DOCUMENT RESUME

ED 147 360 TM 006 814

TITLE Why Should All Those Students Take All Those Tests? (Every-Student Testing or Sampling of Selected Groups?).
INSTITUTION National Education Association, Washington, D.C.
PUB DATE May 75
NOTE 10p.; For related documents, see TM 006 646, ED 084 641, 091 821, and 092 571; also available in ED 146 233
AVAILABLE FROM National Education Association, 1201 16th Street, N.W., Washington, D.C. 20036 (Free of Charge)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Cost Effectiveness; Educational Accountability; Elementary Secondary Education; Evaluation Methods; Group Tests; Item Sampling; *Sampling; School Districts; *Standardized Tests; State Programs; Student Testing; *Testing Problems; *Testing Programs
IDENTIFIERS Alternatives to Standardized Testing; *National Education Association

ABSTRACT

The National Education Association's Task Force on Testing has stated its opinion that standardized tests are overused. The task force suggests that the application of sampling techniques and a variety of alternatives to current testing practices would accomplish the same purposes. Representatives of the testing industry have indicated that the sampling of student populations could be equally effective as the blanket testing of every student. Sampling procedures would also assure the rights to privacy, and conserve time, effort, and cost. Methods for determining whether or not sampling should be used are presented, along with a brief discussion of item sampling. (Author/HV)

WHY SHOULD ALL THOSE STUDENTS TAKE ALL THOSE TESTS?

(EVERY-STUDENT TESTING OR SAMPLING OF SELECTED GROUPS?)

Published by the

NATIONAL EDUCATION ASSOCIATION
1201 16th Street, N. W., Washington, D. C. 20036

May 1975

WHY SHOULD ALL THOSE STUDENTS TAKE ALL THOSE TESTS?

(EVERY-STUDENT TESTING OR SAMPLING OF SELECTED GROUPS?)

The NEA Task Force on Testing, in its first interim report, states:

> The Task Force believes there is overkill in the use of standardized tests and that the intended purposes of testing can be accomplished through less use of standardized tests, through sampling techniques where tests are used, and through a variety of alternatives to tests....
>
> Representatives of the testing industry and others told the Task Force that sampling of student populations could be as effective as the blanket application of tests that is now so common. Some suggested that such procedures, in addition to increasing the assurance of privacy rights, would conserve time, effort, and financial expenditure.[1]

The blanket use of tests (every-pupil testing) in some state assessment and local testing programs appears to require inordinate amounts of time and resources on the part of teachers, other personnel involved in test administration and interpretation, and the students themselves.

Criticisms of the blanket use of tests have come from a variety of prominent researchers, evaluators, and other educators.

House, Rivers, and Stufflebeam, in their evaluation of the Michigan accountability system, concurred that in that state:

> Statewide testing as presently executed also raises the question of the feasibility of every pupil testing. This practice appears to be of dubious value when the cost of such an undertaking is compared with the resulting benefits.

[1] In Task Force and Other Reports presented to the Fifty-Second Representative Assembly of the National Education Association, July 3-6, 1973, Portland, Oregon. pp. 26-46.

> to local level personnel.... The local, and hence overall, costs could be reduced by a matrix sampling plan which requires that each student tested take only a few items.... In the long run, a matrix sampling plan will be the only one feasible from a cost and time standpoint. The cost and time required for every pupil testing for the whole state would be horrendous.... We feel that it /strict adherence to a statewide testing model/ will result in useless expenditures of monies and manpower, in addition to producing unwarranted disruptions of the educational programs within a great number of schools.[2]

In a paper entitled "Criteria for Evaluating State Education Accountability Systems," the National Education Association has laid down fifteen basic principles, one of which is as follows:

> If the state desires test data for its own planning purposes, it should use proven matrix sampling techniques which will not reveal schools and which will greatly reduce costs.
>
> Matrix sampling techniques can give an accurate picture of the state by various categories much more efficiently than testing each child with an entire instrument.[3]

It was with such admonitions as these in mind that this paper was developed. And while some procedures are appropriate for evaluating all students in one way or another for particular purposes, it would appear that there is gross over-use of blanket testing procedures.

To help teachers and other educators better understand some main considerations related to sampling, the NEA obtained permission from Dr. Frank Womer, Michigan School Testing Service, University of Michigan, to reproduce

---

[2]House, Ernest R.; Rivers, Wendell; and Stufflebeam, Dan. An Assessment of the Michigan Accountability System. Michigan Education Association and National Education Association, March 1974. pp. 14-16.

[3]National Education Association. "Criteria for Evaluating State Education Accountability Systems." Washington, D. C.: the Association, n.d...

material from a monograph of his on developing assessment programs.[4] In addition, Dr. Womer prepared, especially for this paper, a section on item sampling. Dr. Womer's recommendations follow.

* *

Determining Whether Sampling Is To Be Used

The decision whether to test an entire population or use a sample involves a combination of concerns. Clearly there are policy considerations; clearly there are psychometric[5] considerations; clearly there are data collection considerations; and clearly there are cost considerations. The best possible staff and consultant thinking on this question should be brought to an advisory committee for them to consider very carefully.

Probably the most crucial consideration is a policy one, since psychometrics, data collection, and cost generally would argue on the side of sampling rather than using an entire population. If it is deemed wise for policy reasons to test all students in a population, that preference, typically, will have to be weighed against available resources and technology; so we will consider first the policy implications of the two choices.

One needs to look carefully at the purposes and goals of a specific assessment program in determining whether sampling is appropriate. If all of the specific purposes and objectives of an assessment program can be met by group results, then sampling must be considered.

---

[4] Womer, Frank B. Developing a Large-Scale Assessment Program. Denver: Cooperative Accountability Project, 1973.

[5] Editor's note: Psychometrics in the strictest sense of the definition has to do with the measurement of mental abilities. It has come to be used much more broadly to define a wide range of activities in assessment and evaluation.

The only assessment situation that clearly calls for common data collection on all members of the population is when it is deemed essential, for improved decision making, to have exactly the same test information for every pupil in a given grade in a state (or other assessment unit). It is exactly this situation that has prevailed for years in local school districts that have every-pupil achievement or ability testing at some grade level. Historically, the compulsory state testing programs were examples of this situation; the voluntary programs were not. If a state mandates common testing for all students it is taking over a role that local districts traditionally have held. This may be good or this may be bad depending on one's point of view of the role of a state department of education. It certainly has important policy implications.

There are many facets to this point, but it should be kept clearly in mind that it is not necessary to test every pupil at a given grade level on identical material in order to get a good picture of education outcomes of groups of students; it is necessary only if one feels that each teacher in an entire state at a given grade level must have the same information for each pupil.

Probably the greatest advantage of sampling is that for a given amount of effort (and money) one can gather more usable information than by using an entire population. If the goals of an assessment program are to gather statewide information only, it is hard to conceive of any reason for testing all students in a given grade. For example, if there are 50,000 third-graders in the state of Limbo, and one wants to gather state statistics only, it is very possible that a sample 5,000 students (or even 500) would

be sufficient if they are selected by a probability sample....[6] Or, if one can afford to test all 50,000 third-graders, and if it is deemed wise to do so, one could select ten 5,000-pupil samples and secure information on ten subject areas, or one could go into great depth of information gathering in two or three subject areas. The combinations of possibilities of sampling pupils and content are almost endless.

If one wants district-level information, then sampling becomes a different situation. In a school district with one third grade, sampling of pupils is hardly possible for most assessment purposes. In school districts with many third-graders, sampling could provide a greater variety of information than common testing on every pupil, in the same fashion as at the state level. Specific decisions of how far to carry sampling should be made only after advice from a sampling statistician. Sampling is a highly developed technical field, and the implications of any decisions to sample or not to sample must be reviewed by competent samplers.

Other "compromise" possibilities exist. One could test all students in a population with one short test, while using a sampling approach for other tests. This approach would provide some common information on all students but would allow for greater depth of data collection over a subject area.

> Principle: Sampling of pupils and/or content should be given very serious consideration for all large-scale assessment projects. The only situation where it may not be useful is one where it is deemed essential to collect common information on all students in a statewide population

[6] Editor's note: For information on probability samples, see Womer, op. cit.

of students. Sampling should be used to maximize the collection of usable information for stated assessment purposes at the lowest possible cost and effort.

* * *

Sampling with total tests is less complicated to administer, but since it is likely to be subject to error in administration and consequently less reliable, in some cases item sampling may be more useful. Therefore, Dr. Womer was asked to prepare an additional statement on the purposes and potential of item sampling. His statement follows.

Item Sampling

The process of item sampling in testing is more useful for one of two purposes:

1. to increase the amount of group test results that can be obtained from students in a given period of time; or
2. to decrease the amount of testing time necessary to obtain large amounts of group test information fror students.

For either purpose, it is essential to keep in mind that item sampling is useful for gathering information about groups of students. Thus it is

a technique for use with relatively large groups, not a classroom-sized group or even three or four classes within a building.

Example 1

A school system has 500 students in the sixth grade. A standardized reading test is to be administered for a one-shot systemwide survey. The test takes 45 minutes to administer, which is all the time that can be taken from a busy schedule at the end of the year.

Staff are unhappy that only reading is to be surveyed. Some major changes were made in the mathematics curriculum three years before and they feel it would be valuable to survey this subject also. By randomly selecting only 250 of the students to take the reading test, the other 250 could be given a 45-minute mathematics test at the same time.

Example 2

A school system has 1,000 fourth-graders. It is desired to do an in-depth study of student outcomes for 100 different behavioral objectives in mathematics. Each objective requires the use of eight questions. The total of 800 questions would require one student to spend perhaps 15 hours of testing time to attempt all of them.

By randomly dividing up the objectives and items into five different subtests (each with 20 objectives and 160 items), each subtest could be administered to 200 students (randomly

selected). This would require only 3 hours of testing time per student (manageable) rather than 15 hours (unmanageable), and group results would still be available for all 100 objectives (800 items).

In either example the results will be usable for group analyses. Any slight reduction in accuracy due to sampling error is apt to be much less than errors due to increasing testing time of students beyond some reasonable amount. Systematic errors due to fatigue, disinterest, poor motivation, teacher concern, and other conditions of testing can easily outweigh a small sampling error.