DOCUMENT RESUME

ED 060 576 EA 004 190

AUTHOR Barrows, Thomas S.
TITLE Evaluation Problems in Performance Contracting.
PUB DATE 28 Jan 71
NOTE 11p.; Revision of a speech given before New York State Council for Administrative Leadership. (House of Representatives, January 28, 1971)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Educational Experiments; *Educational Innovation; *Evaluation Criteria; *Evaluation Methods; Item Sampling; *Performance Contracts; Speeches; Testing; *Test Interpretation; Test Selection

ABSTRACT

This speech discusses performance contracting as educational research, notes some evaluation problems, and proposes solutions to these problems. The term performance contracting, according to the report, denotes an administrative rather than an instructional innovation. The author observes that the understanding of instruction and learning derived from contracts is minimal because strategies are chosen on atheoretical bases and no provisions for generalization are made. Aggregate performance indexes are presented as superior to individual ones, and item sampling is suggested. (Author)

EVALUATION PROBLEMS IN PERFORMANCE CONTRACTING*

Thomas S. Barrows
Educational Testing Service
Princeton, New Jersey

I would like to cover four topics this morning which may be of some use to administrators as they approach performance contracting with the misgivings evidenced in this meeting so far. I am sorry that my remarks are not intended to decrease the anxiety you may feel in the face of this heralded innovation. Perhaps, however, the four points will afford some comfort in allowing you to fix your general anxiety to specific aspects of performance contracting as it is currently practiced.

The four points stated as brief questions are:

> First, may we consider performance contracting to be an <u>educational</u> practice and, therefore, inquiry into its processes and effects to be educational research?
>
> Second, are there some general problems of evaluating educational programs that trouble performance contracting?
>
> Third, are there some evaluation problems that are at least partially peculiar to performance contracting?
>
> Fourth, what might be a partial solution to these problems?

I. Addressing the first question, let me point out that performance contracting is not an <u>educational</u> innovation or practice but an

*Revision of an address presented to the New York State Council for Administrative Leadership, House of Representatives, January 28th, 1971.

administrative one. Performance contracting per se is a method of funding as can be seen by examining current performance contracting projects for a common characteristic. It is not the use of incentives or machines or strangely qualified instructional personnel although contractors who can offer these sorts of things seem to have been attracted in many cases.

Perhaps this would have been immediately obvious had we initially thought to ask what Mr. Blashke could have brought from the Department of Defense that might influence what goes on in classrooms or other instructional settings. It is possible to hypothesize that payment for results might somehow lead to increased positive motivation on the part of instructional staff, that this in turn might lead to new and beneficial instructional behavior, and that new learnings might occur as a result. Note, however, that this is a long and tenuous causal chain and that, were we interested in understanding it, we would most certainly not start our studies with its extreme ends. Rather we would look first at changes in motivation of the instructional staff, then at changes in instructional behavior, and so on, testing each hypothesized causal link. Such a series of studies might be valuable if cast as investigations of the effects of various motivational strategies upon teaching personnel and thence upon student achievement. Note, however, that such a series of studies would specifically entertain the possibility of unintended, including negative, outcomes.

I mention negative outcomes and side effects partially because evaluation efforts cannot afford to be blind to them any more than

research can, and partially because my own bias leads me to suspect that negative outcomes are extremely plausible when performance contracting is carried out as it is now. I suggest this because it seems to me that current performance contracting projects are more easily construed as applications of negative reinforcement (e.g. punishment and/or threat) than as applications of positive reinforcement. I am led to this point of view by performance contracting's introduction at a time when most of education's various constituencies are disparaging of education's adequacy. In addition, most of us by now have heard of projects that have been installed in schools or systems without adequate teacher involvement and of how the teachers have received this sort of thing. It may not be hard to understand this reaction when we recall the threat of being displaced that teachers perceived in the early days of programmed and computer assisted instruction.

All of this cannot be made into a conclusive argument for the evidence has not been collected. If, however, my hypothesis were correct--that teachers and other instructional staff do indeed view payment for results as threatening, as a system of negative reinforcement and a potential displacement of the teacher's role--then we would expect the confused and dysfunctional escape responses which usually result from threat and negative reinforcement. I remind you again that all this is unsupported theorizing but could it be that the teaching of test items that occurred in Texarkana is indicative of a general escape response?

To return to the first issue - performance contracting's adequacy

as a focus for evaluative research - and having said that performance contracting is not per se an educational practice or treatment, let us consider the educational programs that have been funded under its aegis. Here we should ask if evaluative study of these instructional programs has value as research. Again I must be negative for the programs are selected on bases which are atheoretical. Instructional techniques are chosen on judged potential and cost rather than as representative treatments in the context of testable hypotheses and generally applicable theories of instructional effectiveness. The evaluative studies thus ask whether an isolated treatment works--not why types of treatments work--and, therefore, little or no information is produced which is transferable or applicable beyond the particular contract.

Although I have thus far been negative about the contributions of performance contracting to our knowledge of education, I do not mean to suggest that performance contracting is not worthy of evaluators' attentions. Clearly evaluation is a crucial ingredient no matter what the value of the practice so let us now turn to general and then specific problems in evaluating contractors' performances.

II. One of the first questions that arises is what criteria shall be used for evaluating the contractor's efforts. We have seen both criterion referenced and standardized normative instruments proposed. Criterion referencing of individual test items to behavioral objectives has certainly gained popularity but it seems to me that some more thought should be expended before we let the current enthusiasm carry us off. There are two points of view on how to approach

the evaluation of instruction. The first assumes that the purpose of instruction is to cause learning of the material presented. The second assumes that the purpose is to influence the development of broader psychological traits which are applied--sometimes after adaptation--in a variety of contexts and to a variety of content. The first point of view which in my mind corresponds to training is served well by evaluation instruments that are referenced to the specific objectives and content of instruction. The latter point of view which I see as education requires that evaluation be based upon changes in the degree to which learners possess both desirable and undesirable traits.

Which approach should be adopted in evaluation? It seems to me that both are necessary. If a program does not provide the specific skills and learnings intended, its value is certainly in question. If, to take the second point of view, broader traits are not influenced, the value of the program for the accomplishment of education's broader developmental goals is in question. If we knew that skill development and changes in broader psychological characteristics were causally interrelated in invariant ways, then both types of criteria would, of course, not be necessary. It seems that things are not that simple however and so we must employ both criterion and trait referenced measures.

A second point that is often forgotten is that we should view each measure as an imperfect indicator of status with regard to the underlying content or trait referrent. Thus we should not make the seductive mistake of equating fallible scores with status on the constructs we wish to measure. An ITBS reading score for example, is an estimate of a

subject's reading ability not his actual reading ability. This point becomes more important when we leave classic achievement and ability measurement for the more esoteric areas of personal/social development and affective reaction. Here we have much less experience from which to judge instruments' validities and there is much less assurance that we are actually measuring the trait we are interested in. We should, therefore, be especially careful in selecting instruments and our evaluations should be designed to provide reliability and validity data when these are otherwise unavailable. Clearly we are not producing much information regarding the effects of an educational treatment if we do not know how well we are actually tapping the criteria intended.

Thirdly, the usual problems of developing sound evaluation designs within the natural settings in which we find educational "experimentation" must be overcome if we are to be able to draw firm conclusions about the effectiveness of performance contracts. The literature of evaluation and the current requests for proposals that we see indicate that a basic misunderstanding of the principles randomization, matching, and covariance are with us still and have been carried into evaluation of performance contracts. Random assignment to treatment or to control conditions establishes initial similarity within chance limits for all characteristics and allows subsequent dissimilarities to be attributed to differential intervening experience. On the other hand, matching and covariance assure similarity on the matching variables or covariates alone and are, therefore, properly employed to achieve experimental precision rather than as substitutes for randomization. The advice is

clear - employ randomization - but it is also impractical for one often cannot. In such situations a large number of quasi-experimental designs are available. Reliance upon them should be guided, however, by an intimate knowledge of their strengths and weaknesses. In most cases school districts would be well advised to seek advice on design for it is the true sine qua non of evaluative information.

Finally, before moving ahead into an evaluation, we must recognize the assumptions of the analytic model proposed for use. One of these assumptions seems especially offensive to me. That is, education is usually modeled as a conjunctive process rather than as a disjunctive one. In a conjunctive educational process, goals are interrelated with "and" so that each student is expected to move toward each goal. In a disjunctive scheme,"either" and "or" connect the goals. For example, tennis is a conjunctive sport--one must serve and volley, etc. Football is disjunctive--one may be either a great passer or runner or lineman, etc. While a conjunctive model may fit the aquisition of some skills, there are many educational treatments which are intended to impart different things to different students. In graduate school I remember realizing that one could either succeed through a knowledge of content or through methodological sophistication. There are countless other examples--in fact I would argue that the disjunctive model is more widely applicable--yet every educational evaluation that I have seen has used analytic techniques which assume a conjunctive model. It is not entirely clear how to solve this problem. Although

the basic approaches have been specified, the techniques have not been developed to the "cookbook" stage which seems necessary to broad use.

So much for general evaluation issues that one must be wary of in performance contracting. Now let me turn to evaluation issues that have arisen specially within the context of performance contracting.

III. One intriguing characteristic of the evaluative work connected with the performance contracts that we have seen so far is the separation of analyses for remuneration and for program evaluation. In the request for proposals generated for the recent OEO study, for example, payment was to be based upon absolute pretest postest differences while program evaluation was to involve treatment and "control" group comparisons. This practice makes little sense. Given the two types of analyses in the example it could be that one or the other is superior or that they are in some way complimentary. In any instance either the single superior analysis or the complimentary combination should be used for both basically evaluative purposes.

A second phenomenon seen repeatedly in performance contracting is the use of individual difference scores for the determination of payment. This type of score is notoriously unreliable depending on complex interrelationships between the first and second tests. It has little to recommend it.

Thirdly, the common use of an all or none cutoff score for payment seems a poor choice. While a moderate amount of statistical sophistication is necessary to handling the unreliability of

distributions of difference scores, the problems gain in complexity when our knowledge must be extended to estimate the unreliability about one specified gain score. I think we should also wonder what teaching strategies are promoted by a cutoff. If teachers were to behave in a way which should maximize payments they would stop teaching students who had reached the cutoff in order to focus on others. This might or might not be desirable pedagogically.

Finally, the problem of teaching test content presents itself. Evaluation requests have suggested guarding against this by spiraling multiple tests and by not allowing instructional personnel to know which students received which test. The mechanical procedure is cumbersome at best. It may not be ineffective for it is still possible for instructional personnel to ascertain what test each child has been given and will get by asking him which sets of pretest items are familiar to him. In the spiraling situation it is also possible that different tests will provide different payoff rates--a situation bound to stir controversy at least.

These four problems in the evaluation design that has repeatedly been suggested in proposal requests needlessly hinder those who might propose an efficient evaluative study. Let's turn now to what might be some of the more important characteristics of such a study's design.

IV. In an efficient design, program evaluation and payment would both be tied to aggregate score differences. These might be

pretest/posttest comparisons, or experimental/control comparisons, or both depending on what type of a design could be assembled within the situational constraints present. Comparisons could be based on means, on entire distributions, or on some combination of points in the aggregate's distributions. For example, a school district might be willing to pay x dollars for a mean achievement level of y and 1 1/2 x for 1 1/3 y. The point is to base payment and evaluation on estimates of group performance which are inherently more reliable, less error-laden, than are estimates of individual performance.

If we can move to aggregate indices of performance, we may next solve a host of problems by employing a matrix sampling approach. It is possible to derive adequate indices of group performance on a pool of test items by giving differing samples of items from the pool to each student in the group. Furthermore, in a pretest/posttest design, it is not necessary to administer identical items to individual students at each testing. What advantages are thus immediately available?

First, the pool of items to be used may be much larger than in a traditional test administration for we are freed of the testing time constraint on number of items. A large item pool is beneficial in allowing us to more adequately cover admittedly large curriculum areas such as reading where we have traditionally had to undersample. Furthermore, should a teacher try to coach students on such a large pool she would have to teach something approximating the true criterion. This,

of course, no longer has the objectionable quality that coaching does of destroying the representativeness of the behavior sample obtained by the test.

Second, there is no possibility of instructors finding out what test items are to be administered to individuals at posttest for this is not a function of pretest content. Each student is administered a random subset of items each time. Teaching to the test is further obviated.

Matrix sampling thus seems to be a possible way to solve some of the evaluation problems that are presented by performance contracting. It certainly will not solve all or even a majority of them, but it is not presented for that purpose. Rather, it should be considered as but one example from the evaluator's kit of tools. When imaginative and competent use is made of that kit's contents the educational community benefits from more efficient and effective evaluation and evaluative research. The more its use is constrained needlessly by restrictive requests for proposals, the longer we will all have put up with the haunting suspicion that contractors are getting something ror nothing.

In conclusion, let me state my own bias concerning the importance of performance contracting and by extension accountability to the future improvement of education. Ed Zigler noted that children learn for the same reason birds fly. When they don't - he said - we should ask why. The question of what to blame is, I think, much more important than whom to blame.