REPORT RESUMES

ED 012 064 CG 000 201

EVALUATING THE USE OF ASSESSMENT PROCEDURES DEVELOPED IN ONE SCHOOL IN OTHER SCHOOLS.

BY- DARLINGTON, RICHARD B.
CORNELL UNIV., ITHACA, N.Y.

REPORT NUMBER BR-5-0607 PUB DATE JAN 67
CONTRACT OEC-6-10-041
EDRS PRICE MF-$0.09 HC-$1.28 32P.

DESCRIPTORS- *TEST CONSTRUCTION, TESTING, *TESTING PROBLEMS, *STATISTICAL DATA, PHI COEFFICIENT, CHILDREN'S PERSONALITY INVENTORY, CROSS VALIDATION, MINNESOTA MULTIPHASIC PERSONALITY INVENTORY, CORRELATIONS, ITHACA

THE PAPER REPORTS ON AN ATTEMPT TO DETERMINE EMPIRICALLY, FOR SEVERAL TEST-CONSTRUCTION PROBLEMS, THE AMOUNT OF IMPROVEMENT RESULTING WHEN TESTS ARE TAILORMADE TO FIT ONE PARTICULAR CHARACTERISTIC OF A LOCAL POPULATION--THE BASE RATES OF THE TWO CRITERION GROUPS WHICH THE TEST IS DESIGNED TO SEPARATE. THE BASIC PROCEDURE USED WAS TO CONSTRUCT A SERIES OF TESTS WHICH WERE ALIKE IN THE ITEM POOL AND ITEM-SELECTION TECHNIQUE USED, AND IN THE TWO CRITERION GROUPS WHICH THE TESTS WERE DESIGNED TO SEPARATE, BUT WHICH DIFFERED IN THE RELATIVE BASE RATES OF THE TWO CRITERION GROUPS ASSUMED IN THE CONSTRUCTION OF THE TESTS. CROSS-VALIDATION SAMPLE DATA WERE THEN USED TO ESTIMATE THE VALUE OF EACH OF THE TESTS IN POPULATIONS WITH EACH OF THE ASSUMED BASE RATES. THE PURPOSE WAS TO ESTIMATE, FOR EACH OF THESE POPULATIONS, THE EXTENT TO WHICH THE TEST TAILORMADE FOR THAT POPULATION EXCEEDED IN VALUE TESTS TAILORMADE FOR POPULATIONS WITH DIFFERENT BASE RATES. THE RESULTS SHOWED NO NOTICEABLE DIFFERENCE IN THE VALUES OF THE VARIOUS TESTS. THESE RESULTS WERE CONSISTENT ACROSS FOUR DIFFERENT TEST-CONSTRUCTION METHODS STUDIED, AND ACROSS THREE DIFFERENT SETS OF DATA WHICH DIFFERED IN THE ITEM POOL AND THE CRITERION GROUPS USED. (AUTHOR)

FINAL REPORT

Project No. 3054

Contract No. OE-6-10-041

EVALUATING THE USE OF ASSESSMENT PROCEDURES DEVELOPED IN ONE SCHOOL IN OTHER SCHOOLS

January, 1967

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Evaluating the use of assessment procedures developed in one school in other schools

Project No. 3054
Contract No. OE-6-10-041

Richard B. Darlington,
Principal Investigator

January, 1967

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

Cornell University

Ithaca, New York

Table of Contents

Acknowledgements

The author is indebted to Mrs. JoAnn T. Barsis, Mr. Winifred Buckwalter, Mr. Bram Goldwater, Mrs. Charlotte S. Otterbein, and Mr. Roy Williams, all of whom served ably as research assistants in the current project, and to Drs. Richard Holroyd, Albert Rosen, and Robert Wirt, who supplied the data to be analyzed.

Introduction

Users of standard psychological tests must regularly face the fact that the population of people for which a standard test was initially designed differs somewhat from the local population to which the user plans to apply the test. Furthermore, the use for which the test was initially designed often differs somewhat from the use which the local user has in mind. These users must regularly ask whether the time and expense involved in constructing a new test, tailor-made to the characteristics of the local population and the local planned use of the test, would be repaid by a noticeable improvement in predictive power. If fitting tests to the characteristics of a local situation regularly results in a large increase in predictive power, then constructors of standard tests also have to begin to consider the possibility of developing several tests to measure each trait in different local situations, where previously they have only constructed one test. If fitting the tests in this manner regularly results in only a small increase in predictive power, then this fact should be known so that test constructors can have more confidence in the present procedure of constructing only a single test for each trait to be measured.

The present paper reports on an attempt to determine empirically, for several test-construction problems, the amount of improvement resulting when tests are tailor-made to fit certain particular characteristics of a local population and a local use for the test. By a test-construction "problem," we mean a particular choice of the following factors: (a) the variable to be predicted by the test (the criterion variable), (b) the set of items from which the test items are to be chosen (the item pool), (c) the method by which items are to be selected from the test (the test-construction method), (d) the population of people from which a sample is drawn. The individual problems used in the study were chosen not from an interest in those specific problems. Rather, the hope was that these problems would be representative of a certain carefully-defined class of problems, and further that the results of the study would be consistent enough across the problems studied so that some generalization could be made with reasonable confidence to the entire class of problems. As we will see, this latter hope was fulfilled; the results were highly consistent across the different test-construction problems studied.

To help define the class of test-construction problems studied, we will give an example of a problem in the class. Consider a situation in which a test is to be used to discriminate between two groups of people. The groups might be, for example, future school dropouts and non-dropouts. Let these two groups be termed the "criterion groups." Suppose that each pupil who takes the test is given, on the basis of his test score, one of two treatments. One treatment, for example, might be placing him in a special class with a teacher trained to deal with potential dropouts, while the other "treatment" would consist of leaving him in his normal class. Suppose the test is constructed by selecting dichotomous (yes-no) items from a large pool of items, on the basis of the items' ability to discriminate between a particular sample of known dropouts and another sample of non-dropouts. The present project is confined to situations with all of the above characteristics: two criterion groups of people which are to be distinguished by a test, two treatments, of which one is to be administered to each subject on the basis of his test score, tests constructed empirically by selecting from a large pool of dichotomous items those items which discriminate well between samples of people from the two criterion groups. We assume also that there is a "flexible quota;" that is, there is no predetermined number of subjects to be assigned to each of the two treatments. Rather, each subject is assigned to the treatment deemed best for him. The opposite, "fixed quota," situation is often found in college admissions, say, where the two treatments are admission and non-admission, and the number of students to be admitted is fixed in advance.

We turn now to a description of the characteristics of local situations to which tests were tailor-made in the present study. There were two such characteristics. Again we will begin with an example.

Suppose a test constructor is constructing a test to identify future dropouts. He selects for his test those items which, on the basis of a previous sample of students, best discriminate between dropouts and non-dropouts. Suppose two Yes-No items are being compared for relative value in this situation. Item #1 is answered "Yes" by all dropouts and by half of all non-dropouts. Item #2 is answered "No" by all non-dropouts and by half of all dropouts. If students answering "Yes" are identified as dropouts and students answering "No" are identified as non-dropouts, then Item #1 misclassifies no dropouts but half of all non-dropouts, while Item #2 misclassifies no non-dropouts but half of all dropouts. Therefore, whether Item #1 or Item #2 misclassifies more people depends upon whether dropouts are more common than non-dropouts. We conclude that one characteristic of a local situation which should be considered in test construction is the relative sizes of the two criterion groups. We will call these relative sizes the base rates. Base rates were one of the two characteristics of a local situation to which tests were tailor-made.

To introduce the second characteristic, we will go on with the last example. Suppose that in school C students identified as dropouts are put in a special "dead end" class which would be quite injurious to the future of a non-dropout, while in school D students identified as dropouts are shown a special movie discouraging dropping out but are otherwise treated as other students. In school C incorrectly identifying a student as a dropout is a much more serious error than in school D. In school C, therefore, it would be much worse to replace item #2 by item #1, which misclassifies more non-dropouts as dropouts, than in school D. In other words, the relative value of the of the two items changes from school to school since the relative seriousness of the two types of treatment error (treating a dropout as a non-dropout vs. treating a non-dropout as a dropout) changes. We will call this characteristic of a local situation the relative seriousness of errors. It was the second of the two characteristics which tests in this project were tailor-made to fit.

Purpose

The purpose of the project was to estimate, for several test-construction problems, the increase in test value (in applying the test to a particular local situation) which results when the above-mentioned two aspects of the local situation (base rates, and relative seriousness of the two types of treatment error) are taken into consideration in the construction of the test. More specifically, the purpose of the project was to estimate, for several test-construction problems, how much more valuable is a test which is constructed using the values of the base rates and seriousness-of-errors factors which apply to the situation in which the test is to be used, than are tests which are used in that same situation but which were constructed assuming other values for these two factors.

Preliminary work

The above formulation of the purpose of this project immediately raises two questions. First, how do we take the base rates and seriousness-of-errors factors into consideration in constructing tests? Second, how do we measure the "value" of a test in a specific situation?

Neither of these two questions can be dealt with rationally unless we first assume that the seriousness of an error can be measured; or at least unless we assume that the relative seriousness of two errors can be measured. Although such measurements are extremely difficult, there is no doubt that in actual practice we make judgments every day which require us to estimate, at least subjectively, the relative seriousness of two different errors. A rational person postpones a drive to another city on a snowy day, not because he thinks he will probably have an accident, but because if he does have an accident the resulting loss is likely to be much greater than the loss of convenience resulting from postponing the trip. A counselor might spend $100 of his time talking to a student who has hinted he might become a dropout, not because the counselor thinks he will probably drop out, but because if he does drop out the loss will be much greater than $100. To ignore the problem of relative seriousness of errors generally leads one to act as if all errors were equally serious. This solution does not avoid the question at all; it simply substitutes an arbitrary and obviously incorrect assumption for whatever alternative assumption might be made by a careful study of the situation. The assumption that some numerical value, however arbitrary, can be assigned to the relative seriousness of two errors is the central assumption underlying decision theory. The general point is expanded by Cronbach and Gleser (1), which is the first work to apply decision theory in a major way to the area of psychological testing. (See especially Ch. 10 for a discussion of the rationality of attempting to measure the relative seriousness of two errors.)

Given that there is some way of measuring, or at least estimating, the relative seriousness of the two types of treatment error in a two-treatment situation, we turn now to the first of the two questions posed above--how do we take into consideration the base rates and seriousness-of-errors factors when we are constructing a test? We shall first consider base rates alone.

Consider a situation in which the two types of treatment error are equally serious (for example, misclassifying a future dropout as a non-dropout is exactly as serious as misclassifying a future non-dropout as a dropout). In this situation, the objective of a test is simply to minimize the total number of errors of classification. In this simplified situation, how should base rates be considered in test construction? The classical method of considering base rates is simply to let the relative sizes of the two samples of people used in the item-selection process be the same as the relative sizes of the groups they represent in the population. Thus, if non-dropouts are three times as common as dropouts, then the classical procedure dictates that these same base rates should be used in the samples of people whose data are used for test-construction.

But suppose the samples of people available to the test-constructor do not have these relative sizes, and there is no practical way to gather more data. Should he simply throw away data from the group which is too large, or is there a better way? We propose that a better way would be as follows.

Suppose a test-constructor is using the phi coefficient as the index which he uses to select items. That is, he computes the phi coefficient (ϕ) showing the ability of each item in an item pool to discriminate between the two samples of people he is using, and selects for his test those items for which phi is highest. As is well known,

$$(1)\ \phi = \frac{ad - bc}{[(a + b)(c + d)(a + c)(b + d)]^{\frac{1}{2}}},$$

where

a = proportion of the total population of people which is in criterion group A and which also answers "yes" to the item in question,

b = proportion of the total population which is in criterion group A and which answers "no" to the item in question, c = proportion of the total population which is in criterion group *B* and which answers "yes" to the item in question,

d = proportion of the total population which is in criterion group _ and which answers "no" to the item in question.

Thus

$$a + b + c + d = 1.$$

Let

p_1 = proportion of criterion group *A* answering "yes" to the item in question,

p_2 = proportion of criterion group *B* answering "yes" to the item,

P = proportion of total group which is in criterion group *A*.

Then we can express a, b, c, and d in terms of p_1, p_2, and P, by the formulas

$$a = p_1P$$

$$b = (1-p_1)P$$

$$c = p_2(1-P)$$

$$d = (1-p_2)(1-P).$$

Thus a test-constructor can use his actual sample data (without throwing any away) to estimate p_1 and p_2, the proportions of the two criterion groups answering "yes" to an item. If he is trying to construct a test for a local population (for example, a particular school) with a particular value of *P*, then he can enter that value of *P*, along with his empirical values of p_1 and p_2, into the last four equations, and thereby estimate the values which a, b, c, and d would assume in that local population. He can then insert those values of a, b, c, and d into formula (1) to find an estimate of the phi coefficient which that item would have in that local population. He can do this even if the relative sizes of the two criterion groups in his samples of people are grossly different from the relative sizes of the criterion groups (measured by *P*) which exist in the local population for which he is developing the test.

We have described in detail the procedure an investigator would use if he were using the phi coefficient as an index for item selection. Procedures analogous to those above can be (and were) developed when indices other than the phi coefficient are used for item selection. These will be described in more detail later.

We have seen above a method for considering criterion-group base rates in constructing a test. These base rates were one of two factors mentioned above which should be considered in constructing a test. The other factor was the relative seriousness of the two types of errors of misclassification--misclassifying a member of criterion group A as a member of group B, and misclassifying a member of group B as a member of group A. How should this second factor be taken into account in test construction?

Suppose each misclassification of a member of the first criterion group is judged to be three times as serious as a misclassification of a member of the second criterion group. Then the choice of treatment for each member of the first group is three times as important as the choice of treatment for each member of the second. Hence it seems intuitively clear that in the test-construction process, each member of the first group should be given three times the weight given to each member of the second group.

For example, suppose a test is being developed for a school in which dropouts and non-dropouts are equally common. Then using the above notation, $P = .5$. But suppose a highly effective counseling program is available, so it is decided that failure to identify a future dropout is three times as serious an error as incorrectly labelling a non-dropout as a dropout. Then in the test-construction process, the way to take into consideration the fact that the correct treatment choice for each potential dropout is three times as important as the treatment choice for each non-dropout, is to pretend, during the test-construction process, that there are actually three times as many dropouts as non-dropouts. In other words, even though the test-constructor knows the base rates are .50 and .50, he should enter into his item-selection formulas base rates of .75 and .25, since .75 is three times as large as .25. This should be done when he is calculating indices of item value. Thus, for example, if he is using the phi coefficient as the index of item value, then when he is calculating a, b, c, and d from p_1, p_2, and P in the manner described earlier, he should let P in the formulas be .75 instead of .5.

We will now state in more general algebraic terms the procedure we have just described in terms of a specific example. Let P' be the P-measure reflecting the actual base rates in a certain local population. (In the example just given, $P' = .5$.) Let P be the P-value which the test-constructor is going to "pretend" exists. (In the above example, $P = .75$.) Let U_A and U_B equal the seriousness of the two types of treatment error. Then the procedure illustrated in the example amounts to finding a value of P such that $P/(1-P)$ (the "pretended" relative base rates of the two groups) exceeds $P'/(1-P')$ (the actual relative base rates) by a factor of U_A/U_B. (In the above example, $U_A/U_B = 3$.) That is, the problem is to find P such that

$$(2) \quad \frac{P}{1-P} = \frac{P'}{1-P'} \cdot \frac{U_A}{U_B} \quad .$$

In the above example, where $P' = .5$, and $U_A/U_B = 3$, this equation becomes

$$\frac{P}{1-P} = \frac{.5}{.5} \cdot 3 \; ,$$

which is fitted when $P = .75$.

Formula (2) can be rearranged algebraically to give

$$(3) \quad P = \frac{\frac{P'}{1-P'} \cdot \frac{U_A}{U_B}}{1 + \frac{P'}{1-P'} \cdot \frac{U_A}{U_B}} \quad .$$

Thus a test-construction procedure which considers both criterion-group base rates and the relative seriousness of errors in a local situation is to first estimate P' and U_A/U_B for that situation, then compute P from (3), and then use this value of P along with p_1 p_2 in computing a phi coefficient or other index of item usefulness.

Thus, a test can be tailor-made to fit a situation with given values of P', U_A, and U_B by constructing a test to fit an imaginary new situation in which the two types of treatment error are equally serious, and in which the base rate of the first criterion group is P, where P is calculated from (3). The same test fits both the real and the imaginary situation, and in fact fits all situations in which that same value of P would be calculated from P', U_A, and U_B.

A preview of the design of the present study

The above considerations suggest the possibility of constructing a set of tests in which the first test in the set is constructed to fit a P of .05, the second test is constructed to fit a P of .10, and so on, with the last and 19th test being constructed to fit a P of .95. Then for any situation, a P-value could be calculated for that situation from the actual values of P', U_A, and U_B in that situation. Then a worker could select from the set of 19 tests the one test constructed to fit the P-value closest to the P-value of the new situation. That test would then be the one which should be applied to the new situation.

We saw above that a particular test-construction method will construct the same test for all test-use situations for which the same value of P is calculated from formula (3). Since constructing a test involves comparing the values of items, another way of stating our conclusion is that the relative values of items in a situation can be computed merely from knowing P for the situation; the values of P', U_A, and U_B need not be known separately. But what is true for items should also be true for tests; it should be possible to compute the relative values of several tests in a situation from knowing only the P-value of the situation. Thus if we have a set of tests, the same test should be the best test in the set for all situations with the same value of P.

We are thus led to the following conclusion. Suppose a table were available in which the first column shows the relative values of each of several tests in a situation in which the base rate of the first criterion group is .05 and in which the two types of treatment error are equally serious. Suppose the second column of the table shows the relative values of those same tests in a situation in which the base rate of the first criterion group is .10 and in which the two types of treatment error are equally serious. Suppose that there are 19 columns of the table altogether, each showing the relative values of the tests in situations with different base rates but keeping the assumption that the two types of treatment error are equally serious. The base rate for the third column would be .15, for the fourth would be .20, and so on, with the base rate for the 19th column being .95. Then if there were a real-life situation in which the two types of treatment error were not equally serious, it would be possible to calculate P for that situation from formula (3), and then go to the column of the table with the base rate closest to that P. The entries in that column would then show quite accurately the relative values of those several tests in that real-life situation.

Suppose further that the several tests whose values were listed in the table were the 19 tests referred to above, constructed to fit 19 different values of P. Then the first entry in the first column of the table would be the value, in a situation in which P = .05, of a test constructed to fit a situation for which P equals that same value of .05. Going down the column would give the values in that same situation of tests constructed to fit increasingly different P-values. Hence the first entry would be expected to be the highest entry in the first column. By the same reasoning, the highest entry in the second column would be expected to be the second entry, and in general the highest entry in each column would be expected to be the entry falling on a diagonal line running from the upper left corner of the table to the lower right corner. In any given column of the table, the farther away an entry is from this diagonal line, the smaller the entry would be expected to be.

But how much smaller would these entries be than the entries near the diagonal? If they are much smaller, it would mean that it is very important, when constructing a test for a specific situation, to use in the test-construction process the exact P-value which applies to that situation. If the entries far from the diagonal are only very slightly smaller than the entries near the diagonal, it would mean that it is not very important to consider a situation's P-value when constructing a test to be used in that situation. It would further imply that a single test could be used in situations with diverse P-values, with results almost as good as could be achieved by constructing many tests to fit the different situations.

The present project constructed 12 such 19 x 19 matrices. Each matrix was the result of applying a different one of four test-construction methods to a different one of three sets of data.

More preliminary work

Before we turn in more detail to the design of the project, we must first consider the second of the two questions posed above. The first question, which we have now answered, is how we take into consideration, in the process of test construction, the base rates and relative seriousness of the two types of treatment errors. The second question, to which we now turn, is how we estimate the "value" of a test, or at least the relative values of two tests, in a situation with specific values of P' and U_A and U_B.

This question was considered in two papers by Darlington and Stauffer (3, 4). The second of these two papers describes a method for finding the optimum cutting point on a test (that is, the point such that people with test scores above the point should receive treatment A while people with test scores below the point should receive treatment B) as a function of the mean test scores of the two groups of people, the standard deviations of the test scores of the two groups, and the relative seriousness of the two types of treatment error. The formula assumes that the test scores of each of the two groups of people are normally distributed.

Once this cutting point has been found, the only characteristics of the test which are relevant to its evaluation are the proportions of each of the two criterion groups with scores falling on each side of the cutting point. Since the test has thus been dichotomized by the cutting point, it can be evaluated by using the same formulas used to evaluate a dichotomous item. The procedures for evaluating such a dichotomous item were described in the first of the two papers by Darlington and Stauffer. That paper is basically an exposition of elementary decision theory as it is applied to the use and evaluation of discrete tests. Fortunately, in the present study we can avoid most of the complications in that paper, because we have seen above that we need to measure test values directly only in situations in which the two types of treatment error are equally serious. In such cases, the obvious measure of test value is simply the number of correct classifications made by the test, expressed as a proportion of the total number of classifications made. In the present project, we subtracted from this number the base rate of the larger criterion group (that is, the larger of the two numbers P and

1-P), since this is the proportion of the total population which could be classified correctly if we simply classified everyone, in the absence of any test information, in the larger of the two criterion groups. Thus the actual measure of test value, which we will call V, is the increase in the overall proportion of correct classifications resulting from use of the test. For example, if a given test classifies correctly .8 of the members of the first criterion group, and .7 of the members of the second, and if P is .6, then the overall proportion of correct classifications made by the test is

$.8 \cdot .6 + .7 \cdot .4$, or .76, so V, the value of the test, is

.76-.6, or .16. If c_A is the proportion of the first criterion group classified correctly by the test, and c_B is the proportion of the second criterion group correctly classified, and if M is defined as the larger of the two numbers P and (1-P), then our definition of V amounts to

$$V = c_A P + c_B(1-P)-M.$$

This formula provides the answer to the second of the two questions raised above--how do we measure the value of a test?

A proof

In the foregoing discussion we have relied on the reader's intuition to establish the point that the relative values of several tests will be the same in all situations for which the same value of P is calculated from formula (3). We will now establish this point more formally. The present subsection can be skipped by readers who are already convinced of the truth of the assertion.

The seriousness of decision errors can be measured in any convenient units. In mental hospital settings, it might be measured in the number of months by which a patient's stay is lengthened as a result of being assigned to an inappropriate treatment. In industrial and commercial settings the unit of measurement is usually dollars. Although a measure in dollar terms is often not appropriate in educational settings, we will nevertheless use this in an example, because of its ready understandability and simplicity.

Consider a situation in which $P' = .6$, $U_A = \$3$, and $U_B = \$5$. Assume that the first criterion group (whose base rate is P') is the group for which treatment 1 is appropriate, and that the second group (whose base rate is 1-P') is the group for which treatment 2 is appropriate. Then suppose treatment 1 were being given to everyone and it was being considered whether treatment 2 would be better. From the above numbers we see that for .6 of the total group, a switch to treatment 2 would produce a loss of $3 per person, while for .4 of the total group it would produce a gain of $5 per person. Thus the average gain resulting from the shift is

$$.4 \cdot \$5 - .6 \cdot \$3,$$

or \$2.00 - \$1.80, or \$.20. Thus the average gain is positive, and treatment 2 is better for the total group if no test is being used.

Suppose now a test is introduced which correctly classifies .8 of group 1 and .7 of group 2. Consider the strategy of giving treatment 1 to everybody above the cutting point (that is, on the side of the cutting point toward which most members of the first group fall), and treatment 2 to everybody below the cutting point. Switching to this strategy and away from the strategy of giving treatment 2 to everybody (which was the best strategy available without use of a test) produces the following results:

(a) the proportion of the first criterion group treated correctly rises from 0 to .8, at an average gain of \$3 per person correctly classified. Thus the mean gain for this group is .8 · \$3, or \$2.40.

(b) the proportion of the second criterion group treated correctly falls from 1.0 to .7, at an average loss of \$5 per person incorrectly classified. Thus the mean loss for this group is .3 · \$5, or \$1.50.

Recalling that the base rates for the two groups are .6 and .4, the mean gain in the two groups taken together resulting from using the test is

$$.6 \cdot \$2.40 - .4 \cdot \$1.50$$

or \$1.44 - \$.60, or \$.84. Since introducing the test has resulted in a mean gain of \$.84 per person in the two groups together, \$.84 is called the mean value per person of the test, which we denote by V. V is the measure of test value used in this report.

Consider now the problem of stating the procedure just described in terms of an algebraic formula. Let c_A and c_B be the proportions of the two criterion groups correctly classified by the test. (In the above example, c_A was .8 and c_B was .7.) Recalling that U_A was \$3 and U_B was \$5, we see that \$2.40 in the above example was $c_A \cdot U_A$, and \$1.50 was $(1 - c_B) \cdot U_B$. Recalling that .6 and .4 represented P' and (1-P') respectively, the above procedure amounted to using the formula

$$V = P' \cdot c_A U_A - (1-P')(1-c_B)U_B \; .$$

This formula simplifies algebraically to

$$(4)\quad V = P'U_Ac_A + (1-P')U_Bc_B - (1-P')U_B.$$

If we were to repeat the above line of reasoning in an example in which the treatment best for group A was the treatment which should be used for everyone in the absence of a test, we would arrive at the formula

$$(5)\quad V = P'U_Ac_A + (1-P')U_Bc_B - P'U_A$$

instead of (4).

Formulas (4) and (5) give the value of a test in situations of the sort considered in this paper. Consider now the ratio between the value of two tests j and k. Let c_{Aj} and c_{Bj} be the proportions of the two criterion groups classified correctly by test j, and let c_{Ak} and c_{Bk} be the proportions of the two criterion groups classified correctly by test k. Starting with (4), the ratio of the values of the two tests is

$$(6)\quad \frac{V_j}{V_k} = \frac{P'U_Ac_{Aj} + (1-P')U_Bc_{Bj} - (1-P')U_B}{P'U_Ac_{Ak} + (1-P')U_Bc_{Bk} - (1-P')U_B} .$$

Dividing both the numerator and denominator of the right side of (6) by $(1-P')U_B$ gives

$$(7)\quad \frac{V_j}{V_k} = \frac{\left(\frac{P'}{1-P'} \cdot \frac{U_A}{U_B}\right)c_{Aj} + c_{Bj} - 1}{\left(\frac{P'}{1-P'} \cdot \frac{U_A}{U_B}\right)c_{Ak} + c_{Bk} - 1} .$$

If we had started with (5) instead of with (4), we would have gotten

$$(8)\quad \frac{V_j}{V_k} = \frac{c_{Aj} + \left(\frac{1-P'}{P'} \cdot \frac{U_B}{U_A}\right) c_{Bj} - 1}{c_{Ak} + \left(\frac{1-P'}{P'} \cdot \frac{U_B}{U_A}\right) c_{Bk} - 1} .$$

Thus (7) gives the ratio of the value of two tests when the treatment appropriate to the second group is the best treatment to give everyone in the absence of a test, and (8) gives that ratio when the treatment appropriate to the first group is the best treatment to give everyone in the absence of a test.

No matter whether (7) or (8) is the appropriate formula for the relative value of two tests, inspection of the formula shows that the relative value of two tests in a specific situation is affected only by the proportions of the two groups correctly classified by each of the tests (these proportions are unaffected by P', U_A, and U_B), and by the value of

$$\frac{P'}{1-P'} \cdot \frac{U_A}{U_B}$$

for the situation. But formula (2) shows that any two situations which have the same P value will also have the same value of

$$\frac{P'}{1-P'} \cdot \frac{U_A}{U_B} .$$

We thus reach the conclusion we sought to prove in the present subsection: the relative values of several tests are the same across all situations which have the same value of P.

Method

In reading the following procedures, it should be remembered that the primary purpose of this project is to make a statement about the general importance, for any type of prediction problem and any test construction method, of tailor-making a test to fit a specific situation. The descriptions of the specific data sets used are thus briefer (in accordance with the request for brevity in the Instructions) than they would be if we were interested in those data for their own sake. To a much smaller extent, this is also true of the descriptions of the test-construction methods used.

Subjects and test-construction problems

Three different test-construction problems were used. The first problem involved using the MMPI to discriminate 96 hospitalized schizophrenics from 250 nonschizophrenic mental hospital inpatients. The second involved discriminating 112 high-IQ children from 115 retarded children, through the responses of each child's mother to the 600-item Children's Personality Inventory. The third involved using the MMPI to discriminate 136 paranoid schizophrenic mental hospital inpatients with low scores on the MMPI paranoid scale, from 134 nonschizophrenic inpatients with low scores on the paranoid scale. In the first set of data, 29 schizophrenics and 83 nonschizophrenics were set aside as a cross-validation sample. In the second set, 40 high-IQ children and 40 low-IQ children were set aside. In the third set, 48 patients from each of the two criterion groups were set aside. The remainder of each set of data was regarded as the test-construction sample in the analyses described below.

Test construction process

For each of the above-described three sets of data, four sets of 19 tests each were constructed. Thus the total number of tests constructed was 3 x 4 x 19, or 228. The four sets of tests differed from each other in that each set used a different test-construction method. The 19 tests within a set differed from each other only in the value of P assumed in the item-selection process. In the first test within each set, P was .05. In the second test within each set, P was .10. For the third, fourth, and other tests within each set, P was .15, .20, ..., .95.

As was just mentioned, four test-construction methods were used. All four of these methods were procedures for selecting, from a pool of several hundred items, a smaller number of items for inclusion in a test. In all cases, the selected items received unit weights; there was no attempt at differential weighting. Two of the four methods of evaluating items used just P and p_1 and p_2, the proportions of the two criterion groups answering "Yes" to the item. The other two methods used additional information which will be described later in this subsection. Within the bounds of the study, the four methods were chosen to represent the major types of test-construction methods in general use.

Method 1 was the method which was described in detail in the Introduction, using the phi coefficient (ϕ). The four numbers a, b, c, and d were computed for each item by the formulas

$$a = p_1 P$$

$$b = (1-p_1)P$$

$$c = p_2(1-P)$$

$$d = (1-p_2)(1-P).$$

For each item, ϕ was then computed by formula (1). The 36 items for which the absolute value of ϕ was highest were selected for the test. Of course, for the items in the 36 for which ϕ was negative, the scoring direction of the item was reversed before the item was included in the test. That is, if ϕ for an item was positive then a "Yes" answer increased the subject's test score. If ϕ for the item was negative, then a "No" answer increased his score.

Method 2 was more complicated. It used a technique described by Darlington and Bishop (2). The intent of the technique is to start with a "first-stage" test, then to add to that first-stage test those items in the item pool which are most able to improve that test. The ability of an item to improve the first-stage test increases with the item's validity, but decreases as the item's correlation with the first-stage test increases. The specific index used to measure an item's ability to improve the first-stage test was the partial correlation between the item and the criterion variable, partialling out the first-stage test. For each of the 19 tests constructed by Method 2 for a particular one of the three sets of data, the test used as a first-stage test was the test constructed by Method 1 for that same data set and that same value of P.

The partial correlation coefficient is computed from the three simple correlation coefficients r_{ci}, r_{ct}, and r_{it}, where c is the criterion variable, t is the first-stage test, and i is the item in question.

It will be recalled that the purpose of the test-construction phase of the project was to construct, from a single set of data, a series of tests in which each test was designed to fit a situation with a different value of P, where P is thought of as the base rate of the first criterion group in an imaginary situation in which the two types of treatment error are equally serious. Thus in Method 2 we face the problem of estimating, from the single set of data, the values which r_{ci}, r_{ct}, and r_{it} would have in situations with different values of P. The item-criterion correlation r_{ci} is the only one of these three correlations which we have essentially already discussed. Since c and i are both dichotomous, the correlation between them is the phi coefficient, and we described in the Introduction the means for estimating the value which a phi coefficient would assume in a population with a given value of P. We must now consider r_{ct} and r_{it}. Since t is a continuous variable (the first-stage test), both r_{ct} and r_{it} are point-biserial correlations. We will consider first the problem of estimating, from a single set of data, the values which r_{ct} would assume in populations with different values of P.

The basic problem is to express the point-biserial correlation coefficient as a function only of P and of several quantities which would not be expected to differ across populations with different values of P. We faced a similar question earlier in connection with estimating the values which phi coefficients would assume in situations with different values of P, and we solved it by expressing the entries in the phi coefficient (that is, a, b, c, and d) as functions of P and p_1 and p_2, the proportions of the two criterion groups answering "Yes" to the item in question. These numbers p_1 and p_2 fit the above-mentioned specification--they would not be expected to differ in populations with different values of P. The problem is to find a similar set of numbers, and a set of formulas for computing the point-biserial correlation coefficient from them.

Expressing r_{ct} in terms of the familiar formula for a point-biserial correlation gives the formula

$$r_{ct} = \frac{\overline{T}_A - \overline{T}_B}{s_t} \sqrt{P(1-P)} \quad ,$$

where $\overline{T}_A$ and $\overline{T}_B$ are the mean test scores of the two criterion groups respectively, and where s_t is the overall standard deviation of the test scores. P is, as throughout this paper, the assumed base rate of the first criterion group in the situation for which the test is to be constructed, and in which the two types of treatment error are imagined to be equally serious. The quantities $\overline{T}_A$ and $\overline{T}_B$ are already expressed in acceptable form, since they would not be expected to vary across several populations which differed only in P. However, s_t is not in acceptable form; in general, the standard deviation of the test would be affected by P. This problem is handed as follows.

First express s_t in terms of the familiar standard deviation formula

$$s_t = \sqrt{\frac{\Sigma T^2}{N} - \overline{T}^2} ,$$

where T's are individual test scores and $\overline{T}$ is the overall mean of the test scores. But $\overline{T}$ can be thought of as a weighted average of $\overline{T}_A$ and $\overline{T}_B$, where the weights are P and (1-P). That is,

$$\overline{T} = P\,\overline{T}_A + (1-P)\,\overline{T}_B.$$

$\overline{T}_A$ and $\overline{T}_B$ fit our stipulation for entries into r_{ci}; they should not vary across populations with different values of P.

The term $\Sigma T^2/N$ which appears in the formula for s_t can be handled in much the same way we just handled $\overline{T}$. It is actually the mean value of T^2 in the population, so it can be expressed by

$$\frac{\Sigma T^2}{N} = P\,\overline{T^2_A} + (1-P)\,\overline{T^2_B} ,$$

where $\overline{T^2_A}$ and $\overline{T^2_B}$ are the mean values of T^2 in the first and second criterion groups respectively.

Thus we proceed as follows in estimating the value which r_{ct} would have in a situation with a given value of P. We enter that value of P into the last two formulas, along with the values of $\overline{T}_A$, $\overline{T}_B$, $\overline{T^2_A}$, and $\overline{T^2_B}$ observed in our sample data. The values of $\overline{T}_A$ and

$\Sigma T^2/N$ thus computed are entered into the above formula for s_t. That value of s_t is then entered into the above formula for r_{ct}, along with the same values of $\overline{T}_A$, $\overline{T}_B$, and P used before.

A similar procedure was used to find r_{it}, which like r_{ci} is a point-biserial correlation coefficient.

Once r_{ci}, r_{ct}, and r_{it} were found by the procedures described above, they were entered into the partial correlation formula

$$r_{ci \cdot t} = \frac{r_{ci} - r_{ct}\, r_{it}}{\sqrt{1 - r_{ct}^2}\,\sqrt{1 - r_{it}^2}} \, .$$

This quantity was computed for each of the several hundred items in the item pool, for each of the 19 values of P. For each value of P, the 9 items for which the absolute value of $r_{ci \cdot t}$ was highest were added to the 36-item first-stage test to form a 45-item second-stage test. For reasons analogous to those described in connection with Method 1, the items for which $r_{ci \cdot t}$ was negative were scored negatively.

Method 3 was quite different from Methods 1 and 2. We saw in the Introduction the formulas which would be used if one were to evaluate a test or a dichotomous item in terms of the increase in the proportion of correct classifications resulting from use of the test or item. These same formulas provided the basis for item selection in Method 3. That is, in each of the three sets of data and for each of the 19 values of P, we computed the estimated increase in the proportion of correct classifications resulting from using each item (alone, not in a test). This is a measure of item usefulness which is similar to, but definitely different from, the phi coefficient relating the item to the criterion variable. Since the formulas used were described in detail in the Introduction, we will not repeat them here, even though there the discussion was in terms of tests and here the discussion is in terms of individual items. The formulas are the same in the two cases. Once these numbers were computed for a given set of data and a given value of P, the 36 items for which the numbers were highest were selected for a test. As in the other methods, the entire procedure was repeated 19 times, for different values of P, within each of the three sets of data.

Method 4 was the only one of the four test-construction methods which was developed for this project. It was by far the most complicated of the four test-construction methods. We found that it was also by far the most consuming of computer time; problems for which Methods 1, 2, and 3 took about 5 minutes each, consumed about 2 hours of the computer's time for Method 4. Later, on cross-validation, the tests constructed by Method 4 were not noticeably better than those constructed by other methods. Method 4 was thus basically a "flop." For completeness, we will nevertheless include here a description of the method.

Method 4 had much the same relation to Method 3 that Method 2 had to Method 1. That is, Methods 1 and 2 were both within a correlational framework, and Method 2 used as first-stage tests the tests constructed by Method 1. Methods 3 and 4 were both within the framework of measuring item value by the increased proportion of correct classifications resulting from use of the item, and Method 4 used as first-stage tests the tests constructed by Method 3. The essence of Method 4 was a method for estimating which items, when added to a test constructed by Method 3, would raise by the largest amount the proportion of correct classifications resulting from use of the test, assuming that the test scores are normally distributed within each of the two criterion groups. This was done by adding each item in the item pool separately to the 36-item first-stage test, then computing the mean and standard deviation of test scores for each of the two criterion groups. These means and standard deviations were then entered into the above-mentioned formula developed by Darlington and Stauffer (4) for finding the optimum cutting point on the test. Once the cutting point was found, normal curve tables were used to compute, from the mean and standard deviation of test scores for each criterion group, the proportion of each criterion group falling on the correct side of the cutting point. These two proportions were then weighted by P and (1-P) to give an estimate of the overall proportion of correct classifications resulting from use of the original 36-item test with the one additional item added to it. This procedure was repeated hundreds of times, each time using as the one additional item a different one of the several hundred items in the item pool. The 9 items for which this statistic (the overall proportion of correct classifications) was highest, were then added all at once, as 9 items had been added all at once in Method 2, to the original 36-item test to form a 45-item test. As in the previous three methods, this procedure was repeated using 19 different values of P in each of three sets of data.

Cutting points

We have mentioned earlier that the correct cutting point for a test depends upon the P-value of the situation in which the test is to be used. Construction of the 19 x 19 value matrices mentioned above involves measuring the value of each of the above tests in situations with different values of P. Hence for each test we calculated the optimum cutting point for each of the 19 values of P. Besides P, the entries in this formula are the means and standard deviations of the test scores for each of the two criterion groups. These means and standard deviations were computed in the test-construction samples, and the 19 cutting points for each test were then calculated.

Cross-validation and the value matrices

We have described above a process by which 4 sets of 19 tests each were constructed for each of three sets of test-construction sample data. As mentioned earlier, corresponding to each of these sets of data was a cross-validation sample of data, which had not been used in the test-construction process. Using exclusively these cross-validation sample data, twelve 19 x 19 matrices of test values were constructed by the method outlined in the introduction. Constructing each of these 12 matrices involved the following steps:

(a) Let c_{Ajk} and c_{Bjk} be the proportions of the two criterion groups classified correctly, in the cross-validation sample, by the j^{th} test with its k^{th} cutting score, where subjects with test scores above the cutting point are classified as being in the first criterion group, and subjects with scores below the cutting point are classified as being in the second criterion group. Then c_{Ajk} and c_{Bjk} were computed for each value of j and each value of k from 1 to 19, forming a 19 x 19 matrix of values for each of the two statistics. c_{Ajk} and c_{Bjk} are the estimated proportions of correct classifications in the two criterion groups achieved by applying the j^{th} of the tests to a population with the k^{th} of the 19 values of P, with the cutting score on the test chosen to maximize the number of correct classifications for that value of P.

(b) If the k^{th} value of P is denoted by P_k, then

$$(9) \quad c_{Ajk}P_k + c_{Bjk}(1-P_k)$$

is the estimated overall proportion of correct classifications by the j^{th} test when used in a population with base rate P_k. The larger of the two values P_k and $(1-P_k)$ is the largest overall proportion of correct classifications possible without the use of a test; it is achieved by classifying all persons as members of the larger criterion group. If we define M_k as the larger of the two values P_k and $(1-P_k)$, then subtracting M_k from (9) gives the estimated increase in the number of correct classifications (expressed as a proportion of the total population) achieved by the use of test j. If we denote this quantity by V_{jk}, and term it the value of the j^{th} test in a population with base rate P_k, then we have

$$(10) \quad V_{jk} = c_{Ajk}\, P_k + c_{Bjk}\, (1-P_k) - M_k.$$

V_{jk} was computed for each of the 19 values of j and 19 values of k, forming a 19 x 19 matrix of test values. Each column of this matrix showed the values of 19 tests in a situation with a specific value of P. The rationale for this procedure was described in detail on pp. 7-10; those pages should be reviewed by readers for whom the present procedure or its rationale seems unclear.

Each of the twelve 19 x 19 matrices constructed by this procedure related to a particular one of the four test-construction methods and a particular one of the three sets of data. Each of the 12 matrices showed the relative values, as estimated from cross-validation sample data, of 19 different tests in each of 19 different situations with different values of P. The 19 different tests in a matrix had all been constructed using the same one of the four test-construction methods, but the 19 tests had been constructed independently so that they were tailor-made to fit situations with different values of P.

Results

As described in the Introduction, the major question of interest in the present project was whether tests constructed for a situation with a given value of P were more valuable in that situation than were tests which had been constructed to fit situations with different values of P. Phrased in different terms, are the entries falling on the upper-left-to-lower-right diagonal of a given one of the 19 x 19 matrices noticeably larger than entries in the same columns but far from the diagonal? The answer to this question is very simple: no. Although Table 1 shows only a small fraction of the total mass of data produced by the analyses described above, it is fully adequate to show the trend of the data. Table 1 contains twelve 2 x 2 matrices. Each of these twelve 2 x 2 matrices contains four elements from a different one of the twelve 19 x 19 matrices; the four elements show the value of each of two tests in situations with two different values of *P*. In the first two of the three sets of data the *P* values used were .3 and .7. In the third set, test values were all very low for those *P* values, so *P* values of .4 and .6 were used. In each of the twelve 2 x 2 matrices, the upper left and lower right elements show the estimated values of two tests in populations with the *P* values for which the tests were designed. The remaining two entries (lower left and upper right) show the estimated value of each test in a population with the *P* value for which the other test was designed. Comparing the two numbers within a column of any of the twelve matrices compares the estimated value, in a population with a given *P*, of a test constructed for that *P*, and a test constructed for a very different *P*. The former of the two numbers would be predicted to be higher, the question being how much higher. Of the 24 such comparisons that can be made in Table 1, only 8 (less than half) even show the difference to be in the predicted direction. Nor are the differences in the predicted direction larger than the others; if the 8 differences in the predicted directions are considered positive and the remaining 16 differences considered negative, the mean of the 24 differences is negative. Further, inspection of Table 1 shows that the 16 negative differences are not concentrated in the results of any one of the four test-construction methods or in the results of any one of the three sets of data; they are distributed across the four test-construction methods with the

Table 1

Values of different tests in each of several situations--major project results summarized in twelve 2 x 2 matrices*

(see text for explanation)

Test-construction methods (in order described in text)	Data set (in order described in text)					
	1		2		3	
1	027	073	232	255	067	067
	012	052	250	237	096	054
2	0	026	240	260	025	004
	027	040	232	242	063	037
3	028	042	223	247	050	033
	035	036	245	227	054	075
4	038	032	275	267	071	033
	014	0	215	245	067	025

*Figures in the table have been multiplied by 1000 to eliminate decimal points.

frequencies 5, 3, 5, 3, and across the three sets of data with the frequencies 5, 6, 5.

Discussion and Conclusions

We conclude that for the types of test-construction problem studied, tests tailor-made to fit the base rates and seriousness of errors of a particular local situation are little (if any) better in that situation than tests which were constructed for other situations. The results consistently supported this conclusion despite the diversity in criterion variables, test-construction methods, item pools, and samples of people studied.

The study, however, was confined to large item pools. Wilks (5) has shown that the correlations among tests constructed by different weighting methods can be expected to increase as the size of the item pool from which the test items are drawn increases. It is thus not certain whether the above conclusions apply when the item pool is substantially smaller than the pools of 550 and 600 items used in the present study.

A Minor Parallel Study

A minor parallel study, using the data and tests already described, was carried out on a question related to the major topic of this project. This second question was whether the choice of cutting point on a test greatly affects the value of the test for a particular population. This part of the project is of limited interest for two reasons: (a) it is quite simple to adjust the cutting point on a test to fit any particular population, using the Darlington-Stauffer technique mentioned above, so that the question of how much is lost by failing to do so is unimportant; (b) it seems obvious that the proper choice of cutting point would have a great effect on test value. This expectation was fully borne out by the data. The project consisted simply of repeating the previous project, with the single change that in this second phase the cutting point for a test was always left at the value originally calculated for it using the P value for which the test was constructed, rather than changing the cutting point to fit new P values. The data in Table 2 are fully adequate to show the trend of the data produced by this part of the project. As in Table 1, the data are arranged in twelve 2 x 2 submatrices, each submatrix corresponding to a different test-construction method and data set. The upper right and lower left entries in each submatrix are the same as in Table 1; they are thus values of tests constructed for one value of P and evaluated in a population with a very different P value. (The same P values were used as were used in Table 1.) Each of these entries should again be compared to the other element in the same column; this other entry is the value

Table 2

Values of different tests in each of several situations, as a function of placement of each test's cutting point*

(see text for explanation)

Test-construction methods (in order described in text)	Data set (in order described in text)					
	1		2		3	
1	0	073	188	255	058	067
	012	0	250	242	096	0
2	0	026	232	260	054	004
	027	0	232	260	063	0
3	0	042	198	247	017	033
	035	0	245	252	054	0
4	0	032	205	267	025	033
	014	0	215	275	067	0

*Figures in the table have been multiplied by 1000 to eliminate decimal points.

of the same test in the same situation as the other entry in that column, but with the cutting point set for the value of P for which the test was originally constructed, rather than for the value of P in the population in which test value is being measured.

Inspection of Table 2 shows that the choice of cutting point makes a large difference in test value. Of the 24 comparisons possible between the two elements within a column of a submatrix in Table 2, all but two comparisons are in the predicted direction, usually by a substantial margin. The two exceptions (on the right side of the last two submatrices in the center column) are both in the data set for which test cross-validities were much higher than are usually found in psychology, due to the nature of the criterion variable (IQ, divided into superior and retarded groups). In such data, most individuals are so far from the optimum cutting point that misplacing the cutting point should not be very serious, thus giving rise to the two observed differences in the non-predicted direction. In this respect, it is reasonable to believe that these data are atypical for psychology.

Summary

Users of standard psychological tests must regularly face the fact that the population of people for which a test was initially designed differs somewhat from the local population to which the test is to be applied. These users must regularly ask whether the time and expense involved in constructing a new test, tailor-made to the characteristics of the local population, would be repaid by a noticeable improvement in predictive power. The present paper reports on an attempt to determine empirically, for several test-construction problems, the amount of improvement resulting when tests are tailor-made to fit one particular characteristic of a local population--the base rates of the two criterion groups which the test is designed to separate. The basic procedure used was to construct a series of tests which were alike in the item pool and item-selection technique used, and in the two criterion groups which the tests were designed to separate, but which differed in the relative base rates of the two criterion groups assumed in the construction of the tests. Cross-validation sample data were then used to estimate the value of each of the tests in populations with each of the assumed base rates. The purpose was to estimate, for each of these populations, the extent to which the test tailor-made for that population exceeded in value tests tailor-made for populations with different base rates. The results showed no noticeable difference in the values of the various tests. These results were consistent across four different test-construction methods studied, and across three different sets of data which differed in the item pool and the criterion groups used. It is shown mathematically that the results imply that when decision-theory methods of test

construction and evaluation are used, no noticeable gain in test value results from explicitly considering the proper base rates and the relative seriousness of the two types of misclassification of subjects.

A minor parallel study showed that the choice of cutting point on a test has a major effect on the test's value.

References

1. Cronbach, L. J., and Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1957.

2. Darlington, Richard B., and Bishop, Carol H. "Increasing test validity by considering interitem correlations," Journal of Applied Psychology. 50, 1966, pp. 322-330.

3. Darlington, Richard B., and Stauffer, Glenn F. "Use and evaluation of discrete test information in decision making," Journal of Applied Psychology. 50, 1966, pp. 125-129.

4. Darlington, Richard B., and Stauffer, Glenn F. "A method for choosing a cutting point on a test," Journal of Applied Psychology. 50, 1966, pp. 229-231.

5. Wilks, S. S. "Weighting systems for linear functions of correlated variables when there is no dependent variable," Psychometrika. 3, 1938, pp. 23-40.