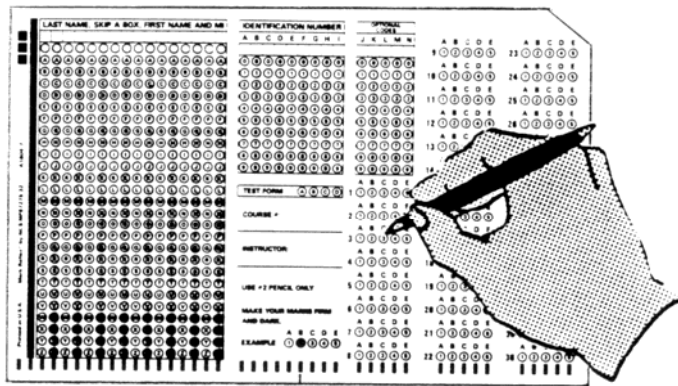


Evidence for NOT Weighting Objective Test Items

EES Memorandum #51



Evaluation and Examination Service
The University of Iowa
(319) 335-0356

Weighting Objective Test Items

The most common method for scoring classroom achievement tests is to weight correct responses +1, and incorrect responses as 0. However, the Evaluation and Examination Service (EES) is often asked by instructors to score tests on which a range of point values has been assigned to either individual items or item options. The assumption underlying differential weighting is that answers to questions are not typically of equal difficulty or “all right” or “all wrong”. Differential weighting allows for students to receive at least partial credit on items for which they don’t know the best answer but supposedly have some knowledge.

On the surface, differential weighting appears to be a reasonable way to evaluate student performance on classroom tests. However, in order for differential weighting to be an effective measurement tool, at least one of the following should occur as a direct consequence of weighted scores:

- ✎ A significant strengthening of the technical qualities of the test through increased reliability and validity.
- ✎ A more accurate ranking of students by test score.

Unfortunately, differential weighting rarely results in more reliable or valid test scores. Most measurement specialists discourage the use of differential weighting because the gain that might be realized is much smaller than the effort required to do the weighting. In the majority of cases, differential weighting results in the same rank order of students as that obtained with simple +1 and 0 scoring. A bibliography has been provided for those readers interested in empirical studies addressing issues related to differential item and response weighting. The remainder of this memo uses final exam test data to underscore the inefficiency of differential weighting.

Data from seven scoring requests (tests 1-4 are from the 1982 memo; tests 5-7 are from 1993 update) is presented in Table 1. Each test was first scored using the weights requested by the instructor, and then rescored using simple +1 and 0 (unit) scoring. A correlation coefficient was computed for each set of differential and unit scores and is shown in column four. For 1993 data, columns five and six report reliability coefficients and mean discrimination indices for differential and unit scoring. These indices were not computed in the original 1982 study.

The high correlation values support the notion that differential item weighting on these tests did not alter the rank ordering of students' scores in any practical way. It is highly likely that students in each case would be assigned the same letter grade using either the weighted or unweighted scores. In addition, data for the 1993 tests show that there were no gains in reliability or mean discrimination when differential weighting was used. Frary (1989) states, "...it would seem that option weighting offers no consistent basis for improving the psychometric quality of test scores unless there is a problem with respect to internal-consistency reliability. These factors along with the cost and complexity of option weighting have contributed to its decline."

Table 2 shows the items from Test VI with assigned weights and item difficulties (p-values %; proportion of examinees answering correctly). Logically, the more difficult the item, the more weight it should be assigned. As can be seen, the most difficult item (#1) and the third most difficult item (#3) were only worth 5 points as compared to items 6 & 7 which were each worth 20 points. Subjectivity in assigning item weights is a severe limitation to differential weighting. Inappropriate weights tend to decrease reliability.

For the typical classroom test there is nothing to be gained either psychometrically or practically by using differential item or item option weighting. If you are considering this option, it is recommended that you spend some time reading the articles listed in the bibliography for a better understanding of the issues involved.

Table 1. Comparison of Differential and Unit Scoring

Test	# of Examinees	# of Items	Differential Weights	Correlation Diff/Unit	Reliability Diff/Unit	X Discrimination Diff/Unit
I	83	41	Rights = 4 Wrongs = -1	.95		
II	50	160	Items 1-40 = 1 Items 141-160 = 3	.92		
III	34	105	Items 1-70 = 1 Items 71-105 = 2	.98		
IV	21	90	Items 1-45 = 1 Items 46-9 = 3	.98		
V	17	60	Rights = 3 Wrongs = -1	.99	0.72/0.78	0.23/0.26
VI	63	7	Items 1-3 = 5 Item 4 = 5 Item 5 = 15 Items 6/7 = 20	.92	0.41/0.48	0.46/0.51
VII	68	22	Rights = 5.66	1.00	0.66/0.66	0.33/0.33

Table 2. Difficulty Values for Weighted Items

Item	Weights	P-value %
1	5	57
2	5	87
3	5	75
4	10	95
5	15	76
6	20	81
7	20	70

Bibliography

- Ebel, R.L., and Frisbie, D.A. (1991). *Essentials of educational measurement* (5th ed.). New Jersey: Prentice Hall.
- Downey, R.G. (1979). Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement*, 3, 453-461.
- Echternacht, G.J. (1976). Reliability and validity of option weighting schemes. *Education and Psychological Measurement*, 36, 301-309.
- Frary, R.B. (1982). Assimilation study of reliability and validity of multiple-choice test scores under six response-scoring modes. *Journal of Educational Statistics*, 7, 333-351.
- Frary R.B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, 2(1), 79-96.
- Wang, M.D., & Stanley, J.C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 40, 663-705.