ED 093 904 TM 003 719

AUTHOR Jaeger, Richard M.
TITLE Some New Developments and Discoveries for Evaluative Analysis.
PUB DATE [Apr 74]
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Data Analysis; *Equated Scores; *Evaluation Methods; Raw Scores; Research Design; *Sampling; Statistical Analysis
IDENTIFIERS Anchor Test Study

ABSTRACT

Two relatively new tools for analysis of data compiled in evaluation studies are presented. The National Test-Equating Study in Reading, known as the Anchor Test Study, produced tables of score-correspondence between the eight reading comprehension and vocabulary tests most widely used in the United States. Two types of tables from this report should be most useful for educational evaluation studies. First are tables that show equivalent raw scores for individual pupils for each of the 28 pairs of reading comprehension and vocabulary tests. The structure of these tables is presented. A second set of equivalency tables, nearly identical in structure, shows equivalent raw-score means for schools, on each of the 28 pairs of tests. Separate tables are provided by the study for individuals and groups of pupils in grades 4-6. To illustrate the use of the Anchor Test Study results in educational evaluations, four designs are considered. A second tool considered is sampling theory, not often used to advantage in educational evaluation. For studies that involve populations exceeding 300 units of analysis, both matrix sampling and respondent sampling should be considered. References are given for further discussions of these methods. (RC)

SOME NEW DEVELOPMENTS AND DISCOVERIES FOR EVALUATIVE ANALYSIS[1]

Richard M. Jaeger
University of South Florida

Claiming novelty for a technical procedure in evaluation is a dangerous business. More often than not, what seems to be an innovation today is in reality a rediscovery of some once-often-used-but-now-lost technique. Those of us in evaluation can make but a sad case for our skill as archivists; or perhaps we should take additional course work in the history of education.

Two examples come readily to mind: The educational assessment movement and criterion-referenced testing. It cannot be denied that large scale educational assessments have enjoyed widespread growth in popularity during the past few years. From Dr. Tyler's leadership with CAPE has come a vital and promising national assessment program. State departments of education have, in just a few short years, moved from open resistance to large-scale comparative assessment, to the investment of thousands, if not millions, in their own assessment programs. Large school districts too, are now contracting with the major testing companies to design and conduct district-wide assessments. One would suspect that these latter activities are unprecedented, and a natural outgrowth of the national and statewide assessment programs. But some day when you find yourself in a good library with an hour to kill, pick up the 1916 Yearbook of the National Society for the Study of Education. In it you will find the report of a district-wide assessment, conducted by Elwood P. Cubberly for Salt Lake City in the previous year. Are the modern assessments different? Certainly. But hardly to a degree that would support a claim to innovation. The case for novelty of criterion-referenced testing was dealt a serious blow in 1972 by Peter Airasian and George Madaus, in an article in Measurement in Education. They cite a 1918 statement by E.L. Thorndike, defining the difference between criterion-referenced and norm-referenced testing, and then note the use of criterion-referenced measurement in a 1916 study by the Boston Public Schools.

[1] Presented at the 1974 Annual Meeting of the American Educational Research Association, Chicago, Illinois.

The Boston study (by F. W. Ballow) is also reported in the Fifteenth Yearbook of the National Society for the Study of Education, interestingly titled Standards and Tests for the Measurement of the Efficiency of Schools and School Systems.

With this background of claimed innovation rent asunder, it is with some trepidation that I propose two relatively new tools for evaluation studies. Let me say at the outset that neither represents entirely new technology; just procedures that haven't been used to a great degree in past evaluation studies. Unless I haven't found the right volume of history.

The National Test-Equating Study in Reading, more commonly known as the Anchor Test Study, was completed in 1973. As reported in an AERA symposium held last year, (Bianchini, 1973; Lovet, 1973; Jaeger, 1973) the Study produced tables of score-correspondence between the eight reading comprehension and vocabulary tests most widely used in the United States. When the final report on the Study is released by the U.S. Office of Education later this year, two types of tables should be most useful for educational evaluation studies. First are tables that show equivalent raw scores for individual pupils for each of twenty-eight pairs of reading comprehension and vocabulary tests. The structure of these tables is shown in Table 1. A second set of equivalency tables, nearly identical in structure, shows equivalent raw-score means for schools, on each of the twenty-eight pairs of tests. Separate tables are provided by the Study for individuals and groups of pupils in grades four, five and six.

To illustrate the use of Anchor Test Study results in educational evaluations, consider four designs proposed by Andrew Porter in an AERA paper delivered last year (Porter, 1973). Consistent with Porter's paper, I will label these designs Case I and Case II, and within each design, specify Situation A and Situation B. Case I represents the rarely-occurring evaluator's dream wherein units of analysis are randomly assigned to an xperimental group and a control group. An observation on a variable of interest s made at the outset of the experiment, the treatment to be evaluated is applied to the experimental group, and a final observation is then made on both the experimental group and the control group. This is nothing more than the randomized pre-post, two-group design suggested by Campbell and Stanley (1963). To use the Anchor Test Study results with this design, the variable observed at the outset of the experiment, at the end of the experiment, or at both times, must be reading comprehension or reading vocabulary. Porter's Case II represents the more frequent situation wherein units of analysis are assigned to an experimental group or to a comparison group in some purposeful (non-random) way. Again, pre-treatment and post-treatment observations are made on both groups, so we have a pre-post two-group design without random assignment.

Situation A occurs when the tests used for pre-treatment measurement and post-treatment measurement are parallel, and Situation B occurs when different, or non-parallel, measurements are made pre-treatment and post-treatment. These designs are shown in the Campbell and Stanley notation in Table 2.

Porter examined the efficiency of four analysis strategies for each of these designs: Analysis of covariance with a random covariate, analysis of variance with an index of response as the dependent variable, repeated measures analysis of variance, and analysis of covariance with estimated true-scores as a random covariate.

For Case I - Situation A designs (random assignment and use of parallel pre- and post-tests), analysis of variance on the dependent variable: Post-test score minus within-group reliability times pre-test score (an index of response) was found to be the most efficient analysis procedure. To be efficient, this procedure requires that the reliability of the pre and post measures be known. When reliability is not known, analysis of covariance using pre-test scores as a covariate is recommended by Porter.

When pupils are the units of analysis, the Anchor Test Study tables of score-correspondence for individuals can facilitate use of the strategies recommended by Porter. Pre-test measures and post-test measures can consist of scores on any of the eight reading comprehension tests or any of the eight vocabulary tests equated in the Anchor Test Study. Prior to analysis, the scores of individual pupils must be converted to scores on a single test, using the equating tables.

The procedure for conversion of scores is quite simple. First, the test to be used in the final analysis must be selected. A logical choice is the test for which the largest number of data are available. Conversion of scores from one test to another adds an error of equating to the omnipresent error of measurement. Although the standard error of equating is typically one-fourth to one-half a raw-score point (compared to standard errors of measurement typically in the range two to four raw-score points), the two types of errors are cumulative and should best be minimized. Once the test to be used in the analysis has been selected, the Anchor Test Study tables are used to convert scores on all other tests to scores on the analysis test. If test results for pupils are available in raw-score form, the Anchor Test Study tables can be used directly. If results are available only in some standard-score form (such as percentile ranks or grade equivalent units), publishers' norms must be used to convert back to equivalent raw scores, prior to using the Anchor Test Study equating tables.

The Anchor Test Study provides precise estimates of parallel-forms test-retest reliabilities for the eight reading comprehension and vocabulary subtests equated. If the experimental and control groups used in the evaluation can be assumed to be equal in heterogeneity on these measures to the nationwide populations of fourth, fifth, or sixth-graders, the Anchor Test reliability estimates can be used with the analysis of variance on index of response. If scores on different tests are converted to scores on a single test, the reliability of the pooled measures might be estimated from a weighted average of the reliabilities reported in the Anchor Test Study. The weights to be applied would equal the proportions of scores available on each test equated. Although this procedure would generate some error, the test-retest reliabilities differ by no more than 0.08 across tests, and for the comprehension subtests used with pupils in grade five, differ by no more than 0.03; the error should therefore be minimal.

When schools are used as units of analysis in a Case I - Situation A evaluation design, the Anchor Test Study equating tables for school means might prove useful. Procedures for conversion of scores would be identical to those used with the tables for individuals. It is a seeming paradox of classical test theory that the reliabilities of group means are no larger than the reliabilities of individual scores, provided individuals are randomly assigned to groups. The standard error of measurement is smaller for group mean scores, but that is of no consequence for the analysis of variance on an index of response. Given the assumption of random assignment then, the Anchor Test Study reliability estimates might still be useful when schools are units of analysis.

For Case I - Situation B designs, Porter recommends use of analysis of covariance with pre-test scores as the covariate. This is somewhat less efficient than analysis of variance on an index of response, particularly when sample sizes are small. However, when the pre-test and post-test measures are different, the regression coefficient of post-test ~~and~~ on pre-test will probably be unknown, making the ANOVA procedure less efficient. Using the Anchor Test Study equating tables, some formerly Situation B designs might be converted to Situation A designs. Eight different tests could be used as pre-treatment or post-treatment measures, and a Situation A design would obtain, provided the proportion of scores obtained on each test was approximately the same for the pre-treatment and post-treatment measurements. Thus the Anchor Test Study tables would permit use of a more efficient analysis procedure.

For both Case II - Situation A and Case II - Situation B designs, Porter argues that analysis of covariance, with estimated true scores on the pre-treatment measure as the covariate, is often the method of choice. An alternative contender in Case II - Situation A designs is analysis of variance of gain scores. This method is best only when the pre-test and post-test measure the same variable with equal reliability. As in Case I, the Anchor Test Study results might permit the use of this simpler analysis procedure where it would otherwise be infeasible. If reading comprehension were measured using any of the eight Anchor Test Study subtests, the assumptions necessary to use ANOVA of gain scores would probably be met, provided equal proportions of pupils received the same subtest as a pre-test and a post-test. Pupils' scores on the eight comprehension subtests would be converted to scores on a single subtest using the Anchor Test Study equating tables, just as described for Case I designs. If it is necessary to use analysis of covariance with estimated true pre-test score as the covariate, the reliability estimates provided by the Anchor Test Study might again prove useful. The necessary assumption is that the test scores of pupils assigned to the experimental and control groups have variance equal to that of the populations of fourth, fifth or sixth-graders in the nation. If a variety of Anchor Test Study subtests are used in the evaluation, the reliability of scores in the converted-score pool could be estimated using the weighted average procedure described above. The parallel-forms test-retest reliabilities provided by the Anchor Test Study are the type needed to properly estimate true-scores on the covariate.

A second tool that deserves more use than it gets in evaluation studies is sampling theory. Although sampling theory has been widely applied in sociological studies, agricultural research and population surveys, it is not often used to advantage in educational evaluation. I'm not claiming that we always collect measurements on entire populations, but that we either use simple random sampling exclusively or treat our data as though we've used simple random sampling. In 1963, Cronbach proposed that evaluations could be conducted more efficiently if samples of pupils completed different samples of test items. Matrix sampling is now used in National Assessment, and in increasing numbers of statewide assessments. Alex Law will describe one such application later in this program. Neither matrix sampling nor sophisticated examinee-sampling designs have been used to great advantage in local assessments or evaluation programs, and their data-collection efficiency has suffered for the omission.

For evaluations that involve populations exceeding three hundred units of analysis (either pupils, classrooms or schools), both matrix sampling and respondent sampling should be considered. Neither topic lends itself well to a two-minute discussion, so I'll merely call your attention to work on matrix sampling by Shoemaker (1970, 1971) and Bunda (1973) reported in the Journal of Educational Measurement, and to my own work on examinee sampling, reported at this meeting last year (Jaeger, 1973). To provide a glance at the latter work, Table 3 shows the numbers of sixth-grade pupils that would have to be sampled, in order to estimate the mean reading achievement in a school district with 1180 sixth-graders, within 0.2 grade equivalent units with 95 percent confidence. Required sample sizes are shown for seventeen different sampling and estimation procedures. It is clear that simple random sampling is far from being the most efficient procedure. If you'd like more details, I'd be happy to send you the entire paper.

Having begun with the contention that not much is really new in the technology of evaluation, I've tried to show how two relatively old tools could be used in new ways. Although the technology of test equating is decades old, its use on a national scale in the Anchor Test Study may well have produced a significant tool for educational evaluators. Sampling theory too, has been available to us for decades. Perhaps we can increase the efficiency of our evaluations by employing it wisely.

Or perhaps there is a new technology of evaluation. I share your anticipation of the coming hour.

REFERENCES

1. Airasian, Peter W. and George F. Madaus, "Criterion-referenced testing in the classroom," Measurement in Education, May, 1972.

2. Bianchini, John C., "The national test-equating study in reading: results of the study", Presented at the 1973 Annual Meeting of the American Educational Research Association, New Orleans, La.

3. Bunda, Mary Anne, "An investigation of an extension of item sampling which yields individual scores," Journal of Educational Measurement, 10, (1973), pp. 117-130.

4. Campbell, Donald T. and Julian C. Stanley, "Experimental and quasi-experimental designs for research on teaching," Handbook of Research on Teaching, ed. N.L. Gage. Chicago: Rand McNally, 1963, pp. 171-246.

5. Cubberly, Elwood P., "Use of standard tests at Salt Lake City, Utah," 15th Yearbook of the National Society for the Study of Education, Chicago: Univ. of Chicago Press, 1916, pp. 107-110.

6. Jaeger, Richard M., "An evaluation of sampling designs for school testing programs," presented at the Annual Meeting of the National Council on Measurement in Education, 1973, New Orleans, La.

7. Jaeger, Richard M., "The national test-equating study in reading: origins of the study and its historical antecedents," presented at the 1973 Annual Meeting of the American Educational Research Association, New Orleans, La.

8. Loret, Peter, "The national test-equating study in reading: administration of the study," presented at the 1973 Annual Meeting of the American Educational Research Association, New Orleans, La.

9. Porter, Andrew C., "Analysis strategies for some common evaluation paradigms", presented at the 1973 Annual Meeting of the American Educational Research Association, New Orleans, La.

10. Shoemaker, David M., "Allocation of items and examinees in estimating a norm distribution by item-sampling", Journal of Educational Measurement, 7, (1970) pp. 123-128.

11. Shoemaker, David M., "Further results on the standard errors of estimate associated with item-examinee sampling procedures," Journal of Educational Measurement, 8, (1971), pp. 215-220.

TABLE 1. ANCHOR TEST EQUATING STUDY: GRADE 4

TEST 1 READING, 1970 EDITION: EQUIVALENT RAW SCORES FOR READING VOCABULARY

PROCEDURE 1, EQUIPERCENTILE METHOD

TEST 1 LEVEL 3 FORM A VOCABULARY RAW SCORE	TEST 2 LEVEL 2 FORM Q VOCABULARY EQUIVALENT SCORE	TEST 3 LEVEL 10 FORM 5 VOCABULARY EQUIVALENT SCORE	TEST 4 ELEMENTARY FORM F WORD KNOWLEDGE EQUIVALENT SCORE	TEST 5 LEVEL 4 FORM A PART I (VOCAB.) EQUIVALENT SCORE	TEST 6 BLUE FORM E VOCABULARY EQUIVALENT SCORE	TEST 7 INTERMEDIATE I FORM W WORD MEANING EQUIVALENT SCORE
40	40	38	50	30	42	38
39	39	37	50	29	42	36
38	38	37	50	28	41	35
37	37	36	49	28	40	34
36	37	35	49	27	39	33
35	36	34	48	27	39	32
34	35	33	48	26	38	31
33	34	33	47	25	37	30
32	34	32	47	25	36	29
31	33	31	46	24	35	28
30	32	30	46	24	34	27
29	31	29	45	23	33	26
28	31	28	45	22	31	25
27	30	27	44	22	30	24
26	29	26	43	21	29	23
25	28	25	43	21	28	22
24	27	24	42	20	26	21
23	26	23	41	19	25	20
22	25	21	40	19	23	19
21	24	20	39	18	22	19
20	23	19	37	18	20	18
19	22	18	36	17	19	17
18	21	17	35	16	18	16
17	20	16	33	16	17	15
16	19	15	31	15	15	14
15	17	14	29	14	14	13
14	16	13	27	13	13	12
13	15	12	24	12	12	11
12	14	11	22	12	12	10
11	13	10	19	11	11	9
10	11	9	17	10	10	9
9	10	9	15	9	9	8
8	9	8	13	8	8	7

Table 2: Designs for Evaluation Studies Where Anchor Test Study Results May Facilitate Analysis

Case I - Situation A	Case I - Situation B
$R\ O\ X\ O$ $R\ O\ -\ O$ (Parallel pre and post-treatment measures, randomized assignment)	$R\ O_1\ X\ O_2$ $R\ O_1\quad O_2$ (Non-parallel pre and post-treatment measures, randomized assignment)
Case II - Situation A	**Case II - Situation B**
$\underline{O\ X\ O}$ $O\quad O$ (Parallel pre and post-treatment measures, non-randomized assignment)	$\underline{O_1\ X\ O_2}$ $O_1\quad O_2$ (Non-parallel pre and post-treatment measures non-randomized assignment)

Table 3: Sizes of Samples Required to Estimate Mean Reading Achievement within ± 0.2 Grade Equivalent Units with 95 Percent Confidence.*

Sampling and Estimation Procedure	Required Sample Size
Simple Random Sampling (SRS)	106 pupils
Stratified Sampling by Lorge-Thorndike Ability Test Scores - Six Strata:	
Proportional Allocation (Strat-prop)	26 pupils
Optimal Allocation (Strat-opt)	25 pupils
Linear Systematic Sampling:	
Alphabetic Order (LSS-alpha)	≤ 59 pupils**
Increasing Order of Lorge-Thorndike Scores (LSS-inc)	≤ 59 pupils**
Increasing Order of of Lorge-Thorndike Scores; End Corrections Used (LSS--E.C.)	≤ 59 pupils**
Order Reversed in Alternate Strata (LSS-O.R)	≤ 59 pupils**
Centrally Located Systematic Samples (CSS)	≤ 59 pupils**
Balanced Systematic Sampling (BSS)	118 pupils
Single Stage Cluster Sampling:	
Unbiased Estimation, Schools Used as Clusters (RSC-schools-unb)	1041 pupils
Ratio Estimation, Schools Used as Clusters (RSC-schools-rat)	394 pupils
Probabilities Proportional to School Enrollments, Schools Used as Clusters (PPS-schools)	577 pupils
Probabilities Proportional to Fifth-Grade SCAT Score Totals, Schools Used as Clusters (PPES-schools)	236 pupils
Unbiased Estimation, Classrooms Used as Clusters (RSC-class-unb)	865 pupils
Ratio Estimation, Classrooms Used as Clusters (RSC-class-rat)	262 pupils
Probabilities Proportional to Classroom Enrollments, Classrooms Used as Clusters (PPS-class)	314 pupils
Probabilities Proportional to Lorge-Thorndike Score Totals, Classrooms Used as Clusters (PPES-class)	53 pupils

*Midcity data, population size = 1180 sixth-grade pupils.

**Five percent is the smallest sampling fraction investigated. Smaller sampling fractions might provide acceptable precision for these sampling methods.