DOCUMENT RESUME

ED 117 161 TM 005 015

AUTHOR Boyd, Ja Mille; And Others
TITLE A Study of Testing Practices in the Royal Oak (Michigan) Public Schools.
INSTITUTION Royal Oak City School District, Mich.
PUB DATE May 75
NOTE 51p.

EDRS PRICE MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS Elementary Secondary Education; Information Dissemination; Interviews; Parent Attitudes; *Program Evaluation; Program Planning; *School Districts; Standardized Tests; Student Attitudes; *Student Testing; Teacher Attitudes; *Testing Problems; *Testing Programs; Test Interpretation; Test Results; Test Validity
IDENTIFIERS Michigan (Royal Oak); *Royal Oak Public Schools

ABSTRACT

The testing program of the Royal Oak (Michigan) School District was examined and evaluated under agreement of the Royal Oak Teachers Association and the Administration of the Royal Oak School District. A committee was formed of people knowledgeable of the program and/or specialists in evaluation and testing. This study represents two days of on-site investigation, the major basis of which was interviews. The committee met for a planning session, examined materials related to the testing program including test manuals and tests themselves, interviewed a broad range of teachers, counselors, school administrators, parents and other community residents, and deliberated on findings and conclusions. Specific findings under these general headings are discussed: planning the testing program; purposes of the testing program; content of the tests; application of the tests; computing, summarizing, and filing results; reporting results; and use of results. Recommendations with the intent of improving the testing program are given. (Author/DEP)

A STUDY OF TESTING PRACTICES
IN THE
ROYAL OAK (MICHIGAN) PUBLIC SCHOOLS

MAY 1975

By:

JA MILLE BOYD, MICHIGAN EDUCATION ASSOCIATION
KEN JACOBSEN, ROYAL OAK SCHOOLS, TEACHER
BERNARD H. MC KENNA, NATIONAL EDUCATION ASSOCIATION
ROBERT E. STAKE, UNIVERSITY OF ILLINOIS
JERRY YASHINSKY, ROYAL OAK SCHOOLS, PRINCIPAL

CONTENTS

The Origin of the Study, Procedures Used by the Panel, And Organization of the Report

The report that follows is a result of an agreement reached between the Royal Oak Teachers Association and the Administration of the Royal Oak School District. (The full agreement appears as Appendix A.) Subsequent to that agreement the Michigan Education Association and the National Education Association became involved in planning and conducting the study.

A panel of five who were either intimately knowledgeable of the local testing program and/or specialists in evaluation and testing was constituted to conduct the study.

They were: JaMille Boyd, Michigan Education Association
Kenneth T. Jacobson, Royal Oak Schools, Teacher
Bernard H. McKenna, National Education Association
Robert E. Stake, University of Illinois
Jerry Yashinsky, Royal Oaks Schools, Principal

The Committee met for a planning session, examined materials related to the testing program including test manuals and tests themselves, interviewed a broad range of teachers, counselors, school administrators, parents and other community residents, and deliberated for a day on findings and conclusions and (initial) drafting of the report.

The total Panel contributed to the report development.

This study was not, nor was it intended to be, a piece of scientific social science research. Rather it represents 2 days of on-site investigation, the major basis of which was interviews.

Two working papers were developed during the deliberations and are available from the authors to anyone wanting further information on the issues discussed.

Working Paper #I: "Cautionary Statements About the Working Paper, the Validity of the Tests" by B. McKenna, National Education Association.

Working Paper #II: "The Validity of the Tests" from R. Stake, 270 Education Building, University of Illinois, Urbana.

The report is divided into four parts.

In order that those examining it may get a quick and complete overview of the study, Part I presents general findings. The reader can get a broad impression from this section of what the Panel found without reading further. This section does not deal with specific tests or contain recommendations.

Part II presents a narrative related to specific areas for consideration related to costs and benefits in terms of human and other resources.* Quotations and examples in this section are only exemplary of salient points made by individuals and not meant to be generalizations.

Part III presents the Panel's findings and comments on specific tests within the Royal Oak Testing Program.

The final section contains recommendations of the Panel.

There are two appendices:

A Memorandum of Agreement for a Study of Testing Practices in the Royal Oak Schools

A Summary of the 1974-75 Testing Program for the Royal Oak Schools

*With the goals and resources available for the study, it was not possible for the Panel to gather and interpret data on financial costs of the Royal Oak Testing Programs.

An Evaluation of the Royal Oak, Michigan School District Testing Program

Part I

General Findings

A. Planning the Testing Program

1. Some parts of the testing program seemed to reflect an appropriate amount of planning by teachers and other specialists and to be based on objectives developed or adapted by them.

2. For other parts, teachers were not involved in determining whether they were to be administered, how often, and for what purposes.

3. Parents and other citizens were not well apprized as the program was developed.

4. Counselors had little or no input in planning and determining the testing program.

B. Purposes of the Testing Program

1. Some of the stated purposes for which the testing program was implemented are commendable--improving instruction, program planning and evaluation, diagnosis of individual learning problems.

2. Some stated purposes are either questionable or of less importance--public relations, establishing pupil potential, vocational guidance.

3. The stated purposes were not clearly communicated to teachers and parents. Both these groups expressed uncertainty and confusion about purposes of the total program and of specific tests.

C. Content of the Tests

1. There was a sincere effort to relate some components of the testing program to goals and objectives of instruction.

2. Some items on some tests were reported as inappropriate for the age-grade levels at which they were administered.

3. Some total tests appeared to be inappropiate for some students to whom they were administered. Some special education students were required to take tests that in teachers' professional judgment they should have been exempt from.

D. Application of the Tests

1. Some tests should have been field tested on small samples of students before broad application.

2. Teachers had little or no in-service to understand test formats and to become familiar with test administering procedures.

3. There appeared to be no provision for developing in students an understanding of the purposes of the tests.

4. The technical quality of some tests (editorial, readability, collating) was so poor as to make their administration difficult.

5. The physical facilities for testing were frequently not conducive to producing an accurate reflection of student potential and knowledge.

6. Too much time in the early part of the school year was spent on testing when teachers needed to work intensively to get educational programs operating on a sound basis and to take advantage of student freshness and enthusiasm for learning.

E. Computing, Summarizing, Filing Results

For some staff, a large amount of time was spent in scoring, graphing, and filing test results, and the potential exists for an inordinate amount of time to be spent if new testing programs are adopted for additional subject areas.

F. Reporting Results

1. The mailing of test scores to parents with little help in interpretation, in some cases, had negative effects.

2. Teachers were not provided manuals, booklets and related materials to use in interpreting tests in parent conferences.

3. Teachers were provided little or no in-service education to assist them in interpretation of results.

4. The return of results for some tests was so late that the results were of little or no use to teachers.

5. The print-outs of some tests were bulky and cumbersome to work with, and were difficult to interpret.

G. Use of Results

1. Teachers and counselors reported that a substantial part of the testing program was of little or no use in diagnosing individual learning needs or in planning for improving instruction.

2. Teachers were able to explain the results of tests to parents only in broad generalizations because they did not have the

time or the in-service opportunities, or manuals and materials for doing so.

3. Numbers of both parents and teachers reported that the results of tests only confirmed what they already knew.

Part II

Costs and Benefits:

Use of Human and Other Resources

Effects on Students, Staff, Curriculum

Teachers frequently reported that secondary students viewed the tests as irrelevant or boring and many expressed displeasure when told that they must take yet another test. Some students were frustrated because their test scores reinforced the negative opinion that others held of their ability.

Some elementary children showed overt emotional effects (e.g. nosebleeds, crying) to the tests, according to several teachers. They saw the children as particularly frustrated by their lack of understanding and success on large numbers of test items which covered subject matter to which the pupil's had not been introduced.

The concern of some parents included the possibility of an overemphasis on test-related subject matter relative to other student needs such as self-esteem and appropriate attitudes. Several parents feared the negative effects of pupil comparisons, increased competition, and the self-fulfilling prophecy phenomenon* which might be generated by the testing program. Also the large amount of time spent in testing relative to instruction was viewed as undesirable.

---

*Teacher expectations of those who did poorly in the tests would be so low that little or nothing would be done to motivate them.

Similarly, the Royal Oak counselors stated their concern that the current practice of testing every child in every grade with one or more tests, "will cause students, parents and teachers to conclude that test scores are scientific, accurate, extremely important and proof positive of each student's educational ability and achievement or lack thereof." The major concern is that the strong emphasis on testing will create unjustified faith in the validity of test scores.

The use of the test results by teachers was quite variable. Some were not used because: (1) results were returned too late in the semester; (2) accuracy of the test scores was questionable; (3) little or no diagnostic information was provided by the normative scores; (4) there was insufficient in-service regarding the meaning and interpretation of test results.

A small number of teachers indicated the usefulness of the test results, particularly the criterion referenced tests, as a guide for planning in collaboration with the reading specialists.

In general, there was little evidence that the testing program had significantly influenced the school curriculum and instruction, but there were some indications that the Objective Referenced Tests (ORTs). were being used to help teachers move toward individualized instruction.

The teachers pointed out the need for an adequate number of staff and more material suitable for instruction.

One person summarized the past year's testing experiences as involving "a tremendous amount of time, energy and concern of all parties, i.e. administrators, students, teachers and secretaries," with little or no attention and training given with respect to the purpose and use of the tests.

Both teachers and parents voiced concern that responsibility might be placed on teachers for test results.

Staff Workload, Involvement and In-Service

Teachers expressed concern about the large amount of time necessary for the administration of the tests, the management (i.e. charting, graphing, filing, etc.) and reporting of test results. The anticipated expansion of the present assessment plan in additional subject areas will increase the time demanded at all stages: administration, utilization, reporting. Thus, the teachers point out that potentially much less time will be given to creative and humanistic educational activities. Elementary teachers were particularly distressed by the amount of time spent in testing at the beginning of the school year at the expense of other desirable goals, e.g. establishing rapport with their new students.

Some parents expressed concern that the test would become the main determinant in how the teachers would spend their time. Others had expectations that time spent in testing was well invested if the data were used as the basis for accelerating or selecting and grouping pupils

according to their abilities. (The Panel points out that there is little evidence that homogeneous grouping improves learning.)

Almost all teachers voiced the need for more sufficient in-service training in all aspects of the testing program. The past year's training was highly inadequate and it was not supplemented even when requested. The unsatisfactory preparation of teachers in interpreting the test results was echoed by many parents. A PTA officer estimated that over 90% of the parents had some difficulty in understanding the test results. One parent thought that the teachers should communicate with parents in everyday rather than technical language.

Teacher Involvement

Several months before the initial use of objective referenced testing, teachers were asked to serve on Curricula Task Forces; the purpose was to write performance objectives. The task was later expanded to the writing of a test instrument. Consensus regarding the performance objectives and test items was reached by obtaining the opinions and feedback of "many, many" teachers. It was the view of one Task Force chairman that locally developed objectives and assessment tests were needed because the state test was not sufficient.

The central office administrative staff views the involvement of teachers as crucial in order that "realistic" objectives which most students can attain are set forth. However, the selection of the normative tests was

made by school administrators. Many teachers felt that these tests were very much put upon them and they had little or no usefulness for them or their students. According to one teacher-observer, the participation of many teachers on the various Task Forces was related to the accountability atmosphere. Teachers lacked the special technical understanding necessary for this task and were simply attempting to counter the state objectives and the anticipated accountability imposed upon them for their students' performance.

PART III

Issues Related to Specific Tests

Standardized testing in the Royal Oak District can be divided into three main components:

1. The tests of the Michigan Assessment Program, developed by the State Department of Education and administered in Royal Oak and all other districts in grades 4 and 7 in the subject matter of reading and arithmetic.

2. Objective Referenced Tests, developed by the teachers of the district, administered (for the first time in the Fall of 1974) to all students in grades 1-9 in the subject matter of communication skills.

3. The Comprehensive Test of Basic Skills, a nationally standardized test battery (including a scholastic aptitude scale) administered in Royal Oak district in grades 4-12 excluding grade 9, and the Differential Aptitude Test, a nationally standardized guidance test administered in the ninth grade.

After taking testimony from teachers, counselors, administrators, parents, and other citizens, and after reviewing the test materials and reports, the Panel makes the following observations on the three components:

1. The Michigan Assessment Tests

The State Assessment program was begun in the late 1960's. Recognizing that the state objectives were not the same as the District objectives and wanting to have local assessment based on local objectives, members of the local staff initiated their own project for stating objectives and later developing tests. It was assumed by some that participation in the state program would not be required if a local assessment system was operating. But participation in the state program continues to be mandated.

Originally, the state assessment program was to provide guidelines for state level policy decisions and program review. As new test procedures were developed and as new political realities became apparent, the purposes of this battery became more oriented to the aid of the teacher and curriculum supervisor in meeting the needs of the individual learner. These new purposes and procedures have not been long in effect, but the testimony of teachers indicated that some diagnostic use could be made of the state tests. However, the administration of them was time-consuming and many local objectives were not covered.

It is the Panel's belief that assessment systems should be developed locally and that participation should be limited to those classrooms, departments, and buildings for which the aims of the tests were properly representative.

The Panel is aware that the continuation and expansion of the state assessment program is being reviewed. It is clear that too many children and teachers are having to contribute time if the ultimate aim of the program is to give comparative information about districts and summary information about the state. As the program administrators know, matrix sampling is a possibility that can be considered.

If the aim is to help individual children and teachers, then each must be involved--but the benefits of the testing are not sufficiently apparent, it would seem, to warrant continuation. The state plan has been that the assessment program would be extended to other grades and other subject areas. The Panel believes that recent postponement of the expansion was wise, and that the continuation of the assessment program should be opposed by the district until a better evaluation of its effects is completed.

2. The Objective Referenced Tests

These tests have been developed to match the district's statement of goals in the communication skills area. Additional tests in other areas are anticipated if this testing activity is found to be useful.

The Panel recognizes that tests developed locally, while being more accomodating to local needs and concerns, frequently do not have the technical quality of most commercially developed tests. Test development ordinarily involves multiple reviews of items, field testing, norming,

and the development of statistical characteristics. More investment in this sort of preliminary work should be made with local tests. Unless appropriate priorities are set, the trade-off between technical competence and relevance to local objectives will continue to be a difficult problem.

It is not unreasonable to conclude, the Panel feels, that these tests have been developed and implemented too hastily or too recently for decisions to be based on their results. Until they have been worked with by persons responsible for the testing and until they have been used for a sufficient period of time to assure their usefulness, it is not reasonable to think of them as sufficiently valid for assessment purposes or for guidance purposes.

Certainly, the goal of involving large numbers of staff in particular areas to clarify instructional objectives and to develop evaluation procedures is commendable. It was not apparent that this activity was carried so far--as it has been in some districts--that the staff found the tasks onerous and completed them perfunctorily.

Some language arts teachers reported that the results of these objective referenced tests helped in diagnosis and planning for instruction. The results seem to be much less successful than what was originally promised, but that is mostly a matter of having set too high an aspiration for the project for the time and resources available.

One of the reasons for the disappointment was that there was less than a perfect agreement--as anyone would have predicted--among the staff as to the importance of the objectives selected. Some staff members reported that they were not sufficiently involved in developing and adapting the objectives. In addition, some teachers reported that there are already more stated objectives than they can teach directly to--and they believed that the procedure was in danger of becoming totally unmanageable. These teachers were not indicating an opposition to clear and explicit objectives--as would be expected, they endorsed them. But it was the large number of statements, the intricate documentation, and the inordinate amount of bookkeeping that troubled them.

A second reason for disappointment with these tests was, as mentioned above, the abbreviated period of development and try-outs. This probably resulted in the impression on the part of a number of teachers that these tests, just as the state tests, do not sufficiently reflect what it is that the teachers are actually trying to accomplish as their interpretation of, or as an addition to, what is stated in the statements of objectives.

The technical quality of these tests is not high. The administration of the tests has not been smooth. Some parts, e.g., the listening parts, are particularly difficult to administer. The school district should provide more in-service education and coordination as needed.

3. The Comprehensive Test of Basic Skills (CTBS)

This battery was recommended for implementation by the central administrative staff and approved by the board. The teachers see that it has little reference either to goals and objectives for the schools or the needs and desires of teachers and other staff in diagnosing problems and planning for instruction.

It is not uncommon for some staff members or citizens to have an interest in how a community or an individual stands with regard to national norms. Since this information had not been obtained recently in the Royal Oak district it was not surprising that the central staff perceived a need for it. Usually, the tests indicate information about the academic skills of individual children that are already well understood by teachers and parents.

This information is not related to how well teachers or the district as a whole are doing their jobs. No matter how much harder a child tries, no matter how much better a teacher is brought in to replace one departed, the results on a test battery of this kind remain about the same.

Parents often expect something else. Some who testified before the Panel here reported confusion; some said they became quite anxious. Others were pleased. But mostly, parents found the results worthy of little attention and found no reason to discuss them with school personnel.

Teachers reported that the administration of this battery was awkward and unsatisfactory. Having so many tests to take in so short a period probably fatigued both examinees and examiners perhaps sufficiently to affect the validity of the testing. A number of teachers felt that the problem of administration alone brought questions to their minds as to what the results could mean.

Students reacted negatively to these tests, the teachers said. It was reported that large numbers of students did not take the tests seriously, did not try, did not complete them--with some responding in a capricious manner. (This is not just a local problem. Examiners across the country report a lessening of concern on the part of the youngsters to make the best possible showing on tests.)

Some teachers objected to this battery on the grounds that the content was inappropriate. But the major concern with it--expressed repeatedly by teachers and counselors--was the lack of usefulness of the results. Many mentioned it specifically. Almost none found its results useful in diagnosing student learning difficulties or in planning for instruction. Most believed that whatever good results it provided were pieces of information already known to the staff.

The Panel recognizes that some parents and some staff members continue to want general skill information on individual students. There are possibilities for providing such information to those who want it.

The Differential Aptitude Test (DAT)

Since this test is applied to a limited number of students (compared to other tests examined) and its results are used by only a few staff, impressions obtained on its application and usefulness were limited.

Some counselors reported they found the DAT useful in advising students on program alternatives.

But serious question was raised about comparing DAT and CTPS scores to show school success as compared to aptitude. It was believed that the differences in development, norming, and the mechanics of the application of the two tests would make any conclusions from such comparisons very risky.

RECOMMENDATIONS

The Panel offers the following recommendations with the expectation that these findings will facilitate the further study, review and improvement of the District's instructional program.

1. We recommend that the staff continue to use student-performance information as only one means to guide and to improve the District's instructional program. Other means should include new instructional procedures, in-service education and appropriate materials and media.

2. We recommend that the staff give greater attention to the limitations of standardized testing, especially when it is being used for purposes for which the validity has not yet been determined.

3. We recommend that in evaluating student performance greater reliance be placed on the expertise and professional judgment of teachers, counselors, and other specialists, support-personnel and less reliance directly on tests and other standardized assessment instruments.

4. We recommend that the amount of classroom teaching time used for testing be reduced, except when the teachers find the testing directly contributing to instruction in ways that justify the time and effort spent.

5. We recommend that the mandatory obligations of teachers to prepare statements of objectives, formalized criteria, and assessment tests be diminished--and that teachers, administrators and other staff jointly accept such obligations only when in their professional judgment they find that such activities contribute to the maintenance of a high quality of instruction.

6. We recommend that the entire staff - teachers, superintendent, administrators, counselors - assume joint and increased responsibility to communicate effectively with the community generally and with the parents individually about student progress and educational activities of the District.

7. We recommend that the entire school staff become more aware of the ways in which assessment information is misunderstood by parents and others, and that they resist crude procedures such as mailing out student test results and offering uninterpreted school averages for publication, and that they make it as easy as possible for parents and citizens to get relevant evaluation information interpreted by someone fully qualified to do so.

8. We recommend that whenever a testing program is operating, an extensive in-service program be provided for <u>all</u> staff involved in developing, implementing, and interpreting evaluation of student progress in whatever ways it is measured.

9. We recommend an annual review of all aspects of the District's testing program involving - directly or indirectly - all who are affected by it.

10. We commend the staff for its beginning work on the Objective Referenced Tests in communications skills, but we remind them that the cautions about testing expressed in the report, apply to their tests, too.

11. We recommend that if ORTs are being considered in additional subject areas, their usefulness should be weighed against the time and effort which increase with each subject area added.

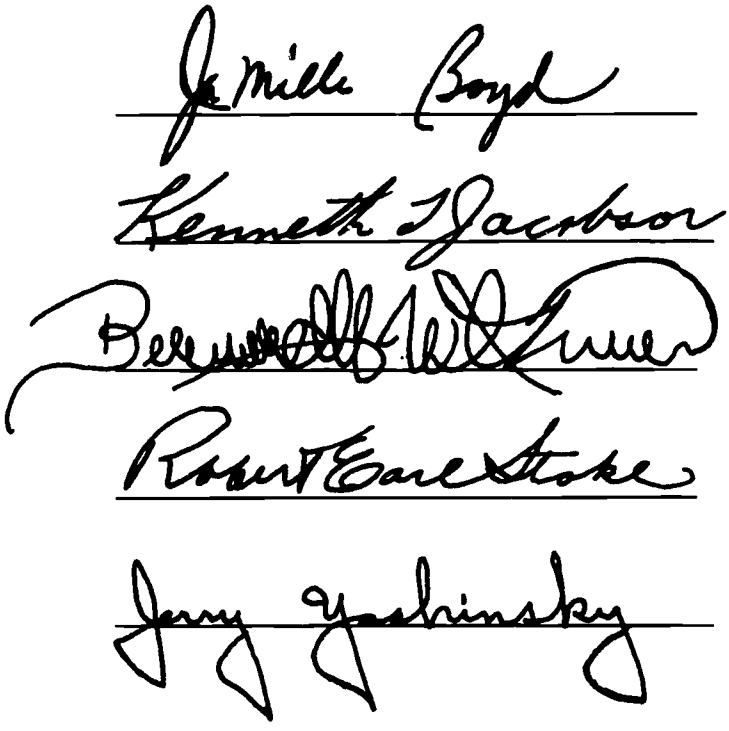
12. We recommend that the decision to use the Comprehensive Test of Basic Skills be reconsidered in terms of its total costs and actual benefits to the District.

13. We commend the staff for recognizing the disparity between State goals and District goals, and recommend that resistance to the State Assessment Program be continued as long as its costs are seen to be higher than its benefits.

14. We commend the staff for a clear understanding that achievement goals for each individual child are quite imperfectly indicated by District goal statements and recommend a continued higher priority orientation to the individual needs of each child.

15. We recommend that the total staff, and particularly teachers and counselors who have major responsibility for using the results of the testing program, be involved from the beginning in any further efforts to determine the purposes for which testing will be used, implementation, interpretation, and use of scores and will have a major voice in decisions to evaluate and revise the District testing programs.

This Report agreed to by members of the Panel, 19 May 1975.



MEMORANDUM OF AGREEMENT FOR

A STUDY OF TESTING PRACTICES IN THE ROYAL OAK PUBLIC SCHOOLS

Purpose

To study the adequacy and utility of the testing program in the Royal Oak Public Schools and to describe the effects of state/federal testing requirements on the school district.

Questions, Concerns to be Addressed:

1. Testing Program, History

   Tests Used, Chronology

   Role of teachers, administrators in selection/design of program

2. Present Testing Program

   Testing requirements (specific tests)

   Tests used

   Grades included in testing program

   Frequency of testing

   Relationship of tests to district curriculum guides, if any

   Time requirements:

      District-wide testing

      Individual diagnostic tests

   Number of tests per student/yr.

   Test results utilization:

      Administrative uses

      Instructional uses

3. Role of, and reasons for, teacher involvement in program development
4. Relationship of MEAP to district program.

Audiences for Report:

Board of Education, Royal Oak Public Schools

Instructional Staff, Royal Oak Public Schools

State Department of Education

Access to Data:

Royal Oak Public Schools will make available to the Panel such information as the Panel believes pertinent.

Study Panel/Resources for:

The study will be conducted by a panel composed of 1 NEA staff member, 1 MEA staff member, 1 ROEA member, 1 representative of the ROPS administration, and 1 outside consultant. Each organization will be responsible for the expenses of its representatives. MEA/NEA will split the costs for outside consultant.

Procedures:

The study panel will examine materials related to the development and

utilization of the testing program in the Royal Oak Public Schools and will interview such students, instructional, administrative, and community personnel as is necessary. Input from SDE staff may also be obtained.

The panel will visit the Royal Oak School District during the week of March 17, 1975. Interviews will be conducted at Emerson.

Report:

In drafting its report, the Panel will be mindful of the need to improve educational services for school children. The primary aim will be to state clearly the concerns of professional as well as administrative staff and to make such recommendations as may be necessary to improve educational programs for children in the Royal Oak Public Schools. Individuals providing testimony to the Panel will not be identified without their prior approval.

It is recommended that the final report be presented to a joint meeting of the Royal Oak Board of Education and the Executive Committee of the Royal Oak Education Association. Following the presentation above, each party to the study may disseminate the report as they see fit.

The formal report will be formulated to include the specific charge for the study, questions and concerns investigated, a discussion of procedures followed, strengths and weaknesses of the present program, conclusions, and recommendations.

SUMMARY OF 1974-75 TESTING PROGRAM FOR THE ROYAL OAK SCHOOLS

NAME OF TEST	GRADE AND FORM	DATES(S) ADMINISTERED	APPROXIMATE STUDENT TIME
State Assessment	4	1974 9-23 to 9-27	Power Test - No Specified Amount of Time Average instruction Time Used By Class Groups: 5 Hours
Comprehensive Test of Basic Skills - C.T.B.S.	4 - Form S Level 1 5&6 - Forms S Level 2	10-7 to 10-25	Total Test Time According to Test Manual: 358 Minutes Actual Instruction Time Used: 6 Hours
Short Form Test of Academic Aptitude - S.F.T.A.A.	4 - Level 2 5&6 - Level 3	10-7 to 10-25	Total TestTime According to Test Manual: 94 Minutes Actual Instruction Time Used: 40 Minutes
Royal Oak Objective Referenced Tests in Reading	Primary Yrs. (Grades 1-3) Transitional Yrs (Grades 4-6)	11-4 to 11-8 9-16 to 9-20 Readminister in May	Power Test - No Specified Amount of Time Average Instruction Time Used by Class Groups: 4 Hours
FLEMENTARY TESTING - 1974-5			

SUMMARY OF 1974-75 TESTING PROGRAM FOR THE ROYAL OAK SCHOOLS

NAME OF TEST	GRADE AND FORM	DATES(S) ADMINISTERED	APPROXIMATE STUDENT TIME
Comprehensive Test of Basic Skills - C.T.B.S.	7 - S3 8 - S3 10 - S4 11 - S4	1974 10-8 to 10-14 10-1 to 10-7 10-23, 24 11-5, 6	Total Test Time According to Test Manual: 373 Minutes Actual Instruction Time Used: Jr. High Schools - 9 - 45 min. periods Sr. High Schools - 2 Blocks of time, 3 hours 15 minutes each
Short Form Test of Academic Aptitude - S.F.T.A.A.	7 - Level 4 8 - Level 4	10-8 to 10-14 10-1 to 10-7	Total Test Time According to Test Manual: 84 Minutes Actual Instruction Time Used: One (1) 45 Minute period
State Assessment	7	9-23 to 9-30	Power Test - No Specified Amount of Time Average Instruction Time Used For Class Groups: 4 - 45 minute periods
Royal Oak Objective Referenced Tests in Reading	7 8 9	9-16, 17 9-16, 17 9-16, 17	Power Test - No Specified Amount of Time Average Instruction Time Used For Class Groups: 4 - 45 minute periods
Differential Aptitude Test - D.A.T. Royal Oak Objective Referenced Tests In Communcations Skills * Pre-Kindergarten Assessment	9	10-1 to 10-4	Total Test Time According To Test Manual: 181 Minutes Actual Instruction Time: 3 hours, 15 minutes

SECONDARY TESTING - 1974-75

About Working Papers I and II

Material in these attachments was prepared during the period that the panel to evaluate the Royal Oak Testing program was at work. Unlike the major sections of this report, the panel did not agree that they should be included either as an integral part of the findings and recommendations or as appendices to them. Therefore, they were prepared as separate working papers.

Working Paper I is a rebuttal to Working Paper II. It has been placed first because the impressions given by first examining the Stake Table (Working Paper II) and accompanying narrative would cause hasty and unwarranted conclusions. This is because, whether intended or not, the Stake Table communicates the assumption that all the "Information Purposes for Which Tests are Sometimes Used" in the right hand column are desirable purposes.

This impression prevails even though Stake says "... the social consequences of these uses were in no way considered as part of the check on validity" and that "we can see from the charts that the tests have been shown to be valid for what was once* their principal jobs... For almost all other purposes of testing, these tests have not been validated statistically," and even though he obliquely indicates that "what was once their principal jobs" may no longer obtain.

If one draws the logical conclusion that (1) what was once their principal jobs (which is shown in Working Paper I to be inappropriate purposes) and

---

*Emphasis mine.

that (2) for all other purposes, as Stake says, they have no validity, then why the Table at all? Why develop a complicated paradigm to show that the tests are useful for unworthy or obsolete purposes?

Since Stake doesn't say much about these purposes ("for what was once their principal job"), they are dealt with in Working Paper I, as well as some of the other purposes listed by him.

In summary, a further illustration using Stake's analogy about handguns may be helpful. The Stake Table, to a considerable degree, seems to suggest something comparable to arguing that certain weapons are excellent for some particularly destructive purposes and highly inaccurate for others. So why use hand guns at all? The reader will want to move back and forth from the Stake Table to the "Cautionary Statements..." of McKenna to fully understand the subtleties suggested here.

Bernard H. McKenna

Working Paper I

CAUTIONARY STATEMENTS ABOUT THE WORKING PAPER ENTITLED "VALIDITY OF THE TESTS"

Bernard H. McKenna
Professional Associate
National Education Association

CAUTIONARY STATEMENTS ABOUT THE WORKING PAPER ENTITLED

"VALIDITY OF THE TESTS"

Bernard H. McKenna

Robert Stake's paper, "Validity of the Tests," implies that several highly questionable testing purposes may be legitimate. Even though it may be possible to show validity of some of the tests for these purposes, it is the purposes themselves that are faulted. It is proper caution about the dangers in the purposes themselves that Stake fails to give sufficient attention. These questionable purposes are:

1. To indicate standing of invididual students with reference to norm groups.
2. To predict future standing of the individual in other reference groups.
4. To measure gain or improvement in skill or knowledge since a previous measurement.
7. To indicate the standing of the entire group.
8. To measure gain or improvement for the group.
10. To evaluate teaching.

Following are deficiencies in six (6) of the ten (10) purposes in Table I that make whatever validity the various tests show for each of those purposes highly questionable and even potentially dangerous.

Purpose I: <u>To Indicate Relative Standing of Individual Students, Possibly With Reference to Norm Groups</u>

It has recently become recognized by numbers of evaluation experts that the dangers that result from comparing the standing of individuals to norms (averages) far outweigh the usefulness of such practice.

Firstly, there is no way of knowing what the averages themselves mean in terms of quality of performance. That is, an average is only a mathematical statistic. Scores below the average might represent satisfactory performmance. Or conversely, scores at the top might represent unacceptable performance. Until some value judgment is placed on average, high and low, test scores have little meaning and their use can easily misrepresent quality of performance. Beyond that, lining students up on a scale in reference to an average assures that half the students will fall below the average, no matter how well individual students may be doing.

Secondly, norms (averages) frequently tend to be used in combination with ranking procedures, which in turn are used to determine pass or fail status of students. The tests have another deficiency which makes this practice dangerous. This deficiency is referred to as the standard error of measurement. As an example of how this works, a student's score of 82, with the measurement error considered, might conceivably fall anywhere between a 77 and 87. But taken at face value, decisions are made about students using a particular score as a cutting point without considering this wide margin of uncertainty.

Purpose II: To Predict Future Standing of the Individual in an Anticipated or Hypothetical Reference Group

A number of factors more significant than test performance, some difficult

to measure, contribute to potential "Future Standing (aptitude)." Using tests to predict future standing is frequently accompanied by sorting and classifying students and grouping them for instruction (homogenous grouping). Students are denied access to some educational programs on the basis of such limited and questionable predictions. Instances are known of students being placed in classes for the mentally retarded on the basis of this testing purpose. A court ruling in the District of Columbia several years ago which abolished tracking (homogenous grouping) of students on the basis of tests was a landmark decision linking standardized testing to denial of equal educational opportunity. In California, a judge recently found that standardized testing causes a disproportionately high percentage of minority students (compared to the general population) to be placed in classes for the educable mentally retarded.

The Task Force on Testing of the National Education Association recommended this year that:

> "Tests must not be used in any way to label and classify students, to track students into homogenous groups, to form the major determinants of educational programs, to perpetuate an elitism, or to maintain some groups and individuals 'in their place' near the bottom of the socio-economic ladder. In short, tests must not be used in any ways that will deny any student full access to equal educational opportunity."

Finally, if the prediction of future standing has to do with how well one will stand on similar tests at other educational levels (for secondary schools

if the student is elementary, for college if the student is secondary) it is suggested that tests and measurements specialists stop developing tests and determining their validity for such purposes. Rather, the testing and measurement community should develop tough-minded alternatives to standardized testing.

Purpose IV: To Measure Gain or Improvement in Skill or Knowledge Since a Previous Measurement

The assumption that tests can accurately measure learning growth over arbitrary time periods is now considered by many authorities to be a misconception.[1] The idea of "a year's growth in a year" is seen as an artificial way of packaging learning activities. No sound justification has been built for 18 weeks (a typical semester) or 10 months (a typical school year) as compared to 11 weeks or 22 weeks or 9 or 11 months.

In addition, different students mature at such different rates and gain learning readiness for different kinds of learning at such widely varying ages that "normal growth" for a variety of educational purposes should be considered only in terms of the individual.

Purpose VII: To Indicate the Relative Standing of the Entire Group Possibly With Reference to a Norm Group

This purpose has most of the same deficiencies as Purpose I.

---

[1] Robert S. Soar and Ruth M. Soar, "Problems in Using Pupil Outcomes for Teacher Evaluation," Washington, D.C., National Education Association, 1975, mimeographed (unpublished).

Purpose VIII: To Measure Gain or Improvement for the Group Since a Previous Measurement

This purpose has most of the same deficiencies as Purpose #4. In addition, those students in a group who initially have less knowledge and skills will appear to make more gains. Conversely, students who start out with more knowledge and skills will appear to make less gain. Group gain scores will be affected, then, by the percentage of the individuals within a group that start with various levels of knowledge of the subject -- more gain appearing to take place if larger percentages of students begin with little knowledge of the subject.[2]

Purpose X: To Evaluate Teachers (or Administrators) as to Competency or Effectiveness of Instruction

This is a totally unacceptable purpose for standardized testing for at least three reasons:

1. The tests are too crude to directly reflect teaching or administrative effects. Even though there is much agreement that teachers, administrators and schools generally contribute much to providing appropriate climates for learning, neither they nor anyone else know what specific instructional techniques directly contribute to gains in learning, as learning is currently measured.

2. Even if specific instructional techniques that produce learning

---

2 Ibid.

were known, differences in experience levels, verbal, and other skills in every classroom of students of school building would have to be exactly equal. Otherwise, teachers and administrators who received students with broader experiences and higher skill levels would appear to be more competent and effective.

3. All the factors other than what teachers and administrators do (and over which they often have little or no control) that cause or inhibit learning would need to be controlled, currently an impossible task, e.g., emotional climate of the home, nutrition, motivation, up-to-dateness of curriculum, availability of media and materials, class size, uninterrupted teaching time, specialist availability, in-service opportunities, plant facilities. There are others.

General Test Deficiencies that Cause the Tests Not to Achieve the Purposes

An over-arching reason which affects all six unacceptable purposes is that the tests frequently don't measure what they are claimed to measure. It has been found in many instances, and with a wide variety of standardized tests, that individual test items are often ambiguous, misleading, and confusing. It has been shown that students who think the most deeply and creatively mark wrong answers because of such test deficiencies. Thus test results produce misleading information rather than what the users want to obtain.[3]

---

3 Jerrold R. Zacharias, "The Trouble With I.Q. Tests," The National Elementary School Principal, Vol. 54, No. 4, March/April, 1975, pp. 23-29.

A second general deficiency of the tests which affects the purposes is that they are most often standardized on student populations that are typically middle class, majority group, and for whom English is the native language. They don't work well for minority students, for those of low socio-economic status, and for those for whom English is a second language.

A Word About Objective Referenced Tests

Objective referenced tests are said to correct many of the deficiencies of standardized (norm referenced) tests. In practice, many of them have not worked that way:

1. They often do not reflect local school district or building or individual classroom objectives any better than conventional standardized tests.

2. The items in these tests are frequently the same items used in standardized tests and contain the same ambiguities cited above for standardized tests.

3. They are quickly and frequently converted for norm referenced measurement or have applied to them pass-fail scores, cutting points, or minimal competency levels.

Working Paper II

"THE VALIDITY OF THE TESTS"

Robert E. Stake
Specialist in Testing and Measurement
University of Illinois

THE VALIDITY OF THE TESTS*

One of the issues of concern to test makers and test users for many years has been the validity of tests. People experienced with educational tests realize that any one test has a different validity for different purposes. A test may be highly valid for some purposes, but for other purposes that same test may have low validity.

A test with high validity is one which obtains--with a high degree of accuracy--the very information the user wants to obtain. A user does not, of course, want just a test score; he wants a test score that indicates something. A test is used at different times to indicate different things. The validity of the test each time depends on what the user wants indicated.

Educational tests do not measure directly the skills or understandings of a child, nor the effectiveness of a curriculum. They are used to indicate these things by measuring what children answer to a small selection of questions. Only a small sample of the many relevant questions is asked.

And even the total sum of all possible questions does not directly indicate what it is that the user wants measured. Educational tests are always indirect measuring instruments. These tests will have low validity if they are inaccurate--but they also will have low validity if they are measuring something that is not a good indicator of what the user wants measured.

For some uses the validities of even the best tests have never been demonstrated. Some tests have been used millions of times without a check on a validity of the

---

*A statement prepared for the Panel on Testing in the Royal Oak (Michigan) School District, May 1975.

most expected usage. For example, the validity of "reading readiness" tests as a guide to beginning or postponing formal reading instruction has not been established. The diagnostic uses of most tests are not based on "demonstrated validity." In other words, the technical studies to show that instruction is more effective when based on the test information have not been done. For many other tests and testing purposes the validity of the test is only assumed. Test developers and researchers have not yet demonstrated their validity.

During the fifty years or so that we have had these tests the users have been interested in only a few possible uses of them. Recently, particularly with the arrival of the "accountability movement," many additional uses of testing have been proposed. It has been implied that tests that have been shown valid for discriminating among students would naturally be valid for assessing the effectiveness of teachers, verifying the quality of textbooks, determining the accountability of a district, deciding on a district's need for specially-trained remedial reading teachers, and for setting national educational policy. It is possible that the tests will be useful for these purposes--but at this time the claims for such testing have not been backed up with validation studies.

The purpose of these statements is not to argue that we should do such validity studies, but that we should not assume that they have been done. The purpose is to urge users of tests to resist the temptation to suppose that the tests will obtain complex information for us that has not yet been obtained elsewhere.

The validity of a test for a particular use is demonstrated by showing (usually in a carefully supervised statistical study) that improved understandings or decisions are reached by using the test. When test scores are used in combination with other observations to reach understandings or decisions, the validity

of the test would be shown by the increase of effectiveness as a result of adding in the test information.

It is not unreasonable for educators to use tests for which the validity has not yet been demonstrated. They of course should use them with greater caution.

A test may be useful to teachers or administrators even when it has not been validated statistically. We sometimes speak of a "clinical" or "experiential" validity. Most test specialists are critical of such nonstatistical bases for decision making, at least if statistical validation is a practical alternative. I know a few professional educators sufficiently knowledgeable about the curriculum and about the tests that they can use the test scores to improve instruction either in the classroom level or for the district as a whole. Our studies show that this is not true of most teachers and administrators. And we do not have a good way of knowing "for which users, in which situations" a test can be said to have a clinical validity.

Most people who are not well acquainted with testing have too high an opinion of the validity of the tests. The testing literature is filled with cautionary statements. They warn of expecting too much from the tests. But many persons, including experienced educational officials, let their yearnings to have instruction fully measured obscure these cautions.

In an effort to summarize estimates of the confidence we might place in tests for obtaining different information I have prepared the following table. The statements of validity for the different tests are based on my experience, reasoning, and reading of the professional literature. I have submitted these esti-

mates to several colleagues who have indicated that--with perhaps a slight difference of opinion in two or three cells--they agree with my estimates.

Some other colleagues have given additional advice. They have said that I should not circulate this chart because it will be misused, used to justify the abusive uses of tests.

Just as an accurate handgun can be used for immoral purposes, so also can a valid test be used for immoral purposes. Either can be hurtful through negligence. An array of test scores can be used to deny equal opportunity, to grant undeserved privileges, or to disguise bigotry.

Ranking students, assigning them to fast or slow groups, or treating them differently in school on the basis of predicted future success are potentially immoral ways of handling students. The educational benefits for these common practices are more apparent than real, and the social costs are potentially high.

For the first, second, and seventh purpose listed on the chart several test types have been demonstrated by psychometrists to be valid. But the social consequences of these uses were in no way considered as part of the check on validity. Each educator and each citizen (as well as each psychometrist) should be questioning the morality of these discriminations.

What we can see from the chart is that the tests have been shown to be valid for what was once their principal jobs: e.g., to indicate the relative standing of youngsters, to grade them, to admit them to special programs, and to predict the level of performance at a later time. For almost all other purposes of testing,

these tests have not been validated statistically. Some of the tests are too new to have gained a demonstrated validity. For some purposes the uses are too diffuse or idiosyncratic to justify the investigation. But for whatever the reason, the validity for most assessment purposes has not been demonstrated.

It surely is as much a mistake to expect too much from tests as it is to fail to accept what help they can be. Many professional persons can benefit from the stimulation tests give to thinking about how to improve the curriculum. Many can use tests to orient students to their work and get them to work harder. And sometimes educators can actually use them to measure what they want to measure. Use of tests by professional persons with a full realization of the ill effects of discrimination, working to improve the opportunities for learning, should be encouraged.

In many places there is a call for using tests to indicate the accountability of the teacher or the school system. For this use no type of test has a demonstrated validity. Such use of tests seems clearly unwarranted at this time.

Robert E. Stake
Specialist in Testing
& Measurement
University of Illinois
at Urbana-Champaign

INFORMATION PURPOSES[2] FOR WHICH TESTS ARE SOMETIMES USED	KIND OF TESTS → Scholastic Aptitude Tests	General Learning-Skills Tests	Specific Competency or Skills Tests	Curriculum-specific Tests of Subject Matter	Broad Co Objectiv Tests
How used in Roval Oak →	SFTAA	CTBS	ORT and Mich Assessment	Teacher-made tests	none
USING SCORES TO REPRESENT THE INDIVIDUAL STUDENT					
...to indicate relative standing of individual students, possibly with reference to norm groups[3]	HIGH VALIDITY	HIGH VALIDITY	VALIDITY NOT KNOWN	VALIDITY VARIABLE AND SELDOM KNOWN	HIG VALI
...to predict future standing of the individual in an anticipated or hypothetical reference group[3]	HIGH VALIDITY[4]	HIGH VALIDITY[4]	VALIDITY NOT KNOWN	VALIDITY VARIABLE AND SELDOM KNOWN	HIG VALI
...to describe the individual student ipsatively with a profile or array of subscores	No SUBSCORES	LOW VALIDITY	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY VARIABLE AND SELDOM KNOWN	LO VALID
...to measure gain or improvement in skill or knowledge since a previous measurement[5]	NOT APPLICABLE	LOW TO MIDDLING VALIDITY	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY NOT KNOWN	LOW MIDD VALI
...to measure attainment directly, using items or tasks of particular interest in and of themselves[6]	NOT APPLICABLE	NOT APPLICABLE	VALIDITY NOT KNOWN BUT PROBABLY HIGH	VALIDITY VARIABLE AND SELDOM KNOWN	NO APPLI
...as a diagnostic device, to identify a needed lesson or instructional tactic for the individual student	NOT APPLICABLE	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY NOT KNOWN	VALID NOT K BUT PROBA LOW
USING MEANS TO REPRESENT AN ENTIRE GROUP					
...to indicate the relative standing of the entire group possibly with reference to a norm group[3]	HIGH VALIDITY	HIGH VALIDITY	VALIDITY NOT KNOWN	VALIDITY VARIABLE AND SELDOM KNOWN	HIGH VALID
...to measure gain or improvement for the group since a previous measurement[5]	NOT APPLICABLE	MIDDLING TO HIGH VALIDITY	VALIDITY NOT KNOWN BUT PROBABLY MIDDLING	VALIDITY NOT KNOWN	MIDDL TO H VALID
...to identify a suitable lesson or instructional tactic or curriculum for the entire group	NOT APPLICABLE	NOT APPLICABLE (AND NO VALIDITY)	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY VARIABLE AND SELDOM KNOWN	PROBA NO VALID
...to evaluate teachers (or administrators) as to competency or effectiveness of instruction	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALIDITY NOT KNOWN BUT PROBABLY LOW	VALID NOT KN BUT PROBAB LOW

Notes:
1 These estimates are for usual situations and would not cover special cases.
2 Tests are used for important purposes other than for getting information.
3 The morality and educational value of gathering this info is in question.
4 High validity here is dependent on an "academic" criterion & a varied group.
5 This is not the same purpose as the last below, "to evaluate teachers or curric."
6 A high validity here would not necessarily mean that these scores represent effectively a broader area of achievement.

APPENDIX C

ROYAL OAK TESTING STUDY

PERSONS INTERVIEWED

Category	Number
Teachers	26
Parents	26
Administrators	8
Board Member	1
Counselors	4
School Psychologist	1
Reading Specialist	1