DOCUMENT RESUME

ED 099 408 TM 004 062

AUTHOR Coffman, William E.
TITLE A Moratorium? What Kind? NCME Measurement in Education. Vol. 5, No. 2, Spring 1974.
INSTITUTION National Council on Measurement in Education, East Lansing, Mich.
PUB DATE 74
NOTE 8p.
AVAILABLE FROM NCME, Office of Evaluation Services, Michigan State Univ., East Lansing, Michigan 48823 (Subscription rate: $2.00 per year; single copies $0.35 ea. in quantities of 25 or more, or $0.50 for a single issue)

EDRS PRICE MF-$0.75 HC Not Available from EDRS. PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; Educational Accountability; *Educational Testing; Groups; Scores; Standardized Tests; *Testing Problems; *Test Interpretation; Weighted Scores
IDENTIFIERS *NEA Moratorium on Testing

ABSTRACT

Many problems in the areas of test interpretation and educational assessment are causing difficulties for educators. On one hand the public and legislators are requesting more state testing programs and assessment programs, while on the other, educators realize the problems concerning testing and test interpretation. Difficulties arise when tests are misinterpreted and misused. A proposed moratorium by the National Education Association is not the answer to the problem since it would destroy the continuum of data and create a critical information gap. Reporting systems based on criterion referenced measurement, the use of computers to find patterns from which to generate interpretations, and further use of adjusted scores can help to alleviate some of the problems. A moratorium on testing would only destroy the continuum of data and create a critical information gap. (Author/SM)

NCME
measurement in education
A SERIES OF SPECIAL REPORTS OF THE NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION
Vol. 5, No. 2 Spring 1974

A MORATORIUM? WHAT KIND?*

William E. Coffman

These are critical times for educational measurement and evaluation. On the one hand, the public is deeply concerned about the quality of education and is demanding that educators become more accountable for the quality of their programs. On the other hand, educators are concerned that increasing emphasis on standardized tests is encouraging teachers to emphasize those outcomes of teaching that are easy to measure and to neglect those outcomes that are hard to measure—or for which no standardized measures exist. On the one hand, legislators are passing laws calling for the establishment of state testing programs. On the other hand, the National Education Association is calling for a moratorium on the use of standardized tests in the schools.

To the layman, it seems only natural to use standardized tests to find out what the schools are accomplishing. After all, aren't the measures able to tell us just where each child stands in relation to the norms? Don't high scores mean we have good schools and low scores that we have poor schools? To the teacher trained in educational measurement, the interpretation of standardized test scores is not so simple. Groups of children differ widely in what they know when they first come to school, and what one can teach a child in a year depends to a considerable degree on how much he knows at the beginning. Furthermore, there's the problem of bias in tests. No standardized test covers all the things a school is trying to teach, and sometimes the tests include things the school, for some very acceptable reason, has decided not to emphasize.

Tests are commonly used in other professional areas, but there is less likelihood than in the educational setting of their results being misinterpreted. For example, one isn't likely to oversimplify the job of interpreting the results of the usual medical test.

This paper is a modified version of the Presidential Address at the annual meeting of the National Council on Measurement in Education, New Orleans, February 27, 1973.

William E. Coffman

BY MICRO FICHE ONLY HAS BEEN GRANTED BY
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

ABOUT THIS REPORT

Along with all the other crises we're bombarded with, here comes the "measurement crisis." Dr. Coffman describes the conflicts besetting the field: Public cries for accountability, but educators' fears of testing what is easy to measure rather than what is important to know. Legislation of state testing programs, on the one hand, and, on the other, the call by the National Education Association for a moratorium on the use of standardized tests in schools. And the always present difficulties of test interpretation leading, too often, to misinterpretation and misuse of results.

The proposed NEA moratorium, Dr. Coffman believes, is not the answer to the problems. He discusses as possible solutions improved reporting systems, perhaps based on criterion referenced measurement (more broadly developed than at present, however, he argues); perhaps making greater use of computers to find patterns from which to generate interpretations; continued exploration of the use of "adjusted scores."

Dr. Coffman concludes with his idea of the kind of moratorium which would be beneficial to students, educators, and parents—not a halt to testing, but a halt to misuse of tests and test results.

Dr. Coffman, 1972 NCME president, knows his field from several angles—high school teacher and principal, and student and developer in the test and measurement field. Now Lindquist Professor of Education and Director of the Iowa Testing Programs at the University of Iowa, he was associated with the test development program at the Educational Testing Service in Princeton, N. J., from 1952 to 1969, serving as associate director and then director of test development, and subsequently as director of research and development, College Board Programs Division. He is currently a member of the Analysis Advisory Committee, National Assessment of Educational Progress, and a consultant for other research and measurement projects.

Would anybody think that the doctor with the healthiest patients was necessarily the best doctor? Certainly not, if he stopped to consider that the most seriously ill patients would be most likely to seek out the physician with the best reputation. Do patients generally demand that the doctor tell them the results of their medical tests in numbers? Not often. Generally, a patient expects his physician to tell him what the results mean; he isn't likely to make the mistake of thinking that he knows what the numbers mean. In contrast, nearly everybody feels confident he knows what a grade score of 6.3 on a reading test means. This is unfortunate, for the meaning of a score on an educational test is every bit as difficult to interpret as the numbers used in recording the results of a medical test.

DISTORTED EXPECTATIONS

I don't think I'm misrepresenting the situation when I say that there is today a serious tendency to oversimplify the interpretation and use of educational tests. Certainly, the lay public—even the elite of that public—have distorted conceptions of what educational tests can do. As Dyer (1973) reported:

> In 1971 the education committee of one of the state legislatures came up with an educational accountability bill that read in part as follows: "If the performance of any school district on any test approved by the state board of education—does not equal or exceed the national performance average for such a test for two successive years, said school district shall not receive any further state financial assistance—until such time as said school district has achieved such national performance average."

"...the gap has been widening between the increasing sophistication of the test makers and the understanding of the test user..."

The fact that the bill did not pass suggests that legislators can be brought to see the error of their ways; but the job of communicating—not only to the lay public, but even to the users of tests within the profession—is a monumental one. Even a superficial search through the testing literature of the last half century will turn up papers in every decade calling attention to the extensive misuse of test results on the part of the profession, and as Dyer so eloquently points out, the gap has been widening between the increasing sophistication of the test makers and the understanding of the test user, who just can't find the time to keep up with the developing literature. We're simply not going to educate the test user to all the subtleties of test interpretation; we're going to have to design more fool-proof reporting systems.

Incidentally, don't think that the problem of keeping informed is one that hounds only the educational profession. Ask your general practitioner how much time he has to keep up with the medical literature and to learn to interpret the results of the hundreds of new tests that are becoming available all the time. Or ask your engineer friend how easy it is to maintain competence in a field where the knowledge explosion is rendering today's competence obsolete in five years.

"We're simply not going to educate the test user to all the subtleties of test interpretation; we're going to have to design more foolproof reporting systems."

ENOUGH, ALREADY!

It may have been, at least in part, a recognition of this widening gap that motivated the NEA to approve the resolution calling for a moratorium on standardized testing. If I read correctly the reports of the discussion that preceded the voting and the explanations that have followed, there was no intention of condemning out of hand the use of tests to monitor the progress of children through the educational system. In fact, there seemed to be a clear recognition of the responsibility of a school system to render an accounting of what it has been up to.

Furthermore, another resolution approved at the same time called upon superintendents and boards of education to refuse to hire graduates of institutions that failed to include in their program of teacher education training in the interpretation and use of standardized tests.

The delegates seemed to be saying, "The gap between what the test makers know about the limitations and possibilities in standardized test scores and what the users are able to apply has become so wide that more harm than good is resulting from the use of standardized tests in the schools. Let's put a stop to it all until the test makers can come up with packages less subject to misuse or until the profession can develop the sophistication needed to prevent serious misinterpretations."

DRAWBACKS OF MORATORIUM

I'm afraid, however, that the solution proposed by the NEA is too drastic. Can we honestly say that we ought to take away from the thousands of knowledgeable and sensitive test users information on which they have come to depend? Or to create a gap in the accumulated record of pupils' progress through the educational system that may some day permit insightful researchers to create a picture of the ebb and flow of the educational tides of the 1970's?

In many instances we're just beginning to learn how to organize our accumulating data files so that the hidden relationships can be worked out. To abandon the data collection because somebody might misuse it would be a mistake of the first magnitude, for it is primarily through a study of patterns of changes in test

"To abandon the data collection because somebody might misuse it would be a mistake of the first magnitude..."

performance over time as changes occur in the inputs and treatments that the researcher is able to form judgments about what may be happening and why. And it won't do simply to select a representative sample of schools on which to collect data. To some extent each school system is unique, and to understand

"To some extent each school system is unique..."

what is going on in a school system, it's necessary to have a finger on the pulse of that system, not simply to apply blindly what has been learned from studying another system.

IMPROVE REPORTING SYSTEMS

Some diagnosticians have suggested that the first thing that needs to be done is to get rid of reporting systems that are subject to misuse. For example, the current draft of the "Standards for the Development and Use of Educational and Psychological Tests" includes this recommendation:

> Interpretive scores which lend themselves to gross misinterpretation, such as mental age or grade equivalent, should be abandoned or their use discouraged. VERY DESIRABLE (APA, 1973)

In general, I think I agree with this recommendation. Even people with considerable experience in dealing with quantitative data can be misled by grade equivalent scores. For example, the report on the study of equality of educational opportunity prepared by James Coleman (1966) and his associates talks about blacks getting further and further behind as they go through school, and the statement is repeated in the recent book by Mosteller and Moynahan (1972).

On the other hand, teachers report, with some justification, I think, that the grade score does give some hint of the level at which one should begin teaching a child, a kind of information not readily inferred from either percentile ranks or standard scores, even those based on a longitudinal growth model. If we are to introduce numbers that permit us to chart the growth of children through the educational system and at the same time avoid the dangers

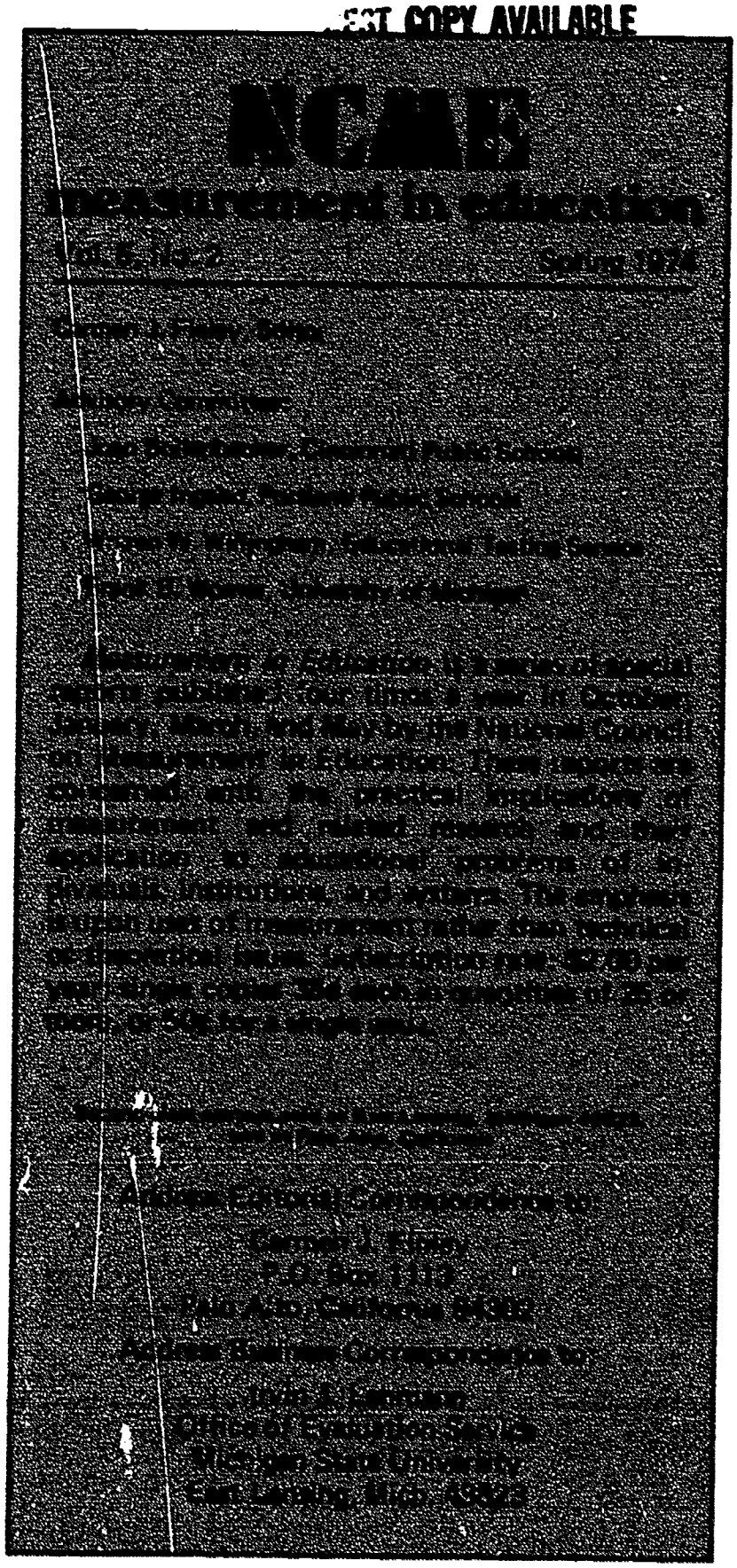
"(Criterion referenced measurement) has generated some output that is potentially as dangerous to education as grade equivalent scores."

inherent in a grade equivalent system, we're going to have to do two things: (1) provide a way of going from the new system back to the old in order to preserve the continuity of the record, and (2) provide a means of bridging the gap between the numbers and the decisions they imply about what we should be doing with individual pupils as a result of knowing their test scores.

CRITERION REFERENCED MEASUREMENT

I suppose the obvious solution to providing interpretation is to provide a criterion referenced interpretation, and let me say that this isn't nearly so modern an idea as some people would have us believe. Even before 1920 E. L. Thorndike was examining the problem of educational measurement in the broader context of measurement in the sciences and concluding that educational measurement would follow physical measurement in developing a system where each score had an immediately meaningful referent. From our perspective of the 1970's we can see that it was probably a mistake to develop scores that answered the question, "What group is this person's performance like the average of?" The discussion in recent years of criterion referenced measurement has, on the whole, I think, been salutary. At the same time, it has generated some output that is potentially as dangerous to education as grade equivalent scores.

The problem, as Krathwohl and Payne so clearly point out in their chapter in the second edition of *Educational Measurement* (1971), involves the conflict between skills which are important to learn, and skills which are most easy to teach and to measure. The kinds of learnings that are most important are those that involve complex skills and understandings and thus that develop slowly over the years. The learnings that respond most rapidly to instruction and are easily demonstrated through responses to test items involve simple skills and recall of information that serve primarily as vehicles for the development of more complex learnings. To the extent that measurement deals only with the simple learnings and ignores the more complex, it encourages the training of simple responses without emphasizing the fact that the way in



which the simple learnings are developed may be crucial for the development of the more complex ones. Thus, to specify, as some have, that the criterion of an acceptable item for a criterion-referenced test is that it be responsive to teaching (over the short pull), is to reinforce a superficial concept of what education is all about.

SIMPLE SKILLS ESSENTIAL

This is not to say that there is no place for the development of simple skills. The problem of breaking the code in beginning reading is one that involves hundreds, if not thousands, of specific learnings. One reason there are so many failures in learning to break the code is simply that the task of checking to see that a particular child has noticed and understood the significance of each one of these critical elements is a monumental one for any teacher faced with 20 to 30 squirming first graders. Any system that will help the teacher to determine that the many messages have been received and responded to is likely to improve the teaching of beginning reading.

But let's not be confused. The reason for giving serious attention to instruction in reading is to get the child ready for the task of taking over, bit by bit, the responsibility, under the guidance of the teacher, of his own education. And one task of a good testing program is to chart the progress of the pupil toward this goal. Before long, the school must get on with the task of helping pupils learn how to use what they are learning in complex meaningful contexts, and the testing program must help professional educators know whether or not this task is being accomplished.

Most of the criterion-referenced tests coming on the market today seem to be concerned primarily with the assessment of progress in the building of a repertory of basic information and simple skills. Before long, we're going to face the necessity of providing criterion references for more complex learnings, or else serious educators will judge our testing profession irrelevant.

"Before long, we're going to face the necessity of providing criterion references for more complex learnings, or else serious educators will judge our testing profession irrelevant."

POSSIBLE SOLUTIONS

In this connection, I find the 1962 paper of Bob Ebel highly significant. You will remember that Bob proposed what he called "content standard scores"—scores that could be interpreted by referring to small groups of illustrative test questions. So far as I have been able to determine, no test publisher has taken up Bob's challenge, but I think it would be possible, using the procedure outlined by Ebel, to develop reference sets of questions for most survey-type tests that now provide only norm-referenced interpretations. The sooner we get going on this task, the better.

Another way of solving the problem of misinterpretation may be to have the expert provide more direct interpretation as part of the score report. Of course, the final decisions need to be made at the local level—by teachers and parents and pupils who have access to much more information about the individual and the learning situation than can ever be reflected in accumulated test data. However, just as the physician is coming to rely more and more on interpretations of test data provided by specialists or on interpretations generated after referring patterns of test scores to the information stored in the memory of a computer, so we may find it possible to make use of the computer to generate interpretative statements that provide an interface between the test scores and the user.

Teaching a computer to do this in a manner and with the qualifications necessary to insure that verbal reports do not become the "modern Gospel" just as grade equivalent scores have been in the past will be quite a challenge, but we are making a beginning. Tomorrow I will be reporting the results of the efforts of a team working in Iowa City last summer under the leadership of Professor Walter Mathews of the University of Mississippi. Before too long, I hope to be able to report on a field testing of a pilot system of computer-generated verbal score reports for the Iowa Test of Basic Skills.

INTERPRETING GROUP SCORES

When we turn from the interpretation of individual scores to the interpretation of scores for groups, we are faced with new problems. By abandoning grade equivalent scores we might get rid of the problem reflected in the newspaper report that "one-fourth of the pupils in grade six in Port City are two years retarded in reading." But whatever system is used, any attempt to make a direct interpretation of average scores for Port City will show that Port City's test performance is low. It's only a step from there to the inference that there must be something wrong with Port City's schools.

Now let's be clear on one point. Assuming we have some sort of criterion interpretation available, the actual performance of the pupils in Port City—or in a particular school in Port City or in a particular classroom in a particular school in Port City—will reflect clearly the educational problem faced by the city. The instruction needs to begin where the pupils are, and one needs to have information about where the pupils are if one is to design an effective educational program for them. But if the test results are to be used for purposes of accountability, the test scores must be placed in context. Some way must be found for answering the question, "Given all we know about the situation in Port City, how do these results stack up?"

Right now the most popular way of reporting test results for a school system is in terms of national norms. I'm afraid this is because such a comparison keeps the fun in testing for all concerned. Half of the

REPORTS AVAILABLE

Back issues of *Measurement in Education* are available at 35¢ each in quantities of 25 or more for a single issue.

Vol. 1, No. 1 *Helping Teachers Use Tests* by Robert L. Thorndike

No. 2 *Interpreting Achievement Profiles—Uses and Warnings* by Eric F. Gardner

No. 3 *Mastery Learning and Mastery Testing* by Samuel T. Mayo

No. 4 *On Reporting Test Results to Community Groups* by Alden W. Badal & Edwin P. Larsen

Vol. 2, No. 1 *National Assessment Says* by Frank B. Womer

No. 2 *The PLAN System for Individualizing Education* by John C. Flanagan

No. 3 *Measurement Aspects of Performance Contracting* by Richard E. Schutz

No. 4 *The History of Grading Practices* by Louise Witmer Cureton

Vol. 3, No. 1 *Using Your Achievement Test Score Reports* by Edwin Gary Joselyn & Jack C. Merwin

No. 2 *An Item Analysis Service for Teachers* by Willard G. Warrington

No. 3 *On the Reliability of Ratings of Essay Examinations* by William E. Coffman

No. 4 *Criterion-Referenced Testing in the Classroom* by Peter W. Airasian and George F. Madaus

Vol. 4, No. 1 *Goals and Objectives in Planning and Evaluation: A Second Generation* by Victor W. Doherty and Walter E. Hathaway

No. 2 *Career Maturity* by John O. Crites

No. 3 *Assessing Educational Achievement in the Affective Domain* by Ralph W. Tyler

No. 4 *The National Test-Equating Study in Reading (The Anchor Test Study)* by Richard M. Jaeger

Vol. 5, No. 1 *The Tangled Web* by Fred F. Harcleroad

systems can report that they're OK since they are above the norm and the other half can report that of course they aren't up to norm, but then look at how different the system is from the average system in the country. Everybody wins and nobody has to ask the hard question.

For most widely used batteries of achievement tests, the publishers have provided norms for various subgroups in addition to national norms—regional norms, norms for large cities, norms by IQ levels, and the like. But these don't really solve the problem of accountability. The logical end point of resort to differential norms is that the appropriate norm for system X is the results for system X since system X is unique. Besides, as the variety of norms proliferates, the task of interpretation becomes more complex. School administrators and politicians and the man-in-the-street can't be blamed for feeling that the testers are trying to obscure the meaning of the test scores. Some way must be found to provide a simpler way of evaluating the test results for a school system.

A CASE FOR ADJUSTED SCORES

Henry Dyer has proposed an evaluation model based on the application of multiple regression techniques (1970). The model has much to recommend it since it does take into account some of the variables that seem to account for differences among school systems but which school systems can't do much about. As Forsyth (1972, 1973) has pointed out, there are still some sticky problems in applying the Dyer model, but it is a step in the right direction. The fact that the model is being proposed for use in New York City and is apparently acceptable to the teachers (Shanker 1973) is certainly encouraging.

"Some way must be found to provide a simpler way of evaluating the test results for a school system."

Application of the Dyer model has the effect of substituting for the scores actually obtained, scores that have been adjusted for differences in the characteristics included in the model. This is not the only instance of efforts to develop adjusted scores.

In the summer and early fall of 1971 I had the opportunity to meet with a group under the leadership of Dr. Selma Mushkin of Georgetown University that was asked by the U. S. Office of Education to look into the feasibility of developing systems of adjusted test scores for reporting summaries of test results for school systems. I discovered that in many areas, statistical reports consist of numbers that have been adjusted to take into account differences from group to group that might obscure the real meaning of the data. For example, raw death rates for cities are not directly comparable because of differences in the age distribution from one city to another. To answer the question, "Which city has the lowest death rate, taking into account the age distribution in the population?" *adjusted death rates* are reported.

More recently, as a member of the Analysis Committee for the National Assessment Project, I've had the opportunity to observe how statisticians like John Tukey and Fred Mosteller go about the business of adjusting statistical data to reduce the likelihood that they will be misinterpreted.

I doubt that we can ever produce reports that are completely resistant to misinterpretation, but I do think that much more can be done to produce reports less subject to misinterpretation. If the resolution of the NEA has the effect of speeding up the development of such reporting systems, it will have had a salutary effect.

ONE KIND OF MORATORIUM

I guess I haven't been saying very much about what kind of moratorium I think is in order, have I? I've been talking about the need for more training in testing and evaluation for teachers and administrators, about more emphasis on criterion-referenced interpretations of test scores, about the need to report test results in ways that are meaningful to those for whom the reports are intended, about the possibilities of adjusting summary statistics for differences in inputs. I could also have talked about the need for developing measures of a whole lot of additional variables to insure that our evaluations are comprehensive, but Jack Merwin did that for us last year (1973) and all I need to add is "Amen".

I've questioned the desirability of calling a complete moratorium on standardized testing in the schools: to interrupt the data collection process while we perfect our evaluation system is to create a critical information gap. But I see nothing wrong at all with encouraging a moratorium on the use of test scores to label children rather than to guide their learning, to classify teachers rather than to identify points where teachers may be helped to become more effective, to pull the wool over the eyes of the public rather than to generate questions about how a school system might go about doing an even better job. Let's not spend too much time deploring the NEA's resolution; let's get on with the business of meeting their demands for better tests, better reporting systems and wiser test use.

REFERENCES

APA MONITOR, September, 1972.

Coleman, J. S., *Equality of educational opportunity*. 2 volumes. Washington, D.C.: Office of Education, U.S. Department of Health, Education, and Welfare, U.S. Government Printing Office, 1966, Volume 1, p. 21.

Dyer, H. S Can we measure the performance of educational systems? *National Association of Secondary School Principals Bulletin*, 1970, *54*, 96 105.

Dyer, H. S., Recycling the problems in testing, in *Proceedings of the 1972 Invitational Conference on Testing Problems*, Princeton, N.J.: Educational Testing Service, 1973.

Ebel, R. L. Content standard scores, *Educational and Psychological Measurement*, 1962, *22*, 15-25.

Forsyth, R. A. Considerations related to the usefulness of the performance indicators in Dyer's student change model of an educational system, Paper presented at the annual convention of the American Educational Research Association, Chicago, April 4-7, 1972.

Forsyth, R. A. Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system, *Journal of Educational Measurement*, 1973, *10* (1), 7-12.

Krathwohl, D. R. & Payne, D. A., Defining and assessing educational objectives, *In Educational Measurement, Second Edition* (R. L. Thorndike, Editor), Chapter 2, Washington: American Council on Education, 1971.

Merwin, J. C. Educational measurement of what characteristics of whom (or what) by whom and why, *Journal of Educational Measurement*, 1973, *10* (1), 1-6.

Mosteller, F. & Moynihan, D., *On equality of educational opportunity*, New York: Vintage Books, 1972. Page 15.

Shanker, A., Accountability: The hazard of blame-placing, paid advertisement in the *New York Times*, Sunday, January 7, 1973, Page E 11.

NCME NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

Second class postage paid
at Palo Alto, California

Office of Evaluation Services
Michigan State University
East Lansing, Michigan 48823

FORTNA RICHARD O
EDUCL TESTING SERV
PRINCETON NJ 08540

A series of special reports of the National Council on Measurement in Education