PSYCHOMETRIC AND SAS® PROGRAMMING ASPECTS OF OBJECTIVE EXAMINATIONS

W. J. Klein and G. R. Emslie
Ryerson Polytechnical Institute

ABSTRACT

Although objective achievement tests lend themselves to machine scoring, educators and programmers have not fully exploited the possibilities. Alternative methods of presenting and scoring objective tests are reviewed and those worthy of further research are identified. Reliability, the reproducibility of measurements when testing conditions remain constant, is a prime requirement. In general, the longer the test and the less prone it is to random guessing, the more reliable. For objective tests guessing is related to the number of alternative answers. Because testing time is limited, a trade-off is necessary between a long test with few alternative answers per question and a test with fewer questions but more alternatives. Commonly, tests require the student to "choose the single best answer provided". However, it is demonstrated that the probability of a correct guess is typically lower when the test instructions require more than one answer to be selected. Theoretically, such tests are expected to show higher reliability than conventional ones. Whether this expectation is supported is a matter for empirical research. The SAS programming required to implement these innovations is discussed.

INTRODUCTION

Objective examinations generally require test-takers to choose among several provided answers. This form lends itself to the use of optical scanners and computer scoring. Optically scanned test protocols produce a data matrix in which the rows identify the students and the fields contain their responses. Additional fields are used to identify the course, the class section, and the instructor. A preliminary record contains the keyed answers (those defined as "correct" by the test constructor). Although educators and computer programmers have responded to the challenge, they have not fully exploited the technology. Alternative ways of presenting and scoring objective tests hold promise.

TYPES OF ANSWER SELECTION TESTS

Among answer selection tests, the most commonly adopted technique is "multiple choice". However, when the most general application is to situations in which the student makes only a single selection for each test item, the term is a misnomer. Moreover, the very popularity of the "choose one from a number of alternatives" presentation has tended to restrict research. Attention has been paid, for example, to the question of how many alternative answers the examiner should provide per item (Wood, 1977) but there has been little systematic study on the optimal number of selections required of the student.

A simple taxonomy of answer selection tasks is presented in Table 1. Three factors determine the general structure of these tasks:

1. How many alternative answers are provided (factor **a**)?
2. How many of the answers are keyed for credit (factor **k**)?
3. How many answers do the respondents select from the menu of alternatives (factor **s**)?

Thus any testing situation (of an answer selection nature) is described by the three digit code, **aks**. For any value of **a**, the number of alternative answers provided, there are three general task categories:

1. Exact Match tasks in which the student must select all the alternatives keyed for credit and only those alternatives (i.e., **s** = **k**). The conventional "multiple choice" task is more correctly described as an exact choice task involving one selection and one keyed answer.

2. Multiple Choice tasks in which the number of choices the student makes exceeds the number of keyed alternatives (**s** > **k**, thus truly "multiple" choice as opposed to the traditional use of the term). Credit is given if all the keyed alternatives are included in the person's selections.

3. Multiple Key tasks in which more alternatives are keyed than the student is required to select (**s** < **k**). Credit is given if all the student's selections are included in the set of keyed alternatives.

Table 1 Options for Choice Tests With **a** Alternative Answers per Question

k = number of keyed answers	**s = number of answers to be selected**							
	1	2				**a**-2	**a**-1	optional
1	e ✓	mc	mc	mc	mc	mc	mc	
2	mk	e	mc	mc	mc	mc	mc	
	mk	mk	e	mc	mc	mc	mc	
	mk	mk	mk	e	mc	mc	mc	
	mk	mk	mk	mk	e	mc	mc	
a-2	mk	mk	mk	mk	mk	e	mc	
a-1	mk	mk	mk	mk	mk	mk	e	
variable								e

Notes:

Each cell, the combination of the number of alternatives keyed for credit by the test constructor and the number of alternatives selected by the test-taker, represents a potential test. Thus the traditional "multiple choice" test (marked ✓) is only one of many tasks that can be fashioned from a given number of alternative answers.

a = the number of alternative answers provided per question;
k = the number of alternatives (per question) keyed for credit;
s = the (maximum) number of alternative answers that subjects are permitted to select per question.

Variable = the number of keyed alternatives varies from question to question;
optional = the number of alternatives to be selected is at the discretion of the student (assuming the goal is to match the key).

e = exact match tests (**s** = **k**);
mc = multiple choice tests (**s** > **k**);
mk = multiple key tests (**s** < **k**).

The presentation is not exhaustive. Restricted discretional tests might be provided by requiring respondents to select an appropriate subset of alternatives, for example, "at least one and no more than two" or "either two or three".

A basic requirement of any test is that it is reliable. It must measure with acceptable precision and its results must be reproducible (within a tolerable margin of error) if the same people are retested under conditions similar to those prevailing at the original examination. Reliability varies with the subject matter (e.g., factual knowledge is typically measured more reliably than conceptual) and with the personal characteristics (age, ability, motivation, etc.) of the test-takers. These major influences aside, two structural variables known to affect reliability are test length and the test's susceptibility to guessing (Magnusson, 1967). The longer the test the more reliable. The more random guessing, the less reliable. The limited availability of testing time usually forces a choice: the examiner either includes many

questions with relatively few alternative answers (i.e., maximizes test length) or has few questions but provides many alternative answers (i.e., minimizes the probability of correct guesses). "True-false" tests are an extreme version of the former approach. A compromise is exemplified by traditional "multiple choice" tests. Among these, a common requirement is that the student "choose the best answer among five". But if five alternatives are available, the traditional task might not be optimal. As shown in Table 2, three nonconventional presentations are potentially superior because the probability of correctly guessing the answer is reduced.

The empirical research required to determine whether this potential superiority is realized is under way.

SCORING OF ANSWER SELECTION TESTS

Generalizations of conventional objective testing methods present a challenge to the programmer. Consider the following example for which the test instruction is to select all the true and only the true alternatives:

New York city:

a. is situated on the eastern seaboard of the United States.
b. is closer to Toronto than it is to Boston.
c. has a population of over 2 million.
d. has a professional baseball team called the Saints.
e. is the largest city in the world.

Table 2 Options for Choice Tests With Five Alternative Answers per Question

k = number of keyed answers among five	**s = number of answers to be selected by student**				
	1	2	3	4	optional
1	TEST 511 **p** = .20	TEST 512 **p** = .40	TEST 513 **p** = .60	TEST 514 **p** = .80	
2	TEST 521 **p** = .40	TEST 522 **p** = .10	TEST 523 **p** = .30	TEST 524 **p** = .60	
3	TEST 531 **p** = .60	TEST 532 **p** = .30	TEST 533 **p** = .10	TEST 534 **p** = .40	
4	TEST 541 **p** = .80	TEST 542 **p** = .60	TEST 543 **p** = .40	TEST 544 **p** = .20	
variable					TEST 5** **p** = .03

Notes:

Tests are identified by the three digit code, **aks** (an asterisk replaces the digit if the value varies across questions).

a = the number of alternative answers provided per question;
k = the number of alternatives (per question) keyed for credit;
s =the (maximum) number of alternative answers subjects are permitted to select per question.

p =probability that a random guess matches the key; for choice tests with a given number of alternatives per question, the probability is lowest when the number of keyed alternatives varies from question to question and the test-takers are instructed that all, some, or none of the alternatives might be keyed; if the test instructions specify a fixed number of selections, the probability of a guess matching the key is lowest when **s** = **k** = **a/2**.

Variable = the number of keyed alternatives varies from question to question;
optional = the number of alternatives to be selected is at the discretion of the student (assuming the goal is to match the key).

Shaded areas represent situations in which the probability of a random guess matching the key is lower than for the traditional multiple choice test (single key, single choice) with the same number of alternatives.

For the discretional response test, the probability is for the case where every possibility (from no alternatives keyed to all a alternatives keyed) is included.

An examiner might indicate that alternative (a) is the correct answer and request conventional scoring. Only students who select (a) obtain a credit. At some later time, it is discovered that alternative (c) is also correct. The examiner then requests that the scoring system accommodate two keyed alternatives and that students be credited for either response. Later still, the examiner decides that perhaps each alternative answer should be scored. What was originally conceived as one question has now become five true-false questions. Then comes a request for differential weighting of the questions and, finally, the fickle professor expresses concerns about guessing and demands that the scoring incorporate various penalties for wrong answers.

Encompassing all these testing options in one SAS program greatly encumbers the complexity of the algorithm. To simplify program logic and reduce the program execution time, the Ryerson Test Response System requires the examiner to choose one of the scoring techniques with external data parameters. A front-end SAS macro uses INCLUDE statements to branch to the appropriate source code module located in a MACLIB library.

A weighting procedure gives the examiner the discretion to make some questions more important than others in their contribution to the final mark. In conventional scoring the weight operates at the level of the question. However, one of the advantages of a generalized approach is that differential weights can be given to the alternatives within a question (without precluding the possibility of also giving separate weights to the questions). This increased flexibility creates a problem in test administration because it is impractical to design an examination sheet that will accept a separate weight for each alternative. For the Ryerson Test Response System an input screen has been created with PROC FSEDIT. The examiner enters the desired weights which are matched with the examiner's data file containing the scanned sheets.

COMPARISON OF ALTERNATIVE TESTING AND SCORING METHODS

The Ryerson system provides various psychometric indices of the quality of measurement. Coefficient alpha (Cronbach, 1951), for example, is a measure of a test's reliability. It is a criterion for whether one testing or scoring method is an improvement over another. An equivalent statistic, the Kuder-Richardson Formula 20 (Kuder and Richardson, 1937), is available for dichotomously scored tests. Test scores obtained under different administration or scoring methods can also be compared using PROC CORR. If the results show high intercorrelations and have equal reliability, then one or more of the tests can be abandoned. If they differ significantly then doubts are raised that the approaches measure the same dimension of knowledge.

CONCLUSION

Objective achievement testing provides opportunities for collaboration between programmers and educators. At Ryerson, requests for psychometric analyses have led to SAS programming innovations which in turn have stimulated interest in alternative testing methods.

REFERENCES

Cronbach, L. J. (1961). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Klein, W. J. and Gordon Emslie (1991). Psychometric applications with optical scanning and analysis. SUGI Conference Proceedings, 16.

Kuder, G. F. and Richardson, M. W. (1937). The theory of estimation of test reliability. Psychometrika, 2, 151-160.

Magnusson, D. (1967). Test Theory. Don Mills, Ontario: Addison-Wesley.

Wood, R. (1977). Multiple choice: A state of the art report. Evaluation in Education: International Progress, 1, 191-280.

W. J. Klein, Ph.D.
Ryerson Polytechnical Institute
350 Victoria Street
Room S-348
Toronto, Ontario
M5B 2K3
Voice 416/979-5000 X 7082
FAX 416/979-5341

Gordon Emslie, Ph.D.
Ryerson Polytechnical Institute
350 Victoria Street
Room A-816
Toronto, Ontario
M5B 2K3
Voice 416/979-5000 X 6198
FAX 416/979-5273

C:\DATA\WP\SUGI\SUGI1801.DOC
April 15, 1993