufficient for identifying judges who are qualified to serve on panels using a judgmental standard setting method such as the Angoff method. Specifically, Jaeger does not include as a criterion that experts must be able to conceptualize the target candidates, nor does he indicate that experts should be able to estimate item difficulty. Being an expert in Jaeger's sense does not mean that a panelist can provide accurate item performance estimates.

In the present study we asked individuals who met Jaeger's criteria to perform tasks similar to the tasks judges perform in standard setting studies. Our objective was to determine if these judges are able to perform the tasks asked of them. Specifically, we used classroom science teachers as SMEs. We asked these teachers to make item-level predictions about their students' performance on an end-of-year examination. These teachers were asked to estimate the performance levels of students in their own classrooms who were "borderline passing" — defined as being D/F students.[1] We also asked the teachers to indicate the performance level of their "class as a whole."

The principal focus of the study was to test the basic assumptions of the Angoff method: that experts are able to (a) conceptualize the minimally competent exam-

inee group by identifying the skills and achievement levels that define this group and (b) make performance estimates for this examinee group. Because the teachers in our study were very familiar with their own students, and because the teachers were also familiar with the nature of the criterion variable (the test), we felt they could serve as a model population to test the validity of these assumptions associated with the Angoff method. If these teachers were unable to make accurate predictions for target groups of students in their classrooms using a familiar test, then questions could be raised about whether the tasks posed by the Angoff method are realistic for experts who are serving as panelists in a typical standard setting study. In the typical standard setting study, panelists often do not have the advantage of high levels of knowledge of the candidate group or much prior exposure to the test.

Method

Participants

Twenty-six sixth-grade science teachers from 10 elementary schools in a large Midwestern school system participated in this study. These teachers were randomly selected from the 150 sixth-grade science teachers in the district. They had, on the average, 9.5 years of service in teaching sixth-grade science (SD = 8.0). The range in science teaching experience was from a low of 1 year (twelve teachers had 5 or fewer years of experience as science teachers) to a high of 25 years (nine teachers had 15 or more years experience as science teachers). In terms of overall teaching experience, the mean was 13.7 years (SD = 8.5) with a range from 3 to 32 years of experience (five teachers had 5 or fewer years of experience, and ten had over 20 years of experience as teachers). There were 5 male and 21 female teachers involved in the study.

Instruments

The school system operates an outcomes-based assessment program, called the Benchmark Assessment Program. For each course in the district, teacher teams identify the expected learning outcomes. These outcomes are subjected to a review and validation process and are adopted by the district when the teachers concur that these outcomes represent the most important learning objectives for the course. The sixth-grade science benchmark test consists of 50 multiple-choice items measuring a total of four expected learning outcomes. Figure 1 shows these learning outcomes and several illustrative items from the test. The sixth-grade science benchmark test had been in use for at least four years prior to this study. The K-R 20 reliability for this 50-item test was estimated to be 0.92.

Procedures

Prior to the regular administration of the sixth-grade science benchmark test, the 26 teachers were contacted and informed of the additional tasks they were being asked to complete as part of the 1994 administration of the test. The directions to these teachers were as follows:

Learning Outcome #1:
Demonstrate competency in measuring skills using the metric system.

a. length **d. mass**
b. area **e. temperature**
c. volume

Illustrative item:
A pencil is found to be 20 units long. Which unit was probably used to measure the pencil?
a. centimeters
b. millimeters
c. feet
d. inches

Learning Outcome #2
Identify and classify matter according to its chemical and physical properties.

a. mixtures **d. mass**
b. solutions **e. temperature**
c. volume

Illustrative item:
The core or center of an atom is called the
a. nucleus
b. electron
c. neutron
d. proton

Learning Outcome #3
Discuss the development of space technology and the benefits to daily life.

a. rockets **d. space stations**
b. satellites **e. space shuttles**
c. space probes

Illustrative item
Astronauts train underwater to prepare for what special condition experienced in space?
a. lack of oxygen
b. temperature control
c. weightlessness
d. bone and muscle weakness

Learning Outcome 4
Recognize the importance of ecology and the need for environmental protection.

Illustrative item
Industrial waste high in the atmosphere reacts with water to form a harmful substance which is then carried to earth by rain. This kind of pollution is called
a. hail
b. acid rain
c. thermal
d. fallout

FIGURE 1. *Expected learning outcomes measured on the sixth-grade science benchmark assessment and illustrative items*

We would like you to make estimates of how students in your sixth-grade science class will perform on the items that make up the ... Benchmark Assessment in Science. We would like you to make these estimations for two different groups of students: once for those students in your class who are just barely passing the class (the borderline D/F student) and second for the class as a whole. For each item, you will be asked to make two projections:

(1) how you expect the "borderline" students to perform on the item, and

(2) how you expect the class as a group to perform on the item.

Here's how we'd like you to make these projections. Imagine 100 students who are just like the typical "borderline D/F" student in your sixth-grade science class. Estimate how many of those 100 students you expect to answer the item correctly. You are to do this for each of the 50 items in the Grade 6 Science Benchmark Test. Likewise, when making projections of how the class as a whole will do on the items of the test, your estimations should be based on your best guess of how many of a class of 100 typical students you would expect to get the item correct.

Teachers recorded their predictions on a form provided by the researchers. After completing this task, teachers returned their rating forms to their instructional supervisor, the curriculm leader for elementary science in the district, who made the teachers' responses available to the researchers. All estimates were made prior to the administration of the test.

Administration of sixth-grade science benchmark test. During May, 1994, all sixth-grade science students in the school system were administered the science benchmark test. Students used a separate, machine-scorable answer sheet to record their answers to the 50 multiple-choice questions. Before sending these answer sheets to the district office for processing, the teachers in this study were instructed to assign a grade in sixth-grade science to each of the students in their classes and to record this grade in a designated place on the student's answer sheet. Thus, teachers did not know the test performance of their students prior to assigning a science grade.

Analyses of accuracy. The researchers, after obtaining the teachers' performance estimates and the results from the administration of the science benchmark test, investigated the degree of relationship between the actual and predicted difficulty of the test questions, for the borderline (D/F grade) student group and for the total group. Analyses were conducted for all the teachers, and also for teachers as a function of the years of experience teaching science in this school district.

In addition, analyses were performed to investigate teachers' level of accuracy in predicting their students' performance on the sixth-grade science benchmark assessment as a function of the actual difficulty level of the items. Three levels of accuracy were defined, as follows: Overestimates were those estimated item difficulties (p values) that were more than .10 over the actual p value; underestimates were those estimates that were more than .10 under the actual p value; and accurate estimates were those estimates within .10 of the actual p value.[2] These classifications were then contrasted as a function of the actual difficulty level of the items. The teachers' predictions—both for their D/F students and for their total groups of students—were classified using this system.

Together, these analyses address the accuracy of these teachers in making performance estimates for their students. Overall level of accuracy is investigated through the analyses of accuracy of item performance estimates by the teachers for their borderline (D/F grade) students and for their total student groups. Whether accuracy is influenced by expertise is considered by looking at accuracy as a function of years of science teaching experience. Finally, the level of accuracy in item performance estimates as a function of level of item difficulty is investigated by consid-

ering the relationship between item difficulty and discrepancy between actual and estimated performance at the individual teacher level.

Results

All 26 teachers completed the tasks. All together, data from these teachers' 724 students who took the sixth-grade science benchmark test were used in this study. Actual item performance for the 95 students who were assigned a D or F grade by their teachers was determined, as was the performance for the total group of 724 students. On the 50-item test the average score for the 724 students was 32.69 (*SD* = 14.89); average performance for the 95 D/F students was 22.52 (*SD* = 15.21). Therefore, students identified as D/F students were, on the average, lower performers on the science benchmark test, which lends credibility to the accuracy of the teachers' classification of these students as "borderline" passing and therefore low-achieving students.

The actual item difficulty values, for the D/F students and for the total group of students, were compared to the teachers' predictions of item performance for the borderline D/F and total groups, respectively. For the total group, teachers predicted students would answer slightly more than 36 items correctly (36.34); actual group performance was 32.69. However, when asked to estimate item performance for their D/F student group, teachers underestimated their performance by 9.51 score points (estimated mean was 13.01, actual mean was 22.52).

Tables 1 and 2 contain information about the level of accuracy of these teachers as a group in making item performance estimates for the borderline students and the total group of students. Table 1 shows summary information about discrepancies between predicted and actual item difficulty across teachers for the D/F students and also for the total group of students. For the total group, the median difference in item performance was 0.087; the 25th percentile value was 0.015, and the 75th percentile value was 0.170. By contrast, when estimating the performance of the D/F students, these teachers' median difference between actual and predicted performance was −0.216; the 25th percentile equaled −0.283, and the 75th percentile equaled −0.069.

TABLE 1
Average difference between predicted and actual (predicted – actual) item difficulty for borderline and total student groups

	Borderline students	Total group
Mean	–0.167	0.092
SD	0.16	0.10
Percentiles		
25th	–0.283	0.015
50th	–0.216	0.087
75th	–0.069	0.170

TABLE 2

Each teacher's average difference, across the 50-item test, between estimated and actual performance (predicted – actual) for students classified as D/F and for the total group, sorted by magnitude of difference for the D/F student estimates

	Indexes of accuracy			
	D/F students		Total group	
	Mean[a]	*SD*[b]	Mean	*SD*
	–0.391	0.149	0.137	0.147
	–0.363	0.156	0.162	0.134
	–0.338	0.182	–0.154	0.273
	–0.310	0.137	0.087	0.124
	–0.291	0.155	–0.001	0.158
	–0.290	0.139	0.153	0.160
	–0.290	0.149	0.000	0.141
	–0.275	0.188	0.042	0.170
	–0.273	0.130	0.004	0.140
	–0.259	0.126	0.006	0.136
	–0.257	0.129	0.014	0.138
	–0.254	0.145	0.016	0.146
	–0.246	0.146	–0.086	0.176
	–0.185	0.145	0.086	0.124
	–0.185	0.144	0.211	0.133
	–0.128	0.174	0.080	0.162
	–0.126	0.199	0.239	0.165
	–0.094	0.181	0.106	0.129
	–0.083	0.183	0.042	0.170
	–0.073	0.234	0.211	0.113
	–0.065	0.192	0.070	0.174
	0.027	0.132	0.177	0.135
	0.055	0.159	0.242	0.141
	0.097	0.154	0.206	0.191
	0.104	0.210	0.210	0.145
	0.169	0.212	0.130	0.146
Mean	–0.167		0.092	
SD	0.157		0.102	

[a]Mean difference between estimated and actual item difficulty across all 50 items; values close to 0 represent higher levels of accuracy.

[b]Standard deviation of differences between estimated and actual item difficulties across the 50 items.

Table 2 provides, for each teacher, the average differences between estimated and actual performance across the 50 items (sorted by degree of difference for the D/F students) both for the D/F students and for the total group. The teachers systematically underestimated the performance of the borderline passing students (the D/F group). Only 5 of the 26 teachers had more overestimates than underestimates for the D/F students, as noted by a positive difference between estimated and

actual item performance. However, this was reversed for the total group of students. All but three teachers had a positive mean difference between their estimated *p* values and the actual *p* values for the total group, which shows that they tended to overestimate the performance of the total group. Twenty of the teachers were more accurate for the total group than for the D/F group, and most of these were substantially more accurate. Therefore, the data in Tables 1 and 2 reinforce that these teachers systematically underestimated the performance of the borderline (D/F) students and overestimated the performance of the total group.

The correlation (across items) between actual performance and estimated performance equaled 0.78 both for the total group and for the D/F students. Thus, as in previous studies, the teachers' rank ordering of item difficulty was moderately accurate, even though their precision in estimating item difficulty was not.

These accuracy values did not seem to vary as a function of years of experience in teaching sixth-grade science in the school district. The correlation between difference in actual and predicted item performance and years of science teaching experience in the school district was –0.08 for the total group and –0.23 for the D/F group. Neither of these values is significantly different from zero.

Tables 3 and 4 summarize the categorization of teachers' accuracy in predicting item performance estimates. Item performance estimates were classified as underestimates if they were more than .10 under the actual student performance, accurate estimates if they fell within .10 of the actual performance level, and overestimates if they were more than .10 over the actual student performance. These values are presented for the D/F students and also for the total student group. Across the total item performance estimates, for the D/F students, the teachers made 858 underestimates, 148 overestimates, and 294 estimates that were within .10 of their actual value. By contrast, for the total group, these teachers provided 151 underestimates, 528 accurate estimates (falling within .10 of actual), and 621 overestimates. Again, these data are consistent with the earlier conclusion that these teachers tended to underestimate the performance of the borderline D/F students and slightly overestimate the performance of the total group of students.

Tables 3 and 4 also display the relationship between the level of item difficulty for the group and the overall degree of accuracy of teacher item performance estimates, classified as underestimates, accurate estimates, or overestimates using the criteria defined above. In these tables, items are classified as easy, moderate, or hard. Items classified as easy had a proportion correct equal to or greater than .67; moderate items had *p* values between .34 and .66 (inclusive), and hard items had *p* values less than .34.

When teachers focused their item performance estimates on the borderline D/F students, 75% of the 52 estimates made for difficult items were underestimates, whereas only one was an overestimate of item difficulty (i.e., actual deviated from predicted by more than 0.10). For items classified as easy, 35% of the teachers' estimates were underestimates of item difficulty, 43% accurate estimates, and 22% overestimates.

By contrast, when these teachers made item performance estimates for difficult items for their class as a whole, only one estimate was an underestimate; virtually all (96%) were overestimates. For the estimates of the easy items for the total

TABLE 3
Frequency of teachers' performance estimates at different levels of accuracy for items of different levels of difficulty for students classified as D/F

	Actual proportion correct			
Type of teacher estimate	< .34	.34–.66	> .66	Total
Overestimate	1	84	63	148
Accurate estimate	12	160	122	294
Underestimate	39	718	101	858
Total	52	962	286	
Number of items	2	37	11	

Note. Overestimate = an estimate more than .10 over the group's actual *p* value; accurate estimate = an estimate within .10 of the group's actual *p* value; underestimate = an estimate more than .10 under the group's actual *p* value.

TABLE 4
Frequency of teachers' performance estimates at different levels of accuracy for items of different levels of difficulty for all students

	Actual proportion correct			
Type of teacher estimate	< .34	.34–.66	> .66	Total
Overestimate	50	347	224	621
Accurate estimate	1	134	393	528
Underestimate	1	39	111	151
Total	52	520	728	
Number of items	2	20	28	

Note. Overestimate = an estimate more than .10 over the group's actual *p* value; accurate estimate = an estimate within .10 of the group's actual *p* value; underestimate = an estimate more than .10 under the group's actual *p* value.

group, 15% were underestimates, 54% were accurate, and 31% were deemed overestimates.

Therefore, the percentages of underestimates, accurate estimates, and overestimates for hard and easy items varied substantially depending on whether the teacher was making item performance estimates for the borderline or total group of students. Thus, although there was systematic variation in the teachers' item performance estimates when focusing on either the borderline or the total group, their pattern across hard and easy items was not consistent across these groups.

Unlike Shepard's (1994) findings, these results did not show a consistent variation in accuracy of prediction simply as a function of item difficulty. That is, these teachers did not systematically overestimate (or underestimate) performance on easy items or overestimate (or underestimate) performance on hard items regardless of target group.

Instead, the systematic over- and underestimation was a function of the teachers' perception of the ability level of the students for whom the estimate was provided. Teachers tended to underestimate for the D/F group and to overestimate for the

total group without regard to item difficulty. Of the 1,300 estimates made for the D/F group (26 teachers for 50 items), 22.6% were accurate (i.e., within .10 of the actual *p* value), and 66% were underestimates (i.e., more than .10 below the actual *p* value). For the total group, however, the teachers estimated accurately 41% of the time and overestimated 48% of the time. Expectedly, for the D/F students, the teachers were most often accurate for the 11 items that were classified as easy for this group (items with *p* values greater than or equal to .67). Similarly, the teachers were most accurate in their estimates of 28 easy items for the total group.

It should be noted, however, that due to the limitation of the proportion-correct score scale (which, by definition, is constrained to a range of 0.00 to 1.00) and the realities of floor and ceiling effects in the total score metric, there is less chance for difficult items to present underestimates (differences in actual and estimated proportion correct in excess of –0.10) and for easy items to yield overestimates. These constraints should be kept in mind when interpreting the results of this study.

Conclusions

In this study we examined the extent to which teachers who are familiar with their students could estimate their students' performance on a district-wide science test the teachers had administered in their classes for the past several years. Teachers estimated performance on the 50-item test for their entire class and for the students classified as borderline passing (D/F students). The purpose of the study was to examine the underlying assumptions of judgmental standard setting methods such as the one proposed by Angoff (1971). Such methods assume that judges can conceptualize an appropriate target group of examinees and also estimate accurately item performance by the identified group(s) of examinees.

In testing these assumptions, we examined (a) whether teachers' identification of the borderline group of students was consistent with the performance of the students identified as borderline when compared with the total group of students and (b) the extent to which teachers would be able to estimate actual student performance both for a borderline group of examinees and for the typical student. In addition to the major focus, we also examined (c) whether there were systematic differences in the accuracy of performance estimation by type of student group (borderline or not) and if this was influenced by difficulties of items (hard, moderate, or easy).

The conclusions associated with the major purpose of this study are mixed. The students who were identified as borderline performed substantially lower than students who were not identified as borderline. This suggests that these teachers' can, at least to a limited degree, discriminate between low-performing students and others. That is, these teachers can identify, with some degree of accuracy, individuals whose performance is consistent with performance that might be expected from those who belong to the target group identified as the minimally competent (defined for the purposes of this study as the "borderline D/F students").

However, these teachers also markedly underestimate how these low-performing students would actually perform on the test. In fact, the teachers' estimates of the average proportion correct for the D/F students were for the most part quite inaccurate. Of 1,300 estimates made for this group of students, only 294 (23%)

were accurate (defined as within .10 of actual item performance). More importantly, the errors of estimation were not balanced around these accurate estimates; item performance was systematically underestimated. If this were a standard setting study, results such as these would raise serious questions about the validity of the standard.

Kane (1986, 1994) suggested that one way to examine the validity of a standard is to compare, for examinees who score at or near the passing score, actual examinee performance on each item with the performance estimates for each item, averaged across judges. The *p* values on each item for this group of examinees should be similar to the average predicted *p* value across judges. Rather than using Kane's criterion, we had teachers identify their science students whose competence was borderline (i.e., D/F). We found that the performance of the borderline group of students was not consistent with teachers' performance estimates for these students. Students in this group scored systematically higher than the teachers predicted for almost all items.

Teachers, on the average, were somewhat more accurate in their estimates of the total group of students. Some teachers were, on average, extremely accurate in their performance estimates of the total group of students. Unfortunately, the teachers who were extremely accurate for the total group were not equally accurate in their performance estimates for the borderline group of students.

One of our expectations was that teachers who were accurate for the total group would also be accurate for the borderline group. This might permit using information about accuracy in making total group item performance estimates to screen judges for a standard setting study. If teachers were equally accurate for the total group and for the borderline group, then they could be asked to estimate performance for the total group (for which data are often available), and those who were most accurate would be most defensible as judges for setting the standard for minimal competency. Unfortunately, for this sample of teachers there was little correspondence between estimating performance for the total group and doing so for the borderline group.

We believe that these teachers represent the characteristics of experts as described by Jaeger (1991). Moreover, unlike many panelists in standard setting studies, these teachers were very familiar both with the examinees whose performance was estimated and with the test used to measure performance. It is clear to us that these teachers were not able to perform accurately the task of estimating item performance for the borderline group of students. We are not arguing that Jaeger's list of characteristics is inappropriate, only that it is incomplete as a set of criteria for selecting panelists. We also suggest that although it may be true that experts are more capable than novices of identifying minimally competent candidates and of classifying the difficulty of problems, the assumption that experts are able to perform the tasks required in the Angoff method is not necessarily accurate and needs to be verified.

Further, our results were consistent with previous research (Bejar, 1983; Lorge & Kruglov, 1953; Thorndike, 1980) suggesting that estimating item difficulty accurately is quite difficult. Thus, asking judges to perform this task in a standard setting context may be unreasonable. Our findings were not consistent, however,

with those of Shepard (1994), who found that judges systematically overestimated performance on difficult items and underestimated performance on easy items.

The most salient conclusion we can draw from this study is that the use of a judgmental standard setting procedure that requires judges to estimate proportion-correct values, such as that proposed by Angoff (1971), may be questionable. The teachers in this study performed the estimation task in such a way that if their performance estimates were used to set a standard, the validity of the standard used to identify borderline students would be in question. If teachers who have been with their students for most of the school year are unable to estimate student performance accurately using a test that is familiar to them, how can we expect other judges who may be less familiar with examinees to estimate item performance on a test those judges may never have seen before?

Notes

[1]The use of this characterization is a modification of Nedelsky's (1954) representation of students with "borderline knowledge between F and D" (p. 5).

[2]The selection of ±.10 was based on the observation that most teachers' estimates were divisible by .5 (e.g., .60, .65), and we felt that any estimate that was within rough rounding error was sufficiently close to be considered accurate (e.g., .74 and .65 both round to .70).

References

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.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303–310.

Ebel, R. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Washington, DC: American Council on Education.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3–6, 10.

Kane, M. T. (1986). *The interpretability of passing scores* (Tech. Bulletin No. 52). Iowa City, IA: American College Testing Program.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Lorge, I., & Kruglov, L. K. (1953). The improvement of the estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34–46.

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

Reid, J. B. (1991). Training judges to provide standard-setting data. *Educational Measurement: Issues and Practice, 10*(2), 11–14.

Shepard, L. A. (1994, October). *Implications for standard setting of the NAE evaluation of NAEP achievement levels.* Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.

Thorndike, R. L. (1980). *Item and score conversion by pooled judgment.* Paper presented at the Educational Testing Service Conference on Test Equating, Princeton, NJ.

Authors

JAMES C. IMPARA is Director, Buros Institute for Assessment Consultation and Outreach, 135 Bancroft Hall, University of Nebraska–Lincoln, Lincoln, NE 68588-0353; jimpara@unl.edu. *Degrees*: BA, MS, PhD, Florida State University. *Specialization:* applied measurement.

BARBARA S. PLAKE is Director, Buros Center for Testing, 135 Bancroft Hall, University of Nebraska–Lincoln, Lincoln, NE 68588-0353; bplake@unl.edu. *Degrees*: BA, University of Colorado; MA, PhD, University of Iowa. *Specialization:* educational measurement.