The Use of Principal Component Analysis to Increase the Ability of Multiple Choice Examinations to Distinguish Among Students*

ROGER E. FLORA AND WALTER H. CARTER, JR.

Department of Biometry,
Medical College of Virginia,
Richmond, 23219

Introduction

Multiple choice examinations are widely used in standard examinations such as the GRE, MCAT, or Medical Board examinations. This type of examination has also gained considerable popularity as a means of examining students on the material presented in a formal course of instruction, especially when classes are large. This popularity is due at least in part to the ease with which such examinations can be graded. Indeed, it is often possible through the use of standard answer forms to have the examination graded entirely by computer (Rosinski and Hamilton, 1966).

The widespread use of the multiple choice examination has led to much discussion of its ability to evaluate examinees' knowledge of the material on which they are examined. The effectiveness of the evaluation depends, of course, on the objectives of the examiner. Various practical and/or philosophical reasons have been given for such evaluations (Karsner, 1937). We shall not discuss these but shall assume that one purpose of administering an examination is to distinguish among examinees with respect to knowledge of a specific subject area. Attention is focused here on the examination's ability to achieve this purpose; in particular, a method of scoring is presented which provides a greater distinction among examinees than the usual methods of scoring.

An objection often voiced against multiple choice examinations is that questions are sometimes stated ambiguously so that two or more choices appear to be equally correct. Neither these questions nor those for which either all or none of the examinees know the correct answer contribute to the distinction among examinees. Indeed, if all questions were of these types, the only variation among test scores would be chance variation due to guessing. On the other hand, questions for which some but not all individuals know the correct answer reflect variation among individuals' knowledge of the material covered on the examination as well as some chance variation due to guessing. Thus, it would seem desirable to eliminate all questions which contribute only chance variation to the test scores. The result would be a set of test scores with a higher variance which is influenced less by chance variation due to guessing.

Pratt and Ingersoll (1968) present a method which attempts to detect questions which contribute little to the distinction among examinees' knowledge. These questions are then eliminated and test scores based only on the remaining good questions are obtained. Their procedure is based primarily on intuition and uses a quite arbitrary criterion for determining "good" questions. However, for the two tests to which they applied the method, they were successful in obtaining test scores based only on "good" questions which had a considerably larger variance than the scores based on all questions.

In this paper we present a method of scoring, based on the statistical technique of principal components, which yields a set of test scores having maximum variance. In addition, it is demonstrated that this method can indicate questions which contribute little or nothing to the distinction among examinees. This procedure is applied to two tests, and

* This investigation was supported in part by Public Health Service Research Grant RR00016-08 from the National Institutes of Health.

the scores obtained are compared with the usual percentage scores. The scores are also compared to those obtained by the method of Pratt and Ingersoll for one of the tests.

Methods

In the description of methods, we shall have need to refer to an individual's score on a single question or item as well as to his score on the entire test. In order that the distinction be clear, the score on the entire test will be referred to as the test score and denoted by T, while the score for a specific item will be referred to as an item score and denoted by S. For the usual methods of scoring, the test score can be written as a linear combination or "weighted" sum of the individual's item scores. Thus

$$T = \sum wS \qquad (1)$$

where w represents the weight assigned to a specific item. For example, a common method of scoring is to assign as the test score for an individual, the percentage of test questions answered correctly. This can be obtained from the above formula by assigning an item score of 1 or 0 depending on whether the answer given is correct or incorrect and by assigning a weight of $w = 100/n$ to each question, where n is the number of questions on the examination.

In order that test scores be comparable, they should be based on the same method of assigning item scores. We shall adopt the convention of assigning an item score of 1 for a correct answer and an item score of 0 for an incorrect answer. The only difference among tests scores obtained by different methods will then be determined by the weights assigned to the specific items.

Since we shall apply the method of Pratt and Ingersoll, as well as our own, a description of their method is in order. They first divide the examinees into an upper and lower half based on the percentage of correct responses for the entire test. An item is then declared to be a "good" question if the percentage of upper half students giving the correct answer is at least eight percentage points higher than the percentage of lower half students answering the question correctly. The test score assigned for an individual is then the percentage of "good" questions answered correctly. This score can be obtained from (1) by assigning all "good" questions a weight of $w = 100/n_g$, where n_g is the number of good questions, and a weight of $w = 0$ to all other questions.

Note that the method of Pratt and Ingersoll constitutes a differential weighting of item scores. This is done in an attempt to obtain a set of test scores with greater variance than the usual percentage scores, thereby making the distinction among students clearer. However, some "good" questions are certainly better than others for purposes of distinguishing among students. Also, some "good" questions are only slightly better than those not considered to be "good" questions. It would seem therefore that a more judicious assignment of differential item weights might yield test scores with an even larger variance. In fact, one might assign item weights in such a way that the variance among test scores is a maximum for all methods of scoring of the type $T = \Sigma\, ws$. Such a set of weights can be obtained from a principal components analysis of the item scores. The weights used are those corresponding to the first principal component of the item scores. A description of principal components analysis can be found in most texts on multivariate analysis (Morrison, 1967; Seal, 1962). Standard computer programs for performing such an analysis are also available at most scientific computing installations.

In assigning item weights, the relationship between item scores for different items should be taken into consideration as well as the variance of item scores for a specific item. For example, if the covariance (and hence the correlation) between two items is large, both items are likely to be measuring the same aspect of knowledge. Consequently, both questions, even though they may be good at distinguishing among individuals, should not be given large weights relative to other questions. The method of principal components takes care of this situation by either assigning a larger weight to the better of the two questions or by assigning intermediate weights to each of the two questions. Thus, the actual value of a weight cannot be taken as a complete indication of a question's utility in distinguishing among students. In general, however, a very small weight will indicate a question that is of little value in distinguishing among examinees.

In the present application, it is desirable that all weights be positive to insure that no examinee be penalized for correctly answering a question. In general, not all weights obtained through the method of principal components are positive. However, a theorem due to Perron (1970) asserts that if all covariances (and thus correlations) between items are positive the weights associated with the first principal component will also be positive. Certainly all correlations between test items should be positive, for a negative correlation would indicate that a correct response to one question tends to be associated with an incorrect response to another. This should only be the case if an incorrect choice is indicated as the correct answer for one of the questions. However, we are dealing with estimates of the true covariance between items, and it is possible to obtain a negative value as an estimate even though the true covariance is positive. This will generally occur only if the true covariance is near zero, which usually

results from one of the items being such that very few if any of the examinees know the correct answer. Thus, negative covariances, and hence negative weights, indicate questions that are of little value in distinguishing among examinees. These questions are therefore assigned a weight of zero. However, since negative weights may be indicative of questions for which an incorrect choice is designated as the correct answer, a closer scrutiny of these questions may be warranted.

The method of scoring based on a principal components analysis of item scores was applied to two tests. The first test was an examination given to 127 first year medical students at the Medical College of Virginia during the 1968–1969 school term. The examination, consisting of 45 multiple choice questions on statistics and mathematics, was given prior to a course in biostatistics. The purpose of the examination was to divide the class into an advanced and elementary section if the students' backgrounds varied greatly. Thus, the examination contained several questions for which few students knew the correct answers. The second test, consisting of 50 multiple choice questions, was the final examination in biostatistics given to 134 first year medical students during the 1969–1970 school term. For purposes of comparison, the method of Pratt and Ingersoll was also applied to the first test.

Results and Discussion

As noted previously, the variation among test scores is a useful index of the scores' utility in distinguishing among examinees. A measure of this variation is provided by the standard deviation of the test scores. For the methods of scoring applied to the first test described above, the standard deviations were 8.8 for the method of percentage scores, 12.8 for the method of Pratt and Ingersoll, and 37.0 for the method based on a principal components analysis of item scores. While the method of Pratt and Ingersoll offers some improvement over the method of percentage scores (the improvement observed here is consistent with that reported in Pratt and Ingersoll, 1968) the method based on principal components offers a substantial improvement over either of the other methods.

The percentage scores and the corresponding principal components scores for the 15 students with the highest and lowest percentage of correct answers on the first test are presented in the Table. In addition to the greater variation among the scores based on principal components, it may be observed that the principal components score is higher than the corresponding percentage score for all except two of the top 15 students. Similarly, the principal components score is lower than the corresponding percentage score for all but one of the bottom 15 students. In

TABLE

Percentage scores and principal components scores for the first and last fifteen students on the basis of their percentage scores on test 1. (A total of 127 students took the examination.)

Top 15 Students		Bottom 15 Students	
Percentage scores	Principal components scores	Percentage scores	Principal components scores
80.0	91.9	44.4	42.7
77.8	87.8	42.2	33.4
73.3	79.6	42.2	41.5
73.3	75.4	42.2	40.8
71.1	71.2	42.2	31.7
68.9	77.3	40.0	24.2
68.9	80.0	40.0	26.7
68.9	67.9	40.0	29.8
68.9	83.3	37.8	34.6
66.7	83.9	37.8	33.5
66.7	79.1	37.8	36.9
66.7	80.6	37.8	20.6
66.7	84.8	37.8	27.7
66.7	64.2	35.6	18.8
66.7	74.4	33.3	36.7
Mean 65.3	78.8	39.4	32.0
Range 13.3	27.7	11.1	23.9

general, therefore, a student near the top (bottom) of the class with respect to the percentage scores will also be near the top (bottom) with respect to the principal components scores. It should be noted, however, that within the top 15 students, the students' ranks based on the two methods of scoring differ considerably, and similarly for the bottom students. This is to be expected since some students are more fortunate than others in guessing the correct answer to non-discriminating questions such as those for which no students know the correct answer.

It may also be observed from the Table that six of the top 15 students received identical percentage scores of 66.7. On the other hand, the principal components scores for these students ranged from 64.2 to 84.8. Thus, where no distinction was possible on the basis of their percentage scores, a fairly wide distinction is possible on the basis of their principal components scores. Further, examples of this type are available from an inspection of the data in the Table.

Application of the method of principal components to the first test considered yielded 13 negative weights. An item analysis of the question corresponding to these weights revealed that the percentage of correct responses to all except one of these questions was quite low. In fact, the percentage was about what would be expected had all students made their choices completely at random. For the other question to which a negative weight was assigned, the percentage of correct responses for the lower half of the class was 46% as compared to 52% for the upper half. Further investigation of the responses to this question indicated that the majority of the class was able to narrow the alternatives to two choices but had to guess between these two. Thus, it is apparent that these 13 questions could contribute nothing to the distinction among students. It would thus appear that it is indeed appropriate, as indicated in the methods section, to assign these questions a weight of zero for purposes of computing test scores. It should also be noted that each of these questions was declared to be a "poor" question and thus assigned a zero weight by the method of Pratt and Ingersoll as well.

Further investigation of the manner in which item weights were assigned by the method of principal components revealed that the larger weights were associated with questions for which the difference between the percentage of correct responses by the upper and lower halves of the class was greatest. For example, the largest weight was assigned to the question having the greatest difference between the percentage of correct responses (71% for the upper half versus 32% for the lower half). Similarly, lower weights were associated with questions for which the differences between the upper and lower halves of the class were small.

Application of the method of scoring based on principal components to the second test described above produced results similar to those for the first test. The standard deviations of test scores was 13.0 for the percentage scores and 38.8 for the principal components scores. The relationship between the magnitude of item weights assigned by the method of principal components and the difference between the percentage of correct responses by the upper and lower halves of the class was also similar.

On the second test, however, no negative weights were encountered. This is apparently due to the fact that there were few if any questions for which no students knew the correct answer. This is to be expected since this was a final examination on which much care was taken not to include questions on material which had not been covered in class. On the other hand, the first test was given in an effort to determine students' backgrounds prior to presenting a course of instruction and consequently contained several questions for which very few students knew the correct answer. This indicates that with careful construction of multiple choice examinations, questions with no discriminating ability can be avoided; however, even then test scores based on a principal components analysis of item scores are markedly superior to percentage scores for distinguishing among examinees.

Summary

A method of scoring multiple choice examinations based on the statistical technique of principal components analysis is described. This test is applied to two tests. The results indicate that the method yields test scores which have a marked advantage over the usual percentage scores in distinguishing among examinees with respect to knowledge of subject matter. In addition, the method provides an indication of whether a question is "good" or "poor" for purposes of distinguishing among examinees.

References

Karsner HT: Philosophical comments on examinations. JAMA 108: 1022, 1937

Morrison DF: Multivariate Statistical Methods. New York, McGraw-Hill, Inc., 1967

Perron O: Zur Theorie der Matrices. Math Ann 64: 248, 1907

Pratt PC, Ingersoll RW: A method for increasing the reliability and validity of multiple choice examinations. J Med Educ 43: 1238, 1968

Rosinski EF, Hamilton DL: Examination procedures as part of a new curriculum. J Med Educ 41: 135, 1966

Seal HL: Multivariate Statistical Analysis for Biologists. New York, John Wiley and Sons, Inc., 1962