SOME STATISTICAL ASPECTS OF EDUCATIONAL EVALUATION

Charles Schotta, Virginia Polytechnic Institute
Richard B. Hoffman, State University of New York at Buffalo

Introduction

This paper discusses some statistical considerations underlying educational evaluation. We first point out the objectives of educational evaluation. Then we evaluate the existing set of procedures for producing such estimates from the standpoint of classical probability theory. Next we present some empirical evidence in support of our criticism of existing methods of securing educational evaluation. Some major alternatives to existing evaluation practices are then discussed. We conclude by exploring the essentially Bayesian nature of educational evaluation practices and by delineating some important directions for future research efforts.

I. The Objectives of Educational Evaluation Systems

Typically an educational evaluation, the "grade", is an attempt to measure or to identify the result of a course of instruction. Additionally this measure is used to discriminate between students for a number of purposes.[1] It appears that the manner of securing this educational evaluation can be described as follows: during a course of instruction of some time-interval, say a semester of 18 weeks, the process of instruction (which we, as economists, interpret as the production of human capital) is interrupted several times to administer a test instrument.[2] We will discuss the interpretation of the test instrument in our view of educational evaluation in a later section of this paper. The numerical (or alphabetic code) result of each instrument is weighted by some explicit or implicit function to produce a point estimate. This estimate is taken to represent the measure of the quality of human capital created in the course of instruction, an identification of the product itself, a specification of the intensity of the student input, or various combinations of these. The exact form of the relationship between these and the "grade" does not appear to be a simple one.

This description appears to be an accurate one regardless of the type of test instruments administered. To sum up, the essential features are (1) a rather large population of "learning" or "teaching" or "production" periods, (2) a fixed, usually small (relative to the above) number of "test points", (3) some weighted mean of the value of the "test points" is represented as a meaningful measure of the value of the units in (1), (4) the "grade" produced as (3) is used to rank whatever it is the course of instruction produced, that is, we differentiate the human capital that the student ends up with by using the grade set produced for a class of students.

The striking feature of this procedure is that, on the basis of a sampling technique of this kind, inferences about differences in human capital, with respect to quality and quantity, are made and rather precise decisions are based on these inferences. The usual consideration of variance of the estimates is totally lacking. Here an example might be useful. Suppose we have an introductory statistics course in which we hold 45 lectures during the semester. We sample the work with some instrument three times during the semester. We secure a numerical average for each student. Using some conversion scheme we translate this set of numerical averages into relative alphabetic-coded rankings. Invariably one must make a decision whether to assign a "B" code to the 79 and a "C" code to the 78 or to define the code class demarcation point one, two, or some number of units lower or higher. Once, however, the "grade" has been assigned, it is clearly inferred that you have "different" commodities.

We believe that one can properly describe the objective of educational evaluation as the securing of a ranking of the quality and quantity of the creation of human capital during a course of instruction which will serve as a proxy for the measurement of the efficiency of output of this capital "production" process. Regardless of the initial intent of evaluators, the result of their evaluation is typically used in resource allocation decisions.

II. A Sampling View of Educational Evaluation

Suppose we have the "grade" determined as we suggest it presently is determined. This is a point estimate with no variance estimate. Let us represent a course where test instruments x_1, x_2, x_3 are given and the explicit weighting function is .25, .25, .5. These test instrument measures might be thought of as identifying the state of each student at, say, points y_{12}, y_{36}, and y_{45} for a course of instruction, Y with 45 time points.

$$(1) \quad E = .25\, x_1 + .25\, x_2 + .5\, x_3$$

and is the estimate of the mean, y, of the actual value of points y_1 through y_{45}. E then has, although almost no educational evaluation procedures consider it, a variance estimate.

If one calculates this variance and expresses it as the standard deviation of E, then, based on the sample size, one can, at various probability levels, calculate confidence intervals around E for each student.

Suppose for the i^{th} student we have

$$(2) \quad E_i \text{ with } \sigma_E^i$$

and we wish to form the confidence interval at the .95 probability level. In the above example, as is usual in educational evaluation, the sample (the number of test instruments) is less than 10 per cent of the population (the number of pro-

duction points) hence no population correction factor needs be applied?[3]

Let us represent the true measure of the human capital creation which occurs in the 45 units as G_i.

Then, (3) $P[E_i - t\sigma^i_x/\ n^{1/2} < G_i < t\sigma^i_x/\ n^{1/2} + E_i\ ] = 1- \alpha$

For the case in question for the .95 probability level this would be

(4) $P[E_i - 1.96\sigma^i_x/\ n^{1/2} < G_i < 1.96\sigma^i_x/\ n^{1/2}] = .95.$

Although in a later section of this paper we present some summary results of viewing an educational evaluation in this way, we present below a hypothetical numerical example to illustrate our point. Assume: A = 90 - 100, B = 80-90, C = 70-80, n = 3, at the 1 - α = .95, the grades would be, for the given means,

	Mean Score	SD	Grade Rank
Student 1	75	4	C
Student 2	78	4	C or B
Student 3	85	7	A, B, or C

These examples illustrate that when one views the testing process as a sampling process, when inferences are required at a specified level of probability, it is often quite unclear that class rankings are unequivocal. Throughout our analysis we have assumed that the numerical measure attained on a given test instrument represents a zero variance point estimate of the state of the mind being tested?[4] This represents the most unfavorable assumption which can be made for the consequences of non-zero variance between test instrument scores upon which our criticism of current evaluation methods rests. If one relaxes this assumption to more closely reflect reality, then our criticism is broadened to include the necessity for probabilistic evaluation of the numerical measures for each test instrument.

From the above examples it can be seen that there exists a situation where rankings have questionable meanings and alphabetic codings appear to be representations of rankings which are themselves of doubtful value in further decisions.

III. Empirical Evidence on the Validity of Some Grade Rankings

The following section describes the results of applying the foregoing statistical analysis to several large section social science courses where three instruments were applied during a semester. Two different weighting functions were applied, one consisting of a pattern of .25, .25, and .5; the other an equal weight system. Both sets of instruments consisted of two 50 (MC) multiple choice item instruments and one 120 (MC) item instrument. The instruments were not applied at a random time during the course of instruction. It is assumed that the numerical score for each student for each instrument is an estimate of the state of the human capital created at the time of the application of the instrument. It is further assumed that the numerical score on each instrument has an associated variance of zero.

When evaluated in the fashion suggested in Secion III, out of a total number of respondents of 560, in four classes of size 80, 180, 145, 155, only 20 per cent of the students had confidence intervals at the .95 probability level which were contained entirely within the pre-determined numerical intervals for conversion to alphabetic grades.

When the students were grouped into sets according to whether their numerical averages and the associated confidence band (1) overlapped a grade interval higher than that in which their mean lay, (2) overlapped a grade interval lower than that in which their mean lay, and (3) overlapped both as in both (1) and (2) above, the distribution between the three classes was

	(1)	(2)	(3)	no overlap
number	203	156	89	112
Percent	36	28	16	20

In the test set of students, it was far more likely that either a grade too high or a grade too low was assigned than was the case that there was a probability that any one of three grades was probable.

Indeed, the cases (1) and (2) are the interesting problems since in case (3) it was, on the average, only necessary to lower the probability level to approximately .7 to shrink the confidence interval to lie either entirely within the grade interval or to join cases (1) or (2).

On the other hand, it was necessary to lower the probability level to .4 to eliminate overlapping in case (1), while it was necessary to lower the probability level to .5 to eliminate overlapping in case (2).

It should be stressed that even then, the problem of assigning different grades to individuals who, using some test for difference between two estimates of means of different populations, each with a variance estimate, do not appear to have different means, remains and is a major obstacle to a clear-cut interpretation of the rankings which result from alphabetic grade-code assignment.[5]

It was on the basis of the empirical evidence cited above that we came to our considerable skepticism concerning the quality of educational evaluations produced by the system of educational evaluation outlined in Section I which we take to be wide-spread,

IV. Some Alternatives to Present Educational Evaluation Systems

In this section we wish to consider some major alternatives to the present system of educational evaluation and to evaluate these proposed alternatives against the simple framework of sampling analysis in which the present system was presented. While some of the alternatives are in reality modifications to the present system to circumvent difficulties we have pointed out above, others represent radical proposals for reform of educational evaluation.

A simple modification of present practice which would greatly reduce the variance of the numerical grade estimate is an increase in the number of instruments applied during a specific course of instruction. If this is also coupled with random selection of the specific points of sampling one can more easily reconcile the interpretation of the grade with practice in statistical quality control.

It must be clearly recognized that the assumption that the items on an instrument produce an estimate of the state of the population being sampled with a zero variance is unlikely to be true in practice. Hence the overall quality of the measure of what occurs in a course of instruction is affected not only by the number of instruments and the manner of their application and weighting, but also by the number of items per instrument and the variance associated with the "score" on each instrument. Unfortunately, although we can estimate the sample variance for the results of several instruments,we cannot do so for the items on an individual instrument. Many educational psychologists choose to regard the numerical measure of the items on an instrument as data without observation errors. If this is the approach selected then this is tantamount to accepting the zero variance nature of an instrument "score". The quality of the educational evaluation for a course of instruction is then, from the standpoint of the approach we are taking toward present procedure, independent of the number of items on a given instrument and is determined only by the number of instruments and the variance particular to each respondent being evaluated.

The suggestion that the number of instruments applied be increased leads to an interesting conclusion. Suppose that the application of an instrument is analogous to "destructive testing" in quality control in the sense that while testing the production of human capital does not proceed. There is then a trade-off between additional accuracy and capital building. If we reduce each instrument to one item and and structure the course of instruction in such a way that after each information bit is presented an item is presented, we have programmed instruction. One completes such a course of instruction by a time pattern of binary conditions which eventually leads to the last information bit and item in the sequence. Clearly the results of such a procedure must be evaluated in some manner additional to the items which follow each information bit. Usually the procedure is similar to the standard system we describe with formal instruments consisting of a large number of items being applied several times during the course. In principle, then, there is no difference between the result of such a procedure and that from a conventional course. Occasionally a "grade" for PI course will be derived from a measure of the percentage of successes on individual items. Although this procedure has some interesting aspects, we do not examine it.

Another major alternative is the system where only one instrument is applied for a course of instruction or for a specified sequence of courses. Such measures cannot be used for inferences about the capital creation process but only infer something about a state of the human capital created and in existence at the end point. If we take the view of the human capital which has been created and the physical equipment as being analogous to software, data, and hardware, then we are sampling the contents of bits in core, the ability of a program to call out the correct subroutines, and the logical structure of the circuitry. We think this is a reasonable view to take of such an examination procedure since the most usual objective given for the "comprehensive exam" is that "we will find out what he knows and doesn't know."

V. Educational Evaluation as a Bayesian Decision

There is another interpretation of the educational evaluation process as it is practiced today which might be related to Bayesian decision theory. Suppose we regard the "course average" as the most probable number describing whatever we wish to measure rank for a student in a course of instruction. Given the numerical interval for alphabetic coding, if a student score lies in the middle of the grade band, we are likely to answer the implicit question "what is the probability that his 'true' measure (if one exists) is high enough (or low enough) to give him the higher (or lower) grade" by saying "very low" and to regard this student as a "solid C" or whatever grade is in question. When, however, we have a student whose "course average" is on or near one or another end of the grade band, the a priori probability that his grade could be the higher or lower grade is considerably increased. For the student on the endpoint of the grade band, we might even feel that either grade could be correct. We seek additional information to make the decision. Almost invariably we do not replicate the experiment which produced the score under consideration. Rather we look at the pattern of scores on the instruments to see if it

is "rising", "falling" or some such thing. We try to think about the personality of the student. We consider "special" factors[6] and make a decision. The empirical evidence concerning the degree of overlap of the grade interval and the confidence interval presented in Section III suggests that this type of decision process is the most prevelant manner of resolving much difficulties discussed above lends support to the interpretation of the grading process as a Bayesian decision.

V. Research Areas

We would like to mention what we see as some areas for further research into the question of educational evaluation. We see this as important simply because of the resourse allocation decisions which are made on the basis of educational evaluations. Resource misallocation will result to the extent these evaluations are erroneous.

First, underlying the realiability of evaluations is the reliability of instruments. We are forced to regard the "score" for an instrument as zero-variance estimate or as data. The problem of the reliability of items and of instrument construction are not within our province but must be recognized as fundamental to successful evaluation.

It must, however, be recognized that the whole area of the design of the sampling plan, the selection of the weighting function, and the selection of the method of securing aggregation of course grade measures into larger measures is crucial to the production of quality measures for use in resource allocation[7]

---

1. Based on the relative ranking in a class, a conversion to some conventional grading system such as letter grades will be made. These letter grades will be subjected to additional transformations and operations and the results will be used for such purposes as deciding whether fellowships and scholarships will be given, whether or not a person will be continued in the educational process, what position to assign to a person in the job structure, whether a given male will be drafted or not, etc.

2. For a discussion of education as the process of "creating" human capital see T. W. Schultz, "Investment in Human Capital," Am. Econ. Rev. (March, 1961).

3. One must recognize here that number of items per test instrument may be substituted to an extremely limited degree for additional test instruments. However, since the number of items on an instrument must necessarily be extremely small relative to the population of information to be sampled and the items are not likely to be independent, the effective sample size is not likely to be markedly affected by an increase in the number of items per test instrument.

4. Essentially the score for an instrument represents the weighted sum and remainder of a set of binary values.

5. These results were obtained as a by product of computer programs written to implement weighting functions and alphabetic grade-code conversions. It was merely necessary to add variance calculations and confidence interval calculations and grouping operations to produce the above results.

6. It should be noted that considering such factors for one student violates what we call the principal of "horizontal equity". This term, widely used in public finance theory in connection with tax loads, simply means that one must treat equals alike. As it is applied in taxation much attention is given to determining classes of equals so that they can be taxed alike. In connection with educational evaluation we interpret its application to mean that unless information of the same type is available and considered for every student with the same weight, it should not be considered for any. For example, consider the student who "blows" the final exam, then comes to you with the news the next day that a relative died or that he had a case of the 24-hour flu prior to the exam. To adjust his grade by assigning a positive weight to the new information would violate the principle of horizontal equity unless you accumulated information on the state of all other students who are included in the ranking with respect to these two conditions. Rigid adherence to this principle would, of course, have the very undesirable effect of preventing the evaluator from considering information of relevance for a decision.

7. For the results of a simulation model of the educational process which is designed to measure the impact of alternative sampling plans, weighting functions, and aggregation procedures on the final amount of human capital created by ascertaining the effect on decisions on who remains in the system and who is ejected from the system, see C. J. Goetz and C. Schotta, Jr., "Quality Control in the Production of Human Capital: A Simulation Study," paper to be presented at the Operations Research Society of America meetings, Philadelphia, November 4, 1968. This model ascertains the results of items in such a way that exam scores are approximately zero variance data since the "value" of each simulated "mind" in the simulated "population" can be ascertained and compared with the "value" derived from the instrument. The simulation study is based on the central theme of our earlier paper, Schotta and Hoffman, "A Priori Decision Functions for Education Evaluation" presented at the Operations Research Society of America meetings, New York, May 31, 1967.