DOCUMENT RESUME

ED 104 896 TM 004 338

AUTHOR Jaeger, Richard M.

TITLE The National Test-Equating Study in Reading (The Anchor Test Study). National Council on Measurement in Education, Vol. 4, No. 4, Summer 1973.

INSTITUTION National Council on Measurement in Education, East Lansing, Mich.

PUB DATE 73

NOTE 9p.; For the Anchor Test Study, see ED 092 601-634 and 094 008

AVAILABLE FROM National Council on Measurement in Education, Office of Evaluation Services, Michigan State Univ., East Lansing, Mich. 48823 (Subscription rate: $2.00 per year; single copies $0.25 each in quantities of 25 or more)

EDRS PRICE MF-$0.76 HC Not Available from EDRS..PLUS POSTAGE

DESCRIPTORS *Elementary School Students; *Equated Scores; Measurement Techniques; Reading; Reading Comprehension; *Reading Tests; Scores; *Standardized Tests; *Testing; Test Reliability; Vocabulary

IDENTIFIERS *Anchor Test Study

ABSTRACT

The Anchor Test Study provides a means for translating a pupil's score on any one of seven widely used standardized reading tests to a corresponding score on any of the other tests for students in grades 4, 5, and 6. In addition, the study provides new estimates of alternate form reliability for each of the seven tests, provides estimates of the intercorrelations among the tests, and explores empirically, some methodological questions in test-equating. The tests used are: California Achievement Tests (1970); Comparative Tests of Basic Skills (1968); Iowa Test of Basic Skills (1970); Metropolitan Reading Tests (1970); Sequential Tests of Educational Progress, STEP Series II (1969); SRA Achievement Series (1970); and Stanford Reading Tests (1964). (The Gates MacGinitie (1964) was added later in the study. Ed.) This newsletter contains four major sections. In the first are the objectives of the study. In the second, the thirty-year history of discussion on the need for a test-equating study and the technical problems inherent in the equating of non-parallel tests are considered. The third section describes the planning of the Anchor Test Study and the results of a feasibility study conducted in 1969. The fourth section is a description of the design of the study and some aspects of its implementation. (Author/RC)

Vol 4, No 4

NCME

measurement in [illegible]

A SERIES OF SPECIAL REPORTS OF THE NATIONAL [illegible]

The National Test - Equating Study in Reading (The Anchor Test Study)

Richard M. Jaeger

The study reported in this issue will soon make it possible to compare directly the results obtained on two or more different reading tests — at least in grades 4, 5, and 6. Heretofore if Johnny took the Stanford Achievement Test one year and took the Iowa Test of Basic Skills the next, we never really knew what the difference between those two measures meant. Likewise, if District A wanted to compare itself with District B but each had taken different tests, observed differences in results were clouded by the fact that different norming samples were used.

The Anchor Test Study is a national undertaking involving 300,000 children in 1,650 elementary schools in 50 states and is designed to overcome the problems of making comparisons when different tests are used.

Author Richard M. Jaeger during his tenure as Chief of Evaluation Design and Chief of Evaluation Methodology in the Bureau of Elementary and Secondary Education, U. S. Office of Education, had a major role in planning and implementing the study. He has held other positions in the area of mathematical statistics and research in a variety of private and public agencies including Space Technology Laboratories and The Aerospace Corporation, General Motors Research Laboratories, Stanford Research Institute, and the National Center for Educational Statistics. He is currently an Associate Professor in Applied Educational Sciences [illegible]niversity of South Florida.

CJF

Richard M. Jaeger

Measurement in Education ordinarily reports on a measurement area of broad interest, such as assessment in the affective domain (Tyler, 1973) criterion-referenced testing (Airasian and Madaus, 1972) or the measurement problems posed by performance contracting (Schutz, 1971). This issue, though, is devoted to a report on a specific measurement project. To warrant this coverage, a project should make an important contribution to measurement theory or practice, and must offer findings potentially useful to a large number of practitioners. The National Test-Equating Study in Reading, a.k.a. The Anchor Test Study, qualifies by virtue of: 1) Its fulfillment of a long-standing objective of the measurement community, i.e., the equating of widely used achievement tests in reading comprehension and vocabulary. 2) Its scope—it required administration of nearly 500,000 reading comprehension and vocabulary tests to over 300,000 children in 1,650 elementary schools in all 50 states. 3) Its widespread support—the study carried the endorsement of the U. S. Commissioner of Education, 49 of the nation's Chief State School Officers, and district superintendents and principals representing more than 1,600 schools in all 50 states. 4) Its quality—it provides new national norms for the achievement tests used, based on an unprecedented school cooperation rate of over 90 percent and a sample more nearly representative of children enrolled in U.S. public and private elementary schools than ever before achieved.

This report on the Anchor Test Study contains four major sections. In the first are the objectives of the study. In the second, the thirty-year history of discussion on the need for a test-equating study and the technical problems inherent in the equating of non-parallel tests are considered. The third section describes the planning of the Anchor Test Study and the results of a feasibility study conducted in 1969. The fourth section is a description of the design of the study and some aspects of its implementation.

WHAT IS THE ANCHOR TEST STUDY?

The Anchor Test Study had two major objectives and several minor objectives. The study was to provide a method for translating a pupil's score on any one of seven widely used standardized reading tests to a corresponding score on any of the other tests. It was also to provide new nationally representative norms for each of the seven tests. These were the major objectives. In addition, the study was to provide new estimates of parallel forms reliability for each of the seven tests, to provide estimates of the intercorrelations among the tests, and to explore empirically, some methodological questions in test-equating.

Comparable nationally representative norms were to be provided for the reading comprehension and vocabulary subtests of seven tests recommended for use with pupils in grades four, five and six. The development of these norms was to be accomplished in the *restandardization* phase of the study.

The ability to translate a pupil's score on the reading comprehension or vocabulary subtest of one test, to an equivalent score on the corresponding subtest of another test, was to be provided in the *equating* phase of the study. Estimates of parallel forms reliabilities and inter-test correlations were to result from the equating phase of the study, as were numerous statistics for the comparison of various test-equating methods and procedures.

A BIT OF HISTORY

The idea of conducting a major test-equating study did not spring full-blown from the mind of a Washington bureaucrat. In fact, the idea appears in the psychometric literature quite frequently, over a period spanning more than three decades.

One of the psychometric issues that motivated the Anchor Test Study was the recognition of differences in the supposedly "national" norms developed for leading achievement tests. The first public acknowledgement of this problem was probably that of H. A. Toops, who in 1939 called for the development of test norms on a "standard million" examinees. He proposed that the characteristics of norms samples — stratification variables, representation within strata, etc. — be the same for all standardizations. In a paper delivered before the 1940 Conference of State Testing Leaders (published in the 1941 Harvard Educational Review), Edward Cureton strongly suggested the development of an anchor test. He had in mind the equating of scores on widely used achievement tests and tests of general ability. If we share Cureton's view of the importance of an anchor test study, this study attains a landmark. For he said: *"Such an achievement will mark the date of maturity of educational and mental measurements as a science, and of educational counseling and guidance as a profession."*

In a review of the quality of test norms, titled *"Norms: 1963"*, Roger Lennon spoke of the progress in norming practice achieved over the preceding decades. He noted that substantial differences remained in the definitions of populations, sampling procedures, sample sizes and participation rates associated with the national norms of leading achievement and aptitude tests. Lennon renewed Cureton's plea for a common set of norms:*"For both general mental ability or scholastic aptitude measures and for achievement tests, there is surely a place and a need for a single, comprehensively based, broadly descriptive set of norms. . ."*.

The Anchor Test Study responds to yet another psychometric issue. In a 1930 article in the *Journal of Educational Psychology*, E. F. Lindquist concluded that a substantial part of the variability in norms distributions was, in the analysis-of-variance jargon, between schools rather than within schools. Of course, he found the variance among school means to be appreciably smaller than the variance among individual test scores. Since the average of achievement test scores for children in the same school is often used to describe school-wide performance, Lindquist cited the need for school mean norms. If the mean raw score for children in a single school is converted to a percentile rank using norms for individuals, substantial errors may result. For school means above the mean of the norm distribution, percentile ranks will be underestimated; for school means below the mean of the norm distribution, percential ranks will be overestimated. Apparently, the testing industry didn't respond to Lindquist, for he reiterated his plea in a paper presented before the 1948 Invitational Conference on Testing Problems. Publishers have since provided school mean norms for only two of the seven tests used in the Anchor Test Study. The study yields comparable school mean norms for all seven tests.

Consideration of major equating studies has not been limited to grade-school applications. The abundance of college admissions tests has, more than once, led to the demand that equating tables be established. An important problem is the lack of parallelism among such tests, and the resulting variability of equating lines over groups. William Angoff discussed this problem in a 1954 address before the American Psychological Association. It was also the subject of an NCME symposium in 1964, and was reported in the very first issue of the *Journal of Educational Measurement*. In that symposium, John Flanagan, William Angoff, E. F. Lindquist and Roger Lennon agreed on the need for equating studies, and on the seemingly insurmountable technical difficulties attending the equating of non-parallel tests.

PLANNING FOR THE STUDY

Planning for the Anchor Test Study was initiated by the U. S. Office of Education in the Spring of 1968. Major purposes and the broad design of the study were formulated after a standard review of relevant literature. Then those who had contributed to the methodological literature on test-equating were asked to comment on the objectives and feasibility of the proposed design. The results of this inquiry were interesting and informative, but inconclusive. Nearly everyone who was asked for an opinion gave one. The reactions fell into two distinct categories: those strongly in favor of the study, and those vehemently opposed. The flavor of these reactions is characterized by the following excerpts: "*I must say at once that I am extremely dubious about the possibility of achieving any worthwhile objectives at any reasonable cost through a study of this kind. I am sure that you are aware that what you are proposing is not only extremely difficult, both technically and politically, but that it involves serious dangers as well.*" And from another respondent: "*It is difficult for me, as one interested in educational testing and educational research, to question the study you are proposing. The inadequacy of test norms, and disparities between norms on similar tests, have long been matters of concern. The project outlined in your letter should help to correct these situations, and should provide interesting additional data. If you can get financial support, and the support of educational leadership, I would urge that the study be done.*"

Given such divergence, it was decided to conduct a small-scale feasibility study before committing the funds necessary to complete a full-scale project. The major question to be explored was the technical feasibility of equating tests that were not designed to be parallel. Two kinds of investigations were made. The first was a judgmental analysis of the similarity of the constructs measured by five[1] of the tests. The second provided estimates of the correlations among these tests.

Three investigators who had written reviews of the five tests for O. K. Buros' *Mental Measurements Yearbook* analyzed the tests. They sought to assess the degree to which the tests measured the same reading skills and abilities. The investigators first developed a set of categories of skills required by the tests. Having agreed upon the skill categories, each investigator independently assigned each test item to a category. The percent of the items in each test assigned to each category by each rater was then computed, and degree of agreement among raters was determined by computing a coefficient of concordance for each test. These ranged from a low of 0.54 for the Stanford Achievement Test, Intermediate Level, to a high of 0.94 for the Metropolitan Achievement Test, Elementary Level. The investigators agreed that the tests required many of the same skills and abilities, but an overall coefficient of concordance equal to 0.67 reflected their judgment of less-than-complete overlap among the tests. To compute this coefficient, the skill categories played the role of ranked items, and the tests played the role of "raters."

In the second part of the feasibility study, the reading comprehension subtests of the five test batteries were administered to over 800 pupils in grades four, five and six. Each pupil completed subtests from three test batteries, arranged in a random order. The pupils who completed the tests attended a purposively-selected sample of schools in the states that surround Washington, D. C. Schools were chosen to represent a wide range of socio-economic compositions and degrees of urbanism. The allocation of tests to schools was carefully balanced so that the groups of pupils that completed each test were similar. The resulting data were used to compute correlation coefficients among the five tests and to complete a principal components analysis of the correlation matrix. The tests were found to be highly correlated. For fourth grade pupils, correlation coefficients ranged from 0.81 to 0.91, and when corrected for unreliability, ranged from 0.86 to 0.96. For fifth grade and sixth grade pupils, the correlations were somewhat lower, with minimum observed values of 0.72 and 0.75. However, when corrected for unreliability, six of the ten inter-test correlations for fifth graders and seven of the ten correlations for sixth graders were above 0.85.

Since the three raters' agreement on the distribution of items in a single test was not much better than their judgment of the similarities among the tests, the correlational data were seen as more compelling than the judgmental data. It appeared that the tests had enough in common to make equating technically worthwhile.

The political feasibility of an anchor test study was another matter of concern since development of representative norms was deemed critical. The major test publishers try conscientiously to secure representative national norms for their tests. Most of their norms samples, if not as efficient as possible, are designed to represent the national population of school children (Lennon, 1964). But between the design and the execution, sample bias inevitably results. Usually a fourth to a third of the school systems invited to participate in norming studies decline. The remaining school systems are only marginally representative of the national population. Most often, positive responses to norming invitations are received from very large school systems (who seem to participate through a feeling of *noblese oblige*), and very small school systems (who seem to participate becuase the incentives offered by the test publishers are judged to be worthwhile). The response rates among middle sized school systems are often less satisfactory.

In order to build a high rate of participation in the Anchor Test Study, a group in the Office of Education worked systematically to inform the nation's most influential education policy-makers of

------

[1] The tests used to investigate the feasibility of an equating study were the Metropolitan Achievement Test, the Stanford Achievement Test, the Iowa Test of Basic Skills, the SRA Achievement Series, and the Sequential Tests of Educational Progress. Two tests not considered in the feasibility study were normed and equated in the main study.

the benefits that would be derived from the study. The potential benefits to the U. S. Office of Education were clear, and the U. S. Commissioner of Education not only endorsed the study, but helped to persuade others of its merit. An important body to be convinced was the Council of Chief State School Officers, a group composed of the top educational administrators from each state. Of major concern to the "Chiefs" was the prospect of comparing pupils' achievement in various states. Presumably, once the major achievement tests were equated, a state using the Stanford Achievement Test, say, could be compared to a state using the Iowa Test of Basic Skills. It is interesting to note that cooperation in National Assessment was being debated at about the same time (1968), and much of its early opposition arose from a desire to avoid interstate comparisons of achievement. In early 1969, the Council of Chief State School Officers signed an agreement with the U. S. Commissioner of Education that called for cooperative evaluation of federally supported educational prog. ams. The only project specifically mentioned in that agreement was the Anchor Test Study, and it carried a proviso that results from a test-equating study would not be used to compare achievement in different states. Thus a major political barrier was removed.

After completion of the feasibility study in 1970, the details of the major study were formulated during a nine month period of intensive planning. Through the work of psychometricians and survey statisticians within the Office of Education, and a distinguished cadre of consultants outside the Office, a prescription for conducting every phase of the study was developed. In June 1971, following competitive bidding, Educational Testing Service was awarded a contract to conduct the Anchor Test Study according to specifications developed. Regardless of the ultimate judgment of its worth, the Anchor Test Study stands as a monument to good planning on the part of the U.S. Office of Education, and to outstanding execution on the part of Educational Testing Service.

DESIGN OF THE ANCHOR TEST STUDY

The choice of reading achievement test most widely used at upper elementary grades for restandardization and equating had been made prior to conducting the feasibility study. Because reading is fundamental to accomplishment in most academic subjects, it seemed reasonable to begin what might be a series of anchor test studies by focusing on reading tests. The decision to begin with tests for pupils in grades four, five and six was based on the recognition that curricula often extend beyond the basic skills in these grades. A child must read competently to benefit from these broadened curricula. Also, a survey on test usage conducted by the U. S. Office of Education (USOE) showed that school systems use achievement tests more often in the upper elementary grades than at any other grade level.

The seven tests selected for the Anchor Test Study are those most widely used in the United States, according to data collected by USOE. A survey showed these tests to be used with well over 90 percent of the 4th, 5th and 6th grade pupils who were given achievement tests.

The specific tests used in the study were the latest available from test publishers at the time tests were distributed to schools. The levels of tests used were those recommended by publishers for pupils in grades four, five and six, and the forms used were those identified by publishers as being used "most often by school systems." The test editions, levels and forms used are as follows:

			Level Used For Grade*		
Title/Edition/Form	Publisher	Subtests	4	5	6
California Achievement Tests (1970) – Reading Forms A and B	CTB McGraw-Hill	Reading Vocabulary Reading Comprehension	Level 3	Level 3	Level 4
Comprehensive Tests of Basic Skills (1968), Forms Q and R	CTB/McGraw Hill	Reading Vocabulary Reading Comprehension	Level 2	Level 2	Level 3
Iowa Test of Basic Skills (1970), Forms 5 and 6	Houghton Mifflin	Vocabulary Reading Comprehension	Level 10	Level 11	Level 12
Metropolitan Reading Tests (1970), Forms F and G	Harcourt Brace Jovanovich	Word Analysis Reading	Elementary	Intermediate	Intermediate
Sequential Tests of Educational Progress, STEP Series II (1969) Forms A and B	Educational Testing Service	Reading	Level 4	Level 4	Level 4
SRA Achievement Series (1970) Forms E and F	Science Research Associates	Vocabulary Reading	Blue Edition	Blue Edition	Green Edition
Stanford Reading Tests (1964), Forms W and X	Harcourt Brace Jovanovich	Word Meaning Paragraph Meaning	Intermediate I	Intermediate II	Intermediate II

THE RESTANDARDIZATION PHASE

The restandardization phase of the Study provided new national norms for the reading comprehension and vocabulary subtests of the 1970 Metropolitan Achievement Test. To develop these norms, more than 192,000 fourth grade, fifth grade and sixth grade pupils were tested in April, 1972. One product of the restandardization phase is a set of tables that associate with each possible raw score on the word analysis subtest and reading subtest of the MAT, a mid-percentile rank and a stanine value. Mid-percentiles and stanines are also associated with each possible total of word analysis and reading subtest scores. Separate tables are provided for fourth graders, fifth graders and sixth graders.

Another product of the restandardization phase is a set of school mean norms for the word analysis and reading subtests of the MAT. In these tables, mid-percentile ranks and stanines associated with each possible raw score provide a normative reference for the average of scores earned by pupils in a given school grade. For example, the average score on the reading subtest for fourth grade pupils in George Washington School can be compared to the average scores of fourth graders in schools throughout the nation. A percentile rank of 65, say, should be interpreted as follows: *"The average reading comprehension of George Washington School's* *fourth graders exceeded the average scores earned by fourth graders in 65 percent of the nation's schools."* The statistic thus applies to a school, not to an individual pupil. Separate school-mean norm tables are provided for fourth graders, fifth-graders and sixth-graders, for the word analysis subtest, the reading subtest and the total of scores on these subtests.

At first glance compiling national norms for a test might appear to be a simple task. All that's required is the administration of tests to a sample of pupils, and the conversion of raw scores in the resulting distribution to percentile ranks, stanines, or some other derived score. However, the vision of simplicity is illusory and purely conceptual. The operational complexity of a good norming study is overwhelming. The sample of pupils tested must be painstakingly designed. It must precisely represent a reference-population of interest, and must be so designed that the derived scores estimated for the sample have only small random deviations from those that would result if the entire population were tested. This last property is known as efficiency. It results from a sample that is designed using all available information on the population of interest. Once the sample is selected a large proportion of those selected must participate in the norming. Materials must be carefully packaged and delivered to schools throughout the nation, on time and in sufficient quantities. Tests must be administered precisely in accordance with standardized directions. The materials must then be assembled correctly, labeled and returned to the study director for scoring, editing and analysis. In the analysis process, raw test scores must be smoothed, weighted and properly compiled before derived scores can be computed.

In the Anchor Test Study, the process of building cooperation began with a letter of endorsement jointly signed by the U. S. Commissioner of Education and the President of the Council of Chief State School Officers. Chief State School Officers in 49 of the 50 states responded affirmatively to this letter[2], and urged school superintendents in their states to participate in the study. Letters of invitation were sent to a sample of school superintendents, and in many cases follow-up telephone calls were completed. Next, the principals of sampled schools were sent personal invitations to participate. The process of securing participants began in November 1971 and ended in April 1972; when it ended, over 91 percent of the schools originally asked to participate in the study had agreed to do so.

Since a high response rate was crucial to the success of the restandardization phase of the study, it was decided that separate samples of pupils would be used for restandardization and equating. Because the equating phase imposed a greater testing burden (each participating pupil completed the reading comprehension and vocabulary subtests of two test batteries), it seemed likely that a smaller proportion of school systems and schools would agree to participate in the equating phase than in the restandardization phase. The data confirmed this assumption.

The restandardization phase was to provide new norms for the reading comprehension and vocabulary subtests of the MAT, based on a probability sample of pupils enrolled in public and non-public schools throughout the nation. A complete description of the design of this sample required forty-three pages in the Final Report on the Anchor Test Study (1972); only a brief overview can be provided here.

The populations from which the restandardization sample was drawn were listed on computer tapes provided by the U. S. Office of Education and the National Catholic Education Association. When sampling efficiency was balanced against cost and convenience, it was decided to use schools as primary sampling units. Later in the study, an attempt was made to update the USOE school listings in a random sample of school systems, and to augment the sampled schools with some that would have been systematically omitted for various reasons.

The design for the restandardization phase called for a sample of 940 schools that had at least one of grades four, five and six. Each school contained in this sample (called a primary school) was to have five similar, randomly-selected, potential substitutes.

Each school in the sample was selected with probability proportional to its size, from a population of schools stratified by type of control (public or private), and a number of other variables. Public schools were stratified by the size of the county or district containing the school, percent of minority enrollment (or if not available, percent of minority population in the school's district), income level of the community where the school is located, and the degree of urbanization of the school's location. Catholic schools (constituting about 11 percent of the total U. S. enrollment in grades four, five and six) were stratified by median income of the pupil's parents, Census region and degree of urbanization. Non-Catholic private schools were stratified only by geographic region. In all, 470 strata were defined, and two primary schools were selected from each stratum.

When the 940 schools were selected for the restandardization phase, most agreed to participate, some were found to be ineligible (e.g., no pupils in grades four, five or six), and some declined the invitation. The figures are as follows:

[2] While the one dissenting Chief State School Officer will not be named, suffice it to say that a correlation of unity exists between electoral college support of the current administration in the 1972 esidential election, and Chief State School Officer support for the nchor Test Study. No causal inference is implied.

826 primary schools participated in the study
27 primary schools were found to be ineligible, and were dropped from the study
87 primary schools declined participation and were replaced by 80 randomly-selected back-up schools

Thus considering the primary sample of schools found eligible to participate in the study, nearly 91 percent agreed to participate. This figure exceeds the goal of 90 percent set by the study's designers, and greatly exceeds the participation rates typical in national norming studies.

Within sampled schools, all pupils enrolled in grades four, five and six who were present on the day of testing were tested. There was no sampling of pupils within schools, even though economy would have dictated such sampling. The decision to test all pupils was based on a desire to reduce operational errors, and to provide the incentive of returned test records for all pupils in participating schools.

The distribution, handling and receipt of test materials for both the restandardization phase and the equating phase of the study was guided by an elaborate computerized management system. Every participating state had a coordinator for the study, as did every participating district and school. In accordance with USOE design specifications, Educational Testing Service established monitoring points for the receipt and mailing of materials, and regularly produced updated computer listings showing the location of each set of test materials. With this system, district coordinators were informed of the location of materials in their districts, and state coordinators were informed of the location of materials in their respective states. The U. S. Office of Education was provided regular summary listings showing the number of states, school districts, and schools that had received materials, completed testing, returned materials, etc. The system worked well; 99.7 percent of those schools that had agreed to participate in the study provided usable test scores. It is a credit to the materials-handling system and to the excellent work of coordinators and teachers in over 1600 schools that usable test results were obtained for 95.1 percent of the fourth, fifth and sixth graders in those schools that provided restandardization data.

THE EQUATING PHASE

The equating phase of the study provided many kinds of results. A principal product is a set of tables that allows translation of an observed reading comprehension or vocabulary score on one of the seven tests, to an equivalent raw score on another of the seven tests. The study provides separate equating tables for vocabulary scores, for reading comprehension scores and for totals of these subtest scores, for pupils in each of grades four, five and six.

For the practitioner who uses tests, the other principal product of the equating phase is a set of norm tables for the six tests other than the MAT. The tables provided are parallel to those described for the MAT. To produce these tables, tests other than the MAT were not restandardized directly. The MAT norms produced in the restandardization phase were combined with the score translation tables produced in the equating phase in the following way. Through the equating process, each score on a non-MAT subtest was associated with its equivalent score on the corresponding MAT subtest. For example, an observed score of 25 on the Vocabulary Subtest of the Iowa Test of Basic Skills (ITBS) for fourth graders might be found to be equivalent to a raw score of 18 on the corresponding subtest of the MAT. In the restandardization process, let us suppose that a score of 18 on the fourth grade word analysis subtest of the MAT was found to correspond to a percentile rank of 34 and a stanine of 4. These derived scores would then be associated with an observed score of 25 on the fourth grade vocabulary subtest of the ITBS. Norming the non-MAT tests through the equating process provided equivalent norms for all corresponding subtests, in the sense that derived scores are referenced to the same representative sample of pupils.

The other results provided by the equating phase of the study are of interest to psychometric researchers. Scores on two tests can be defined as "equivalent" in many different ways. Various mathematical definitions are reviewed by Angoff in the second edition of *Educational Measurement* (Thorndike, 1971). The Anchor Test Study provided an empirical comparison of two definitions of equivalence, and several different equating procedures. By one definition, two raw scores are said to be equivalent if their percentile ranks are equal. This definition is used in the equi-percentile method of equating. In the linear equating method, two raw scores are said to be equivalent if they are the same number of standard deviation units above (or below) the means of their respective score distributions. The linear method of equating assumes that two distributions of scores have the same basic shape, and differ only in their means and variances. The translation of a score from the scale of one distribution to the scale of another involves a simple linear transformation that adjusts for differences in means and variances. One technical objective of the study was the comparison of equating tables that resulted from the linear equating method and the equi-percentile equating method. For many test pairs, the linear method resulted in impossible score values (negative test scores, or scores that exceeded the number of items in the subtest); the equi-percentile method was judged to be best.

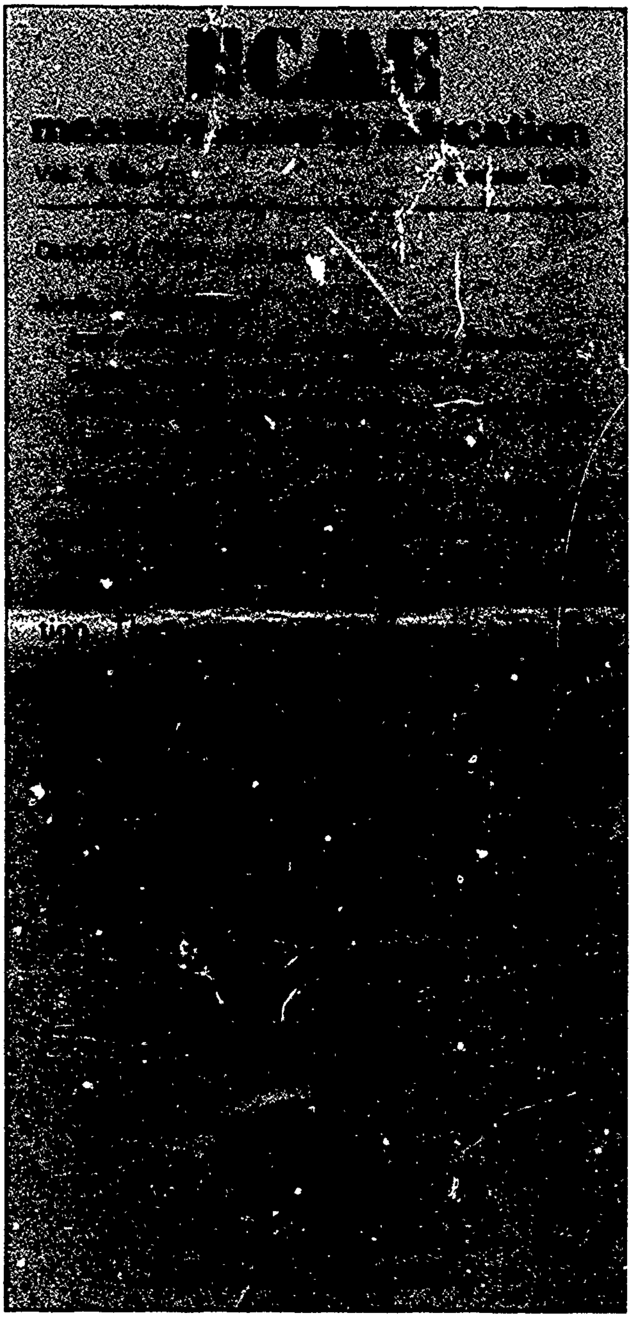
Another technical investigation involved different procedures for compiling the distributions of scores to be equated. Four procedures were investigated, and the linear and equi-percentile methods of equating were used with each procedure. To describe the procedures, it is necessary to consider the way in which pupils were sampled for the equating phase of the study.

Every pupil who participated in the equating phase of the study completed the reading comprehension and vocabulary subtests of two test batteries, usually one battery in the morning and the other in the afternoon. Since there were seven tests in the study, each could be paired with the remaining six tests, for a total of 42 combinations when order of administration is considered. In order to estimate the test-retest reliabilities, each test was also paired with its most commonly used parallel form. If the tests are labeled A through G, the possible pairings can be described schematically as follows:

TEST	A	B	C	D	E	F	G
A	AA' A'A	AB	AC	AD	AE	AF	AG
B	BA	BB' B'B	BC	BD	BE	BF	BG
C	CA	CB	CC' C'C	CD	CE	CF	CG
D	DA	DB	DC	DD' D'D	DE	DF	DG
E	EA	EB	EC	ED	EE' E'E	EF	EG
F	FA	FB	FC	FD	FE	FF' F'F	FG
G	GA	GB	GC	GD	GE	GF	GG' G'G

An entry in this table represents a pair of tests administered in the order indicated. For example, "BD" means that tests B and D were administered, with B administered first. The entry "AA' " denotes the administration of test A and its most commonly used parallel form A' with form A administered first. The way the equating sample was designed, a given school was to administer a pair of tests (such as tests A and C) to two class-sections in each of grades four, five and six.[3] In one class-section per grade, the tests were to be administered in the order AC, in the other section, they were to be administered in the order CA. Thus twenty-one schools were needed to administer the test combinations denoted by the non-diagonal entries. An additional seven schools were needed for the diagonal entries. In these schools, two class-sections in each of grades four, five and six completed two parallel forms of a given test (such as test A); the order of administration was A followed by A' in one class-section per grade, and A' followed by A in the other section. In all, twenty-eight schools were needed to complete all of the test pairings shown in the table. It was decided to replicate the design shown in the table sixteen times, so that each pair of tests (or parallel forms) would be administered in both orders in sixteen different schools. The sample of schools was completely separate from those participating in the restandardization phase of the study.

One procedure for compiling score distributions pooled the test scores of all pupils who completed a given test, whether the pupil completed the test as the first member of a pair or the second member of a pair. For example, all of the scores for test A were pooled using data from all pairs in the table where test A appears. When this procedure is used, data from one seventh of the non diagonal pairs in the table are used to compute equivalent scores. Frederic Lord developed two additional procedures that used data



from all of the pairs in the table to compute equivalent scores for each pair of tests. The two procedures were nearly identical, the only difference between them being the order in which combinations of tests are used. Unfortunately, these estimation methods are far too complex to be described in a short report; a complete description can be found in the Final Report on the Anchor Test Study (1972). The final procedure explored in the study averaged the equating results produced by the two procedures developed by Lord.

To compare the four equating procedures and two equating methods, two kinds of error measures were estimated. One measure reflected the degree to which the equating results would vary if the same equating procedure and method were applied to different representative samples of pupils. Another error measure, called the conditional root-mean-square error of equating, reflected the degree to which a score read from the equating tables would differ from the score a pupil would have earned, had he been given the equated test. When these results were analyzed, it was found

[3] Some of the sampled schools did not have two class-sections in each grade. In these cases, similar schools were combined to create the requisite school units or "pseudo-schools." To complete this cess, over 700 ind al schools were sampled.

that the average of Lord's two equating procedures provided the smallest errors most often.

AVAILABILITY OF THE RESULTS

At the time this article was written, evaluation of the Anchor Test Study results was still in progress and the U. S. Office of Education had not provided public release of the findings. By publication time, results of the study may well be available.

The U. S. Office of Education plans to make the results of the study available in several forms. A complete technical report containing all data tables (thirty-one volumes containing over ten thousand pages) is to be placed in the ERIC Information System. A "User's Manual for Interpreting Scores" is to be sent to all state departments of education, school districts and schools that participated in the study, and made available to others through the U. S. Superintendent of Documents. It will also be placed in the ERIC Information System. The User's Manual is intended to provide those results most useful for educational practitioners, and will be written in non-technical language.

SUMMARY AND ACKNOWLEDGEMENTS

This report on the Anchor Test Study reviews its history and its highlights. A host of methodological investigations and technical findings beyond those mentioned here are reported in the Project Report of the Final Report on the Anchor Test Study (1972). A wealth of additional information can be found in the thousands of tables and graphs that make up the Final Report.

During the five years that the Anchor Test Study was being considered, planned and finally conducted, the staffs of two private companies, one government agency, innumerable independent consultants and thousands of schoolmen at all levels of government worked with dedication and devotion. Not every contributor can be mentioned here, but to ignore some would constitute a grave injustice, for I have borrowed heavily from their ideas and their writings in compiling this report. In the U. S. Office of Education; Dr. Charles Hammer and Mr. Harold Nisselson have contributed immeasurably to the technical excellence and operational feasibility of the study. Dr. David Orr of Scientific Educational Systems, Incorporated directed the implementation of the feasibility study. Drs. David Chapman, Morris Hansen and Thomas McKenna worked with Mr. Nisselson to formulate the sampling design for the study and to complete all phases of sampling. The design of the study owes much to the contributions of Drs. Frederic Lord, William Angoff and Walter Durost. Dr. Peter Loret directed the implementation of the Anchor Test Study for Educational Testing Service. Dr. John Bianchini served as principal investigator, Dr. Alan Seder coordinated the distribution and receipt of materials, and Dr. Carol Vale was associate investigator. It is largely through their efforts that a design became reality.

REFERENCES

Airasian, Peter W. and George F. Madaus, Criterion-referenced testing in the classroom, *NCME Measurement in Education*, 1972, 3, 4.

Anchor Test Study Final Report — Project Report, Princeton, New Jersey: Educational Testing Service, 1972. Note: This report is not publically available.

Angoff, William, The equating of non-parallel tests , paper presented before the 1954 Annual Meeting of the American Psychological Association.

Angoff, William, Equating non-parallel tests , *Journal of Educational Measurement*, 1, 1, 1964.

Angoff, William, Scales, norms and equivalent scores , in R. L. Thorndike (ed.), *Educational Measurement*, 2nd. Ed., Washington: American Council on Education, 1971, p. 508-600

Cureton, Edward E., Minimum requirements in establishing and reporting norms on educational tests , *Harvard Educational Review*, 11, 1941, p. 287-300; also presented before the 1940 Conference of State Testing Leaders.

Flanagan, John, Equating non-parallel tests , *Journal of Educational Measurement*, 1, 1, 1964.

Lennon, Roger, Equating non-parallel tests , *Journal of Educational Measurement*, 1; 1, 1964.

Lennon, Roger, Norms: 1963 , *Proceedings of the 1963 Invitational Conference on Testing Problems*, Princeton, New Jersey: Educational Testing Service, 1964.

Lindquist, E. F., Factors determining the reliability of test norms , *Journal of Educational Psychology*, 21, 1930, p. 512-520.

Lindquist, E. F., Equating non-parallel tests , *Journal of Educational Measurement*, 1, 1, 1964.

Schutz, Richard E., Measurement aspects of performance contracting , *NCME Measurement News*, 1971, 2, 3.

Toops, H. A., A proposal for a standard million in compiling norms , *Ohio College Association Bulletin*, No. 125, 1939; also presented before the 1939 Conference of State Testing Leaders.

Tyler, Ralph W., Assessing educational achievement in the affective domain , *NCME Measurement in Education*, 1973, 4, 3.