DOCUMENT RESUME

ED 135 842 95 TM 006 069

AUTHOR Rasp, Alfred Jr.
TITLE Using Anchor Test Study Tables in State Assessment Programs.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-58
PUB DATE Dec 76
CONTRACT 400-75-0015
NOTE 8p.; For the Anchor Test Study, see ED 092 601-634

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Educational Assessment; Elementary Education; *Equated Scores; Grade 4; Grade5; Grade 6; *Norms; Raw Scores; Reading Achievement; *Reading Tests; Standardized Tests; *State Programs; Test Interpretation
IDENTIFIERS *Anchor Test Study

ABSTRACT

This paper focuses on three topics. The first introduces the original Anchor Test Study conducted and reported by Educational Testing Service (ETS) from 1971 to 1974. This study, involving the testing of more than 300,000 children, produced raw score equivalency tables for eight commonly used reading tests and new individual and school-mean norms tables for grades 4, 5, and 6. The second part describes Washington State's 1973-74 use of the Anchor Test Study tables to conduct a reading assessment based on a statewide sample of sixth-grade students and 1974-75 efforts to develop computer programs to facilitate greater practical application of the original tables. The final section describes advantages shown by the Washington experience and presents suggestions aimed at maximizing the potential of the anchor approach to a state-level assessment of reading achievement. (Author/RC)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 58 DECEMBER 1976

USING ANCHOR TEST STUOY TABLES IN STATE ASSESSMENT PROGRAMS

Alfred Rasp Jr.

ABSTRACT

This paper focuses on three topics. The first introduces the reader to the original Anchor Test Study conducted and reported by Educational Testing Service (ETS) from 1971 to 1974. This monumental study, involving tne testing of more than 350,000 children, produced raw score equivalency tables for eight commonly used reading tests and new individual- and school-mean norms tables for grades 4, 5, and 6.

The second part describes Washington State's 1973-74 use of the Anchor Test Study tables to conduct a reading assessment based on a statewide sample of sixth-grade students and 1974-75 efforts to develop computer programs to facilitate greater practical application of the original tables.

The final section describes advantages and disadvantages shown us by the Washington experience and presents suggestions aimed at maximizing the potential of the anchor approach to a state-level assessment of reading achievement.

HE ANCHOR TEST STUDY

The powerful notion of accountability in education is not the direct focus of this paper, but it serves logically as the starting point in a discussion of the development of the Anchor Test Study and the use of its results. Talk about educational accountability has been widespread for several years. The most cursory survey of the literature or the briefest of visits to a school's faculty room or to a local school board meeting or to a legislative budget hearing will confirm the continuing popularity of the concept. And although not everyone using the term can agree on its meaning or what is required to achieve it, two aspects are commonly acknowledged. The first is the general concern for accomplishment. While it may be true that in the past educators concentrated their efforts on measuring and accounting for inputs rather than results in terms of student performance, today it is clear that both public and professional expectations extend well beyond accounting for inputs to an abiding interest in the achievement of students.

The second commonly held idea grows from this concern for results: More and more groups of private citizens and elective bodies are mandating formal and public reporting of the relative effectiveness of various local, state, and federal educational programs.

This general demand for accountability and the special interest in improving achievement and demonstrating program effectiveness has led to the Anchor Test Study. The specific motivating force was the desire to evaluate the success of the Elementary and Secondary Education Act (ESEA) Title I program. The disappointing results of a 1968 evaluation attempt demonstrated vividly to the U.S. Office of Education the basic problems inherent in attempting to aggregate reading achievement data gained from a wide variety of tests lacking statistical comparability. In 1969, the feasibility of equating achievement tests in reading was investigated, and in 1971, a contract was awarded to Educational Testing Service (ETS) to carry out a study using one test as an anchor point for equating and norming other commonly used reading achievement tests. In April of 1972 and 1973, data were collected on the eight tests that ultimately formed the basis of the widely known Anchor Test Study (ATS), published in final form as a technical report consisting of 34 volumes and more than 15,000 pages (1).

In developing the anchor tables, ETS carried out two operations: norming and equating. The norming phase was accomplished by administering the reading subtests of the Metropolitan Achievement Test to a total of more than 200,000 children in grades 4, 5, and 6. In the equating operation, about 150,000 children took pairs of the selected

reading tests. A total of more than 1,700 schools and 350,000 students participated in the study.

The resulting norms tables developed by ETS provide transformations of the raw scores of the eight reading achievement tests to a single table of national percentile ranks and provide national, individual, and school mean norms for grades 4, 5, and 6. A listing of the tests, editions, forms, and levels included in the study is presented in Table 1.

Suggested Uses

The equivalency tables and the individual student and school norms tables provide a versatile array of applications in assessment and evaluation. A concise discussion of alternative uses is found in the "Use of Tables" section of the popular ATS report prepared by Loret, Seder, Bianchini and Vale (2, pp. 3-6). Two examples taken from that discussion indicate the wide range of practical applications.

First, a comparison of individual student performances using scores from different tests:

> Problem—it is desirable to compare the reading achievement of three students: Peter, Alan and Chuck (all 5th graders) have each taken a different reading test. Their Total Reading raw scores are: Peter (49 on CTBS), Alan (44 on GMT), and Chuck (54 on MAT). To compare the Anchor Test Study national percentile ranks for these three pupils, turn to table 26, page 73, to find the norms for Total Reading score, grade 5. Find each pupil's score in the "Raw score" column, then read across until you find the appropriate entry under that test's name. Peter's 49 yields an Anchor Test Study national percentile rank of 32 on the CTBS, Alan's 44 yields 18 on GMT, and Chuck's 54 yields an Anchor Test Study national percentile rank of 50, on MAT. These Anchor Test Study national percentile ranks are now directly comparable because they are derived from the same norms sample.

Second, a comparison of the performance of two or more schools with mean scores based on different tests:

> Problem—to compare the vocabulary performance of 6th grade pupils at Classical Elementary (mean score on SAT, 29) and Lowell Elementary Schools (mean score on CAT, 26): While the raw score school means are available for both schools, they are based on two different tests. Table 31, page 87, contains the Anchor Test Study school mean norms for grade 6, Vocabulary. Locate the mean raw score (29) for Classical Elementary School and find the corresponding Anchor Test Study percentile rank and stanine in the column entitled "SAT" (percentile rank of 72,

TABLE 1

Test Edition	Form	Level Used at Grade:		
		4	5	6
California Achievement Tests (1970 ed.)	A	3	3	4
Comprehensive Tests of Basic Skills (1968 ed.)	Q	2	2	3
Gates-MacGinitie Reading Tests (1964 ed.)	1M	Survey D	Survey D	Survey D
Iowa Tests of Basic Skills (1971 ed.)	5	10	11	12
Metropolitan Achievement Tests (1970 ed.)	F	Elementary	Intermediate	Intermediate
Sequential Tests of Educational Progress, STEP Series II (1969 ed.)	A	4	4	4
SRA Achievement Series (1971 ed.)	E	Blue edition	Blue edition	Green edition
Stanford Achievement Tests (1964 ed.)	W	Intermediate I	Intermediate II	Intermediate II

stanine 6). Now enter the same table, by locating the mean raw score (26) for Lowell Elementary School, and read the Anchor Test Study percentile rank and stanine in the column entitled "CAT" (percentile rank of 89, stanine 7). These Anchor Test Study ranks may now be compared.

A STATEWIDE APPLICATION

If it is possible to compare the achievement of individual students or schools using the ATS norms tables, would it not then be possible to use the same procedures on a larger scale to develop an assessment of reading for an entire state? This was essentially the question asked by the Program Evaluation Section in the office of the Washington State Superintendent of Public Instruction during the summer of 1973 when the unofficial results of the ETS efforts were first being discussed. The Washington deliberations led to a positive course of action, and the desire to develop a state reading profile through the application of the ATS tables was incorporated into the State ESEA Title III needs assessment plan for fiscal year 1974.

Support for this style of assessment rested on an interest both in generating a description of the reading performance of Washington pupils and in studying the feasibility of using the ATS norms tables and local school district data as the basis for constructing a state profile of reading achievement.

When the *Anchor Test Study Users Manual* (unofficial version not including the Gates-MacGinitie) was made available to the Washington Superintendent of Public Instruction in the fall of 1973, the plan was set in motion. In an effort to generalize reading achievement to the state as a whole and to categories arranged by size reflecting school district enrollment, each common school containing grades 4, 5, and 6 was assigned to one of 10 categories. Twenty percent of the schools were drawn randomly from each size category with an additional 10 percent sample of schools drawn as alternates. A questionnaire was prepared and sent to all school districts to collect information related to the use of the tests included in the ATS tables. Because the survey showed that more of the ATS tests were administered at the sixth grade than at the fourth or the fifth, grade 6 was selected for analysis, and the sampled schools were checked to see where replacements would be required.

Requests for the raw scores of sixth-grade students were sent to the selected schools, and as the resulting data were tabulated, four circumstances became apparent: 1) Several districts did not complete the Anchor Test Profile Survey accurately and did not possess information as claimed. 2) The test results were submitted in a greater variety of forms than was anticipated, especially in the way scores were reported, for example, percentiles, stanines, grade equivalencies, and growth scores were received in addition to the raw scores requested. 3) The times of test administrations covered every month from September to June. 4) Although 87 schools and 6,568 students were included, insufficient appropriate data were available to maintain a 20 percent random sample in each of the 10 size categories as a basis for generalization. The problems of data analysis were greatly increased because of the effort to maintain some semblance of a random sample, and in many instances, precision suffered as a consequence of dealing with the lack of compatibility in test forms, levels, editions and time of test administration.

Although the Washington study resulted in a somewhat limited description of reading performance, it did produce a profile of reading achievement and identified a number of procedural problems which could be remedied in future assessment programs. The results of the reading assessment are displayed in Table 2 which shows an analysis of sixth-grade reading scores using school norms. A more complete discussion of the Washington experience is presented in a technical report titled: *Washington Statewide Assessment Using Anchor Test Norms* (4).

The Development of Computer Programs

The outcome of the 1973-74 study was positive enough to encourage the Washington evaluation staff to consider further use of the ATS tables on the state level. In 1975, we developed computer programs to facilitate the use of the ATS tables for both state and local assessment purposes. The Northwest Regional Educational Laboratory assisted the state office in writing programs to provide score transformations among the eight tests and conversions between fall, winter, and spring norms. The resulting programs accomplished three key purposes. First, the ATS equivalency tables were programmed into the computer so that test scores could be equated quickly. However, since the original ATS tables reported only raw scores and spring norms, they were of limited use for large-scale assessments based on existing data. To provide greater flexibility, two additional steps were taken. Tables were developed and programmed to convert fall and winter testing times to spring norms. The testing time conversions assumed linear growth; for example, if a student was achieving at the 46th percentile in the fall, a straight line projection (with score increases spaced equally between intervals) was made to a spring percentile of 46. (This assumption introduces the possibility of error but is commonly used in large-scale assessments and program evaluations.) Tables were also programmed to convert the standard reporting options—for example, grade equivalent scores, percentiles, and scale scores—to raw scores.

The practical utility of the original ATS accomplishments is enhanced by the additional programs. The following is taken from the Washington *User's Guide to the Anchor Test Program* (3 p. 3) to illustrate their usefulness:

For example, School A may report grade equivalent

scores from Fall testing with the California Achievement Tests, while School B reports raw scores for the same time and test. School C may use Spring percentiles from the Iowa Tests of Basic Skills, while School D has Spring raw scores from the Stanford Achievement Tests. By using the Anchor Test Program, these schools can now communicate meaningfully with each other about these test scores.

Efforts are now under way to make the Anchor Test Program available to those Washington school districts and other agencies of the common school district that have computer installations.

TABLE 2
Washington Assessment
Grade Six Total Reading Scores

Estimated State and Size Category Means and Standard Deviations for Six Standardized Tests (School Norms)

District Size	Standardized Reading Tests					
	CTBS	ITBS	MAT	SAT	SRA	STEP II
20,000 and over	46.4 6.6	61.6 8.9	63.5 7.6	62.3 8.8	53.6 7.9	42.1 4.5
10,000—19,999	43.1 7.8	57.0 10.6	60.0 9.2	57.8 10.6	49.7 9.6	40.1 5.6
5,000— 9,999	49.6 4.9	65.8 6.8	67.3 5.2	66.6 6.5	57.4 5.7	44.4 3.0
3,000— 4,999	48.9 4.9	64.7 7.2	66.0 5.3	65.0 6.6	55.6 5.5	43.4 2.7
2,000— 2,999	52.9 7.1	70.7 10.2	70.5 6.9	70.9 9.4	61.0 8.1	47.2 4.0
1,000— 1,999	46.6 4.6	61.9 6.4	64.2 5.3	62.6 6.2	54.5 5.6	42.5 3.1
700— 999	43.6 8.2	58.3 11.7	60.9 9.6	59.4 12.0	50.5 10.2	40.6 5.7
500— 699	46.4 3.3	61.4 4.5	64.0 3.6	62.2 4.3	53.4 4.4	43.1 0.9
300— 499	41.9 17.7	55.8 22.9	57.4 23.1	50.2 18.5	46.2 23.7	38.3 14.2
Under 300	50.2 8.4	67.1 11.4	67.5 9.8	66.5 10.7	58.4 9.9	44.9 6.8
State (All Schools)	47.0 7.4	62.4 10.1	64.1 8.6	62.4 9.4	55.0 9.1	42.6 5.1
National ATS Median Scores	46.8	62.0	64.8	63.0	54.2	43.0

Note—The first number represents the mean. Second number represents the standard deviation. Although scores on CAT were not reported, CAT state and stratum means can be estimated from the data using Educational Testing Service equivalency tables. CAT means for large districts to the state respectively are approximated as follows: 44, 40.5, 46.5, 45.5, 50, 44, 41, 44, 39, 47, and 44.

ADVANTAGES AND DISADVANTAGES HIGHLIGHTED BY THE WASHINGTON EXPERIENCE

The Washington experience has shown us the advantages and disadvantages of using the ATS tables to conduct a statewide reading assessment. Some of the problems faced in Washington State are peculiar to that setting, but others generalize to a broader range of situations. For example, unless a state requires that local districts use tests in the anchor study, you can anticipate a sampling problem. It is highly improbable that the distribution, across known relevant variables, of districts or schools using compatible anchor tests would be wide enough to ensure that a random draw would select only units with the desired test information. Sampling was a major problem in Washington. Even with an initial 20 percent sample in each size category and an additional 10 percent replacement sample, the schools in the final sample ranged from a low of 6.5 percent in one category to a high of 11.8 percent in another. (See Table 3.) This loss of original sample units limits, to an unknown degree, the ability to generalize from the state results. The state profile of reading is overly influenced by those size categories with higher percentages unless the results are weighted to more accurately reflect the populations involved. Certainly in the district size categories where the number and percent of sampled schools is small, the stability of the achievement estimates must be seriously questioned. The ability to generalize to the entire population with confidence is directly affected by the degree to which the sample lacks precision.

Obtaining an accurate description of available test data at the local district or school level presents another problem. Easy use of the ATS tables depends not only on the use of an anchor test but on the use of the appropriate form and level as well. In addition, an accurate record of administration times and the available test-results reporting options—raw scores, percentiles, stanines, and so on—is crucial planning information. The logistics of data collection also pose problems. Not that districts fail to cooperate, but that test data are frequently supplied in many "shapes and sizes" and the clerical sorting task is monumental. The computer programs developed by the Washington State office, however, help to solve many of the processing and analysis problems stemming from the wide array of test results generated at testing times other than spring, and reported in options other than raw scores.

There are other limitations to the use of the ATS in statewide assessments. The tables limit the assessment to the reading areas, total scores and subtests, and to three grade levels. In addition, since the test items are already selected and organized into standardized tests, there is no opportunity to add or subtract items or the objectives they measure. The achievement assessment is limited to what the eight tests cover, and the items in these tests have been used because they discriminate in a norm-referenced way, not because of their relevance to program objectives.

A final limitation stems from the original parameters of the Anchor Test Study itself. Eight test editions served as the basis of the effort. Two of the tests, the CTBS and Stanford, have already been revised, with new editions planned for several others in the near future. Unless the current tables are expanded or the test publishers themselves provide precise bridges between editions (a rather unlikely event), the current tables will soon be outdated and their usefulness limited.

Efficiency a Major Advantage

In the face of these limitations, there is still a very potent advantage inherent in the anchor test approach to the state-level assessment of reading, and this is the efficiency and low cost of this style of testing. The anchor tests were selected for inclusion in the equating and norming procedure because they are widely used achievement tests. It is probable that in any state most schools administering standardized achievement tests make use of one of the popular anchor tests as part of the regular testing program. To the extent that this is true, no new testing is required.

Local sampled schools need only send copies of scores to the state office for tabulation. This means that the state assessment program can build primarily on existing local test data and that no specific test need be mandated by the state agency or legislature. The resulting assessment program presents a low profile, is unobtrusive, and requires only a limited amount of staff time and relatively few dollars. This basic advantage, while not responding to all of the limitations, is extremely powerful in a time when educational resources are becoming scarce and the demand for public accounting widespread and influential.

Suggestions for State Level Assessment

If the purpose of a state reading assessment is to produce statements comparing the state-level performance or achievement of students to national norms and/or to make broad comparisons among selected educational groupings within the state, the low cost and efficiency gained by using the ATS tables are worth careful consideration. The following suggestions point out some of the major steps that can be taken to implement a reading assessment based on the ATS tables that interferes only minimally in the affairs of local schools and requires only limited resources. To avoid peak load problems in staff time, approximately 18 months should be allowed for the process, with the starting point in late winter or early spring. This seemingly long period of time will prove beneficial to both the state office and local districts.

Step 1. An accurate description of each district's standardized testing program for grades 4, 5, and 6 is required. Some states may have this on file, but in most cases, local districts will need to be contacted to gain the necessary

TABLE 3
Washington Assessment

District and Sample Sizes Used in the
Anchor Test Study Data Collection Effort Based on
1972 School Census Data

Number of Pupils in District	Number of Schools with Grade 6	Number of Schools in Sample	Percent in Sample	Number of Students
20,000 and over	226	23	10.2	1747
10,000 – 19,999	162	17	10.5	1469
5,000 – 9,999	142	15	10.6	951
3,000 – 4,999	78	6	7.7	856
2,000 – 2,999	46	5	10.4	764
1,000 – 1,999	59	5	8.5	252
700 – 999	34	4	11.8	229
500 – 699	28	3	10.7	149
300 – 499	42	3	7.1	85
under 300	92	6	6.5	66
TOTALS	911	87		6568

information—and protocol in making the contacts is important. The survey should collect at least the following data by March:

- —names and editions of tests
- —forms and levels of tests
- —grades and times of administration
- —type of "results" available for students and schools
- —an indication of anticipated changes for the next school year
- —the name and phone number of the district test coordinator

Step 2. Since, given the assessment purpose mentioned above, little is gained by testing every pupil at a selected grade level, a sample should be designed that will provide the generalizability and precision desired. If the analysis is to focus on schools, schools should serve as the sample unit, and data can be collected in the form of scores for all students in the sample schools at the selected grade level or in the form of school mean scores for the sampled schools. If the primary interest is in comparing state student achievement to national norms, there may be a special interest in a two-stage sampling process that first selects schools and then selects a sample of students within the schools. This process is more time consuming to implement but requires the involvement of fewer schools and students in establishing the state profile. Perhaps the easiest and most straightforward process, whether the focus is on schools or students, is to collect data from all children in selected schools. In any event the sample should be drawn by March.

Step 3. After the district survey information is analyzed and the test coordinators contacted for necessary clarifications, the sampled schools should be matched with the survey results. This matching process will quickly determine which of the sampled schools lack appropriate testing programs. Since computer programs are available to convert fall and winter data to spring norms (following straight line projections which may add to the unreliability of results) and to transform all standard results-reporting options to raw scores, the crucial elements of the match are correct test editions, forms, and levels.

Step 4. If the number of randomly selected schools without compatible test data is too large (more than 40 to 50 percent), the efficiency advantage of the ATS model will be lost. In this case, another assessment strategy should be investigated. Assuming, however, that a solid majority of schools fits the desired pattern, meetings should be held with officials of the discrepant schools to plan a positive course of action for the coming year. This contact needs to be made in the spring—April or early May—so that adequate implementation time is provided for changes or "add ons" to local testing programs to incorporate one of the ATS tests into the testing schedule. This step holds the key to success and is more a human relations activity than a technical one.

The solution may be unique in each district. In some cases, it may only require a slight alteration in the district or school testing program; in others, the loan of tests from one district to another may be the answer. The state agency may find it convenient to actually provide some of the tests and scoring services. The fact that eight different tests can be used greatly alleviates the problem. As a last resort, if

there is some evidence for supporting the assumption that there is no systematic achievement bias between schools using one of the anchor tests and those which do not, a limited number of alternate schools can be used without seriously affecting the representativeness of the sample.

Step 5. As soon as the final "go" decision is made, the companies publishing the tests in the Anchor Test Study tables should be contacted to provide the related technical manuals. All other necessary materials should also be ordered so that there will be no holdup during the processing or analyzing phases.

Step 6. Early in the fall, all sampled schools should be contacted directly with specific instructions regarding data collection. This memorandum should build on the previous year's survey response and present an exact course of action, always stressing the importance of the schools' contribution to the state assessment. Making the "Hawthorne effect" explicit is an intricate part of the strategy. Since the sampled schools will be using a variety of tests, and testing at different times, the deadline for the submission of data will vary but should be clearly established for each group of schools using a similar pattern.

Step 7. The state office clerical staff should be trained to review the content and quality of the data as they are received and to monitor the due dates. The goal is to routinize the data collection and processing as much as possible. Most of the materials will be accumulating in November and December after the fall testing, and in May and June after the spring testing, so card punch and computer time should be scheduled accordingly.

Step 8. Once the data are processed, the development of the results tables, including means and standard deviations, can take place. This is a technical job, but if the data have been screened carefully as received, there should be little problem. The predominate concern will be to prepare a public report on the assessment that is clear and concise and that does not generalize beyond the power of the data or the rigor of the sampling. The issue of sampling and the power to generalize is crucial in this time of full disclosure, when both the media and the public demand access to information regardless of its technical quality and frequently use it in ways unintended or beyond intent.

A Final Statement

The anchor test approach to reading assessment on a state level is a workable one if one can tolerate its limitations—limitations brought by the uneven distribution of ATS test users, by well-intended but inaccurate information, by the focus on grades 4, 5, and 6, and by the technical and procedural problems previously discussed. If the conditions can be endured or overcome, this approach can produce a reading achievement profile for a state and do it in a way that is not disruptive in local schools or costly at the state level.

REFERENCES

1. Bianchini, J. & Loret, P. *Anchor test study. Final report* Berkeley: Educational Testing Service, 1974. 34 vols.
2. Loret, P., Seder, A., Bianchini, J. & Vale, C. *Anchor test study—equivalence and norms tables for selected reading achievement tests.* Washington, D.C.: U.S. Government Printing Office, 1974.
3. *User's guide to the anchor test program.* Olympia, Washington: Superintendent of Public Instruction, 1975.
4. *Washington statewide assessment using anchor test norms total reading grade six, technical report.* Olympia, Washington: Superintendent of Public Instruction, 1974.