Florida Journal of Educational Research
Fall 1993, Vol. 33,
No. 1, Pp. 21-30.

Scoring Classroom Achievement Tests: What To Do with the Hard Items?

Roger E. Wilk
University of South Florida

ABSTRACT. This study compares the results of applying two commonly used methods of adjusting classroom tests when items are found to be too difficult: (1) dropping difficult items or (2) adding bonus points to the original score. Undergraduate teacher education students in a required measurement class were given the same five achievement tests during the fall (n = 54) and spring (n = 54) semesters. Four methods of adjusting students' scores were applied: two methods dropped items from the test based on the difficulty value and rescored the tests, and two methods added a bonus percent to the unadjusted total score. Although correlations among semester percentage grades for the different methods were all above .97, only the addition of bonus points maintained the order of the students on the original test. The agreement among the methods in assigning letter grades (90 = A, etc.) varied from 13 to 93%. The effect of dropping items on the content validity and the reliability varied among the unit tests, depending on the characteristics of the items dropped.

When a significant proportion of the students in a class, say 50 or 60%, answer a number of test questions incorrectly on a classroom achievement test, instructors typically make adjustments. Two types of adjustments are common. The instructor may drop the "hard" items, rescore the test, and then base the grade on the remaining items. Alternatively, the instructor may add "bonus" points to the test score to compensate, in a manner, for the number of "too difficult" items. This study compares the effect of these two types of adjustments on the quality of the test, its content validity and internal consistency reliability, and on the percent and letter grades assigned the students.

Although the development of standardized tests has evolved into a highly technical, thoroughly studied process, classroom instructors probably use little of the theory or technology to study their own testing procedures. There is a growing number, however, who do have tests electronically scored, and get the reports of scoring, including basic item statistics. The "what to do with the hard items" problem occurs when the instructor reads the scoring report and judges the test to have too many hard items resulting in too many low scores. The adjustments that teachers make because there were too many hard items include assumptions about the reasons for the low scores: (1) the students did not study, (2) the instruction was not sufficiently effective, or (3) the quality of the test items was lacking. Whatever the specific reason(s), which would be difficult to isolate, the teacher is unlikely to be willing to assign grades considered to be disproportionately low when "teacher error" could have made a significant contribution to the low scores. If the teacher retains all of the items, the assumption is that the test and instruction were not the main factor in low scores. Rather, it was the lack of student preparation that was the principal problem. If the teacher deletes the items and rescores the test, the assumption would seem to be that the instruction and/or the test were culpable. If the teacher decides to retain the difficult items and give bonus points to the entire class, then the effect is to give extra weight to those who answered the hard items correctly without seemingly disadvantaging those who did not answer the hard items correctly.

This study compares the two types of adjustments, deleting items or adding bonus points, with the alternative of making no adjustment.

The Setting and the Data

Five objective achievement tests were given to undergraduate teacher education students in a required measurement class during the fall semester, 1990, and the same tests were repeated as a part of the regular course instruction during the spring semester, 1991. Each of the achievement tests covered one of the five instructional units in the course. Table 1 is the Table of Specifications for the five unit tests showing the content and the categories of learning outcomes. As specified by Gagne´ (1984), "state procedures" and "interpret/generalize" are considered to be categories of verbal information, and "classify examples" and "solve/apply" are intellectual skills. Fifty percent of the 24 objectives and 47% of the 147 items required the students to make interpretations and generalizations about the measurement principles considered in five units. Over 40% of the items and objectives required students to use principles and generalizations to solve problems. On the content dimension, about 20% of the course considered how learning outcomes are characterized and measured, and another third of the course content focused on the development and analysis of objective items and tests to measure

Table 1
Table of Specifications for Five Unit Tests for an Undergraduate Measurement Course

		Categories of Learning Outcomes									
		State Procedures		Interpret/ Generalize		Classify Examples		Solve/ Apply		Total	
Test	Measurement Content	Itm	Obj	Itm	Obj	Itm	Obj	Itm	Obj	Itm	Obj
1	Classroom Attitudes	6	1	12	2					18	3
2	Learning Outcomes	6	1	6	1	24	3			36	5
3	Test/Item Development			12	2			16	2	28	4
	Test/Item Analysis			6	1			19	3	25	4
4	Process/Product Evaluation			18	3	6	1			24	4
5	Standardized Tests			20	3			6	1	26	4
	Total Items & Objectives	12	2	74	12	30	4	41	6	157	24
	Percent Items & Objectives	8	8	47	50	19	17	26	25	100	100

those outcomes. Measuring attitudes (unit 1) and measuring learning process and product outcomes (unit 4) comprised about 30% of the course. One sixth of the test objectives and items (unit 5) asked students questions about standardized test scores.

Fifty-eight students were enrolled in the two sections of the course in the fall semester, and 60 were enrolled in the two spring semester sections. Fifty-four students completed the course each term. In both semesters those enrolled were primarily elementary education majors, special education majors, and majors in secondary education subject matter fields.

The Method

To study the effect of the two general approaches to dealing with hard items, two examples of each approach were devised. These four methods were compared with a "let well enough alone" method, method A, which consisted of scoring and retaining all of the items developed for each test. Methods B and C adjusted the total score on each of the five unit tests by dropping the items with low *p*-values

and rescoring the tests. Method B dropped items with *p*-values equal to or less than .40, and method C dropped items with *p*-values equal to or below .50. Methods D and E added bonus points to the original scores, the percent of items answered correctly. Method D added 5% to the original score and method E added 7%. The basis for setting the *p*-value and bonus point criteria was arbitrary. Because the intent of the study was to assess the effect of these methods in typical instructional circumstances, each of these methods was applied separately in the two semesters. Although the tests in the two semesters contained the same items, the results of applying the criteria were different. The results for the spring semester may be considered a replication of the fall semester analysis.

The effect of dropping low *p*-value items on the content validity was studied by relating those items dropped to the Table of Specifications for the tests. The effect of the methods on the reliability was studied by computing Cronbach's alpha for methods A, B, and C. Because adding constants to the scores in methods D and E would not change the reliability of the original test, the reliability of method A would be the reliabilities of methods D and E as well.

Results and Conclusions

Table 2 shows the effect of applying the two *p*-value criteria for dropping hard items on the content of each test objective and on each test. The effects are certainly not uniform. Some objectives were obviously more difficult, or were comprised of more difficult items than others. Unit 2, which focused on the nature of learning outcomes and how these outcomes are sequenced in a learning task analysis, was clearly the most difficult. In each of the five tests there were objectives that lost from one third to one half of their items when the $p < .50$ criterion for deleting items was applied. Small differences between the two semesters in the effect of dropping items can be attributed to differences in the students and variations in the instruction, although the plan for the course remained the same.

Table 3 shows the item and test data that result from applying methods A, B, and C to each of the five tests during the two semesters. Deleting low *p*-value items had the expected and obvious result of increasing the average *p*-value for each of the tests; the tests became easier. The effects on the reliabilities of the tests were variable. Because the criteria for dropping items did not include a consideration of the item discrimination statistic (*d*), there could be no direct, predictable effect on the reliability. The reliability could be affected by retaining hard items if there were a tendency for students to guess at the answers more often. Reliability was not affected by dropping items with low *d*-values, but decreased when items with reasonable *d*-values were dropped (data not reported).

Table 2
Results of Adjusting Tests by Deleting Items with p-values Below .50 by Objective and Test

				Deleted			
				Fall		Spring	
Test	Obj	Learning Category	Items	.40	.50	.40	.50
1	1	State generalizations	6	1	1	0	0
	2	State procedures	6	2	2	2	3
	3	Interpret data	6	0	0	0	0
		Total Test 1	18	3	3	2	3
2	1	Classify learning outcomes	8	2	3	2	3
	2	Classify verbal information	8	0	0	0	0
	3	Classify intellectual skills	8	2	3	1	4
	4	Sequence skills	6	0	0	0	0
	5	Interpret learning outcomes	6	1	1	1	1
		Total Test 2	36	5	7	6	8
3	1	Interpret Table of Specifications	7	1	1	0	0
	2	Determine item validity	10	0	1	0	0
	3	Interpret task analysis	5	0	2	0	1
	4	Apply item format rules	6	0	0	0	1
	5	Interpret data analysis	6	0	1	1	2
	6	Solve item analysis problems	7	0	1	0	0
	7	Solve data analysis problems	6	0	0	0	0
	8	Solve test analysis problems	6	1	1	0	1
		Total Test 3	53	2	7	1	5
4	1	State generalizations	5	0	0	0	0
	2	State generalizations	6	1	1	1	1
	3	Interpret analysis	7	0	0	0	0
	4	Classify examples	6	1	1	0	2
		Total Test 4	24	2	2	1	3
5	1	Interpret std. scores	8	0	3	3	3
	2	State generalizations	6	0	0	0	0
	3	Interpret SEM	6	0	0	0	0
	4	Solve problems	6	0	0	1	1
		Total Test 5	26	0	3	4	4
All	24	Total Items All Tests	157	12	17	14	21

Table 3
Item Difficulty (p) and Reliabilities (alpha) for Five Tests Scored by Three Methods over Two Semesters

		Fall Semester				Spring Semester		
		Method				Method		
		A	B	C		A	B	C
Test 1	*n*	58	58	58	Test 1 *n*	60	60	60
	Items	18	14	14	Items	18	16	15
	Ave (*p*)	.63	.72	.72	Ave (*p*)	.74	.79	.81
	Alpha	.49	.40	.40	Alpha	.56	.45	.38
Test 2	*n*	57	57	57	Test 2 *n*	60	60	60
	Items	36	30	28	Items	36	32	20
	Ave (*p*)	.71	.79	.82	Ave (*p*)	.69	.75	.80
	Alpha	.47	.54	.51	Alpha	.62	.64	.60
Test 3	*n*	57	57	57	Test 3 *n*	55	55	55
	Items	53	51	46	Items	53	52	48
	Ave (*p*)	.70	.72	.74	Ave (*p*)	.73	.73	.75
	Alpha	.72	.72	.70	Alpha	.67	.66	.66
Test 4	*n*	56	56	56	Test 4 *n*	59	59	59
	Items	24	21	21	Items	24	22	20
	Ave (*p*)	.72	.75	.75	Ave (*p*)	.76	.77	.81
	Alpha	.50	.46	.46	Alpha	.57	.53	.53
Test 5	*n*	55	55	55	Test 5 *n*	57	57	57
	Items	26	23	23	Items	26	22	22
	Ave (*p*)	.72	.75	.75	Ave (*p*)	.61	.69	.69
	Alpha	.50	.72	.72	Alpha	.71	.71	.71

Method A: Keep all items
Method B: Keep items if $p > .40$
Method C: Keep items if $p > .50$

Table 4 shows the effect of applying the five adjustment methods to the computation of the students' average percentage grade for the semester. This table shows the means and standard deviations of the percent grades and the correlations among the grades assigned using the five methods. The adjustments did increase the average percentage grade, but the variability was virtually unchanged. Methods B and D had a comparable effect on the grade distribution, and methods C and E produced equivalent results. The correlations indicate that deleting items changed the order of the students, although the change was small. Adding bonus points did not change the order of the students based on the unadjusted test results.

Table 4
Means, Standard Deviations, and Correlations Among the Adjustment Methods for the Two Semesters

	Fall Semester (n = 54)					Spring Semester (n = 54)				
	Method					Method				
Method	A	B	C	D	E	A	B	C	D	E
A	—	.987	.979	1.00	1.00	—	.988	.975	1.00	1.00
B			.990	.987	.987			.992	.988	.988
C				.974	.979				.975	.975
D					1.00					1.00
Mean	70.8	74.9	77.1	75.8	77.8	71.2	75.6	77.9	76.2	78.2
SD	7.4	7.5	7.5	7.4	7.4	8.7	9.0	9.0	8.7	8.7

Method A: Keep all items
Method B: Keep items if $p > .40$
Method C: Keep items if $p > .50$
Method D: Keep all items and add 5% to score
Method E: Keep all items and add 7% to score

Table 5 shows the effect of the five methods on the assignment of letter grades. The standard, an arbitrary though common one, was: 90% and above = A, 80 to 89% = B, 70 to 79% = C, 60 to 69% = D, and below 60% = F. The results are shown by two statistics, the percent of agreement in the grade assigned among the methods, and the

Table 5
Percent Agreement and Correlation Between Five Methods of Assigning Letter Grades (90% = A, 80% = B, 70% = C, 60% = D)

	Fall Semester (n = 54)					Spring Semester (n = 54)				
	Method					Method				
Method	B	C	D	E		Method	B	C	D	E
A: % =	52	24	45	13		A: % =	41	22	35	21
B: % =		72	93	61		B: % =		82	83	80
C: % =			80	85		C: % =			83	91
D: % =				69		D: % =				85
A: r =	.79	.86	.81	.91		A: r =	.85	.89	.86	.90
B: r =		.85	.95	.80		B: r =		.91	.90	.90
C: r =			.88	.90		C: r =			.91	.95
D: r =				.83		D: r =				.93

Method A: Keep all items
Method B: Keep items if $p > .40$
Method C: Keep items if $p > .50$
Method D: Keep all items and add 5% to score
Method E: Keep all items and add 7% to score

correlation (Pearson's r) between the numerical equivalents of the letter grades (A = 5, B = 4, etc.). The percent of agreement statistic indicates the extent to which the two methods assigned the same grades, and the correlation indicates the extent to which the assignment of students' grades maintained the students in a similar rank order. In combination, a low percent of agreement and a high correlation, e.g., method A and method E in the fall semester, indicates that the two methods put the students in the same order, but the assignment of grades was considerably different. Table 6 shows the percent of students assigned each letter grade by each method.

Table 6
Percent of Letter Grades Assigned Using Five Methods of Adjusting for "Hard" Items

	Fall Semester (n = 54)						Spring Semester (n = 54)				
	Method						Method				
Grade	A	B	C	D	E	Grade	A	B	C	D	E
A	0	2	7	6	7	A	0	6	9	9	11
B	11	26	33	22	39	B	13	33	39	30	27
C	41	46	43	50	43	C	46	41	35	41	35
D	41	26	17	22	11	D	30	17	15	19	15
F	7	0	0	0	0	F	11	4	2	2	2
Total	100	100	100	100	100	Total	100	101	100	101	100

Method A: Keep all items
Method B: Keep items if $p > .40$
Method C: Keep items if $p > .50$
Method D: Keep all items and add 5% to score
Method E: Keep all items and add 7% to score

Implications

Although the data are real results from a real instructional program, the limitations are obvious. The particular student population studied and the subject matter of the tests are specific and limited. The arbitrary nature of the adjustment criteria chosen to form the comparison groups limits (or prohibits) any generalizations about other criterion levels or how criteria should be set. When an instructor is making judgments about what to do with hard items, two questions seem of primary importance. First, will the set of items remaining after deleting the hard items adequately represent the original test plan? This, of course, is the content validity question. Second, will changes that occur in the order of the students' test scores be justified? Such changes could, and probably would, affect the percent or letter grades assigned. Unless the hard items are negatively discriminating, what would justify changing the original order of the students' scores? From the comparisons made in this study, it would seem that it would be a better practice to add bonus points to compensate for hard items. The effect would be to give extra points to

those who answered the hard items correctly without penalizing students who missed the items. Rather than jeopardizing the content validity of the test or changing the order of students' achievement scores by dropping items, the method would make adjustments to an often inflexible grading standard.

References

Gagne´, R. M. (1984). Learning outcomes and their effects: Useful categories of human performance. *American Psychologist, 39*, 377-385.

Wilk, Roger E., University of South Florida, FAO100U, Room 295, Tampa, Florida 33606