DOCUMENT RESUME

ED 155 215 95 TM 007 267

AUTHOR	Porter, Andrew C.; And Others
TITLE	Impact on What?: The Importance of Content Ccvered. Research Series No. 2.
INSTITUTION	Michigan State Univ., East Lansing. Inst. for Research on Teaching.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C. Basic Skills Group. Teaching Div.
PUB DATE	Feb 78
CONTRACT	400-76-0073
NOTE	37p.
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Achievement Tests; Arithmetic; *Content Analysis; *Course Content; *Course Evaluaticn; Elementary School Mathematics; Evaluation Criteria; Evaluation Methods; Grade 4; Intermediate Grades; *Item Analysis; Program Evaluation; Standardized Tests; Tests of Significance; *Test Validity
IDENTIFIERS	*Content Validity

ABSTRACT

Defining practical significance in program evaluations is a difficult measurement problem which can only be solved by an intimate familiarity with the measures on which effects are estimated and their content relationship to the prcgram goals. Past attempts to provide general solr·ions tc the size of effect problem have relied on standardized adices which can be estimated and reported without any knowledge c what was measured. Such efforts are viewed here as steps in the wrong direction. Instead, what is called for is a procedure whereby the content goals of the program, the content implied by a test, and the interrelationship between the two are made explicit. The procedure should investigate treatment-by-item interactions and at the same time, describe the measures used so that persons other than the evaluator can reach their own decisions about practical significance. Analysis of the mathematics sections of four major intermediate level standardized tests (Iowa Tests of Basic Skills, Metropolitan Achievement Tests, Stanford Achievement Tests, and California Test of Basic Skills) with their taxonomies indicated rather substantial differences in content tested. It was clear that standardized tests are not well suited to the task of estimating item domain by treatment interactions. (Author/CTM)

Research Series No. 2

IMPACT ON WHAT?:
THE IMPORTANCE OF CONTENT COVERED

by Andrew C. Porter,
William H. Schmidt, Robert E. Floden,
and Donald J. Freeman

Published by

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

February, 1978

The work reported herein in sponsored by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Teaching Division of the National Institute of Education, United States Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-76-0073)

Abstract

Efforts to define the impact of programs have resulted in an important distinction between the statistical question of reliability of effects and the measurement question of size of effects. Presented here is a discussion of the size of effect question. The size of any program effect, however, cannot be interpreted without first knowing how the general effect has been constructed from its components. In this paper, the authors review points made in the literature on the size of effect question and then focus on the more fundamental question of the construction of an aggregate program effect.

In a mathematics program, for example, the same aggregate effect might be produced by a large gain in computational skills or by a large gain in understanding of mathematical concepts. Individuals or school districts, however, may place a higher value on one of these two areas. Similarly, an effect in an area to which considerable program resources were devoted would have different meaning than an effect in an area to which no resources were devoted.

The selection or construction of measures of program effects (standardized tests, for example) is thus a crucial issue in evaluation. Clearly, a small aggregate effect on a test in which all parts are consistent with the program goals has different meaning from a small aggregate effect on a test which has only 50% overlap with the program goals.

Standardized norm referenced tests are typically designed to maximize individual differences and are not necessarily well suited to estimate program impacts. Rather, tests should be chosen or constructed on the basis of the content or goals of the program to be evaluated.

In a current IRT study of the content of fourth grade mathematics, a method of describing content was developed through an iterative process of analysis and classification of items on standardized tests, beginning with the mathematics sections of the most widely used standardized tests: the Stanford Achievement Test (SAT), the Iowa Test of Basic Skills (Iowa), the Metropolitan Achievement Test (MAT), and the California Test of Basic Skills (CTBS).

Substantial differences were found among the standardized tests. On the Iowa, 40% of the items were story problems, compared to 22% for the CTBS, for example. Clearly, the standardized tests selected can interact with the content of instruction in ways that could produce dramatically different aggregate estimates of program impact.

Such analyses of tests and instructional materials lead to new approaches in program evaluation. Test selection and construction can be improved by attention to the content areas emphasized. Analysis of materials can be used to provide a better match between instruction and

evaluation. The analysis might also be extended to the content presented in programs, which might prove useful for comparisons of programs or for studies of program implementation.

Once the content areas covered by a measure and the procedure used to aggregate effects in these areas is understood, the problem of size of effect must still be addressed. But no sensible solution can be offered until the aggregation in the outcome measure is better understood.

Contents

Figures

Tables

IMPACT ON WHAT?: THE IMPORTANCE OF CONTENT COVERED[1]

by Andrew C. Porter,
William H. Schmidt, Robert E. Floden,[2,3]
and Donald J. Freeman

Introduction

When defining the impact of programs it is important to distinguish between the reliability of effects (a statistical question) and the size of effects (a measurement question). The statistical question has already been adequately defined, but the measurement question has not. Traditionally, the measurement question has been stated: "What size must a program effect attain to be practically significant?" Practical significance, in turn, has been deliberately or inadvertently equated with various indices such as statistical significance, strength of an association, or standard deviation units. These efforts to define practical significance, however, disregard the fact that any program effect is estimated with an aggregate measure which cannot be interpreted

[1] Paper presented at the 1977 meeting of the Evaluation Research Society.

[2] The authors wish to thank Jack Schwille for his helpful comments and critique on an earlier draft of this paper.

[3] The authors are senior researchers at the Institute for Research on Teaching and members of the Outcome Measurement research group. In addition, Andrew C. Porter is the group coordinator and a professor of educational psychology. William H. Schmidt is an associate professor of educational psychology, and Robert E. Floden as assistant professor of teacher education. Donald J. Freeman is an associate professor of student teaching and professional development and a professor in the University College, Student Affairs.

without first considering the components aggregated. The primary focus of this paper, therefore, is the fundamental question of the composition of an aggregate program effect. The analysis begins with a critical review of efforts to define practical significance. It then focuses on empirical unidimensionality and item-by-treatment interactions, two concepts which are central to the interpretation of aggregate program effects. The paper concludes with an illustrative content analysis of standardized mathematics tests and a brief discussion of how such content analyses are significant for the size of effects problem.

Past Efforts to Define Practical Significance

Four general problems with past attempts to assess the size of effect can be identified. First, many researchers confuse practical significance with statistical significance; neither type of significance implies the other. This confusion is perhaps the most frequent misinterpretation of the size of program effects. In the behavioral sciences, this confusion may well stem from an historical preoccupation with testing the null hypothesis. Since the results of tests for statistical significance are on a dichotomous scale (significant or not), there is little information immediately available to provide further guidance. Some investigators have attempted to squeeze extra meaning from significance tests by reporting results from "almost significant" to "highly significant." It is well known, however, that any nontrivial null hypothesis can be rejected given sufficient precision of analysis. For example, an F test statistic for differences of means has sample size in its numerator and so can be manipulated quite independently of the effect being investigated. In that sense, the null hypothesis can be thought of as a "straw man" (Kempthorne & Folks, 1971, p. 347) and any failure to reject it as a Type II error. Evidence of statistical significance, therefore, is not sufficient to support policy.

Morrison and Henkel (1970) have provided an interesting collection of articles dealing with the significance test controversy, several of which comment directly on the important distinction between practical and statistical significance. While none of the articles provided an answer for defining practical significance, it was pointed out that value judgments are at issue in defining statistical significance as well as in defining practical significance. In the case of statistical significance, however, the value question is resolved through convention when the investigation agrees on one or two levels of significance.

In that same collection, Gold (1969) indicated a second difficulty in defining the importance of an effect; the importance of an effect of a given size may vary with its location on a scale. Different utilities may be assigned to a fixed increment at different points on a scale. Even for interval scale data, a one unit effect may have different meaning depending upon its location along the scale continuum. The possibility of shifting importance was recognized long ago in another context when Dalton (1920) stated that increments in income have progressively less utility after the base income reaches a certain level.

A third problem with past attempts to define practical significance is that many size of effect indices are influenced by factors independent of the utility of an effect. Thus, single effects may produce widely varying values depending on factors such as population heterogeneity and amount of measurement error. These difficulties plague even such "scale free" estimates as measures of association and measures in standard deviation units.

Many suggest that reporting an index of association which is relatively insensitive to sample size will avoid the problems suggested by equating statistical and practical significance. Eta squared, epsilon squared, omega squared, and the Pearson correlation have all been used in an attempt to indicate the practical importance of observed relationships. A substantial body of literature has evolved surrounding the relative advantages and disadvantages of these indices (e.g., Cohen, 1969; Friedman, 1968; Hays, 1963; Kennedy, 1970). In practice, the small differences among indices are probably of little importance given the imprecision of the data from which they are calculated. Furthermore, the sampling fluctuation of the index is often ignored. Most advocates of measures of association first ask if the relationship is significantly different from zero. Once statistical significance has been observed, however, it is common practice to forget about sampling fluctuations and interpret the point estimate of association as a parameter. Thus, if the criterion for importance is 10% or more of the variance accounted for, a sample R^2 of .10 or larger is taken to meet the criterion.

Glass and Hakstian (1969) have been critical of all indices of assocation, at least for use in designs involving fixed effects. They express concern that researchers will be misled into interpreting a fixed effect as though it were random. To their concern it should be added that the measures of association are all functions which depend upon the heterogeneity of the population selected for investigation and the amount of measurement error in the variables. Neither heterogeneity nor measurement error is likely to covary with practical significance, however defined. Clearly, then, defining practical significance in terms of an index of association is potentially misleading.

Even if measures of association are useful for deciding what constitutes a strong relationship, this question remains: "How large must the index be to be practically significant?" It is difficult to decide how much variance explained is sufficient to have practical value (particularly if the independent variable is qualitative with more than two levels). In reference to estimates of aptitude by treatment interactions, Cronbach and Snow (1977) assert that a .40 difference between standardized regression coefficients "seems likely to be theoretically important," as is "a difference between transformed correlation coefficients of .424" (p. 56). While they provided no substantive rationale for their criterion, they did add the caveat that "costs and utilities could warrant specifying a greater or smaller effect size" (p. 56).

Others who distinguish practical significance from statistical significance have turned to expressing effects in standard deviation units. Criteria such as .5 or more standard deviations have been used when judging the importance of findings from evaluations (e.g., Westinghouse Learning Corporation, 1969). In addition, standard deviation units have nearly universal application in defining the size of effect to be detected in *a priori* power calculations (Brewer, 1972; Cohen, 1969; Subkoviak & Levin, 1977). But which standard deviation should be used? Should it be the square root of the error variance for testing the significance of an effect and so perhaps differ from hypothesis to hypothesis within a study? Should it be the standard deviation defined on individuals even when the unit of analysis is some aggregate of individuals? These and similar questions remain unanswered.

Despite the difficulties, most people concerned with conducting evaluations agree that an evaluator should decide for him/herself (or his/her client) what constitutes practical significance and design the evaluation and report the results accordingly (e.g., Boruch, 1977). Of the three procedures for defining practical significance just reviewed, measures of association provide the metric least sensitive to factors conceptually unrelated to the size of a program effect; in that sense, they seem best suited to the problem of defining practical significance.

Regardless of metric, however, what constitutes an important effect in an evaluation depends on value judgments which may be made in different ways by different parties.

The fourth problem with previous attempts at defining practical significance is poor reporting practice that makes it difficult - if not impossible - to reasonably assess the utility of a program effect without access to the original data. The lack of reported information about the compositions of an outcome measure forces the reader to accept the evaluator's values. Even worse, it may be that the evaluator is naive about the validity of his measures. In reporting results, therefore, it seems reasonable to strive to present sufficient information about the variables so that others might exercise their own values when interpreting the findings.

Unfortunately, all of the methods for defining practical significance considered thus far facilitate the practice of reporting program effects without providing information about the composition of the dependent variable. The extent of this reporting problem is indicated in Anderson's review of 130 articles (in the Journal of Education Psychology and the American Educational Research Journal) from June 1964 to February 1971 in which one or more homemade tests of reading comprehension were used:

> Most investigators reported nothing about their tests beyond such rudimentary information as the number of items and the response made. Several investigators did not hint that a test was used until the analysis of variance was described, at which point, the test was mentioned no more. One investigator characterized his test in a single sentence. "Criterion achievement was measured by the final achievement test." (1972, p. 165)

Dimensionality of Achievement Tests

Aside from the four problems discussed above, the common practice of thinking about the size of an effect in terms of an aggregate measure is, in itself, likely to be misleading. An achievement test generally assesses achievement in a number of content areas. Thus, identical aggregate scores on an achievement test do not necessarily reflect the same level of achievement across all content areas. In a mathematics program, for example, the same aggregate effect might be produced by a large gain in either computational skills or understanding of mathematical concepts. The values placed on each of these two areas may differ, however. Similarly, an effect in an area to which considerable program resources had been devoted would have different meaning than an effect in an area to which no resources had been committed.

Yet, achievement tests (or at least subtests) are constructed to be empirically unidimensional. For example, the mathematics subtests of concepts, computation, and applications on the Stanford Achievement Test (SAT) Intermediate Level I, Form A are reported to have internal consistency reliabilities of .87, .91, and .93, respectively, when given to beginning fifth graders. Evidence of internal consistency has been taken as evidence that all items measure a single trait; this brings into question the utility of identifying subsets of items (e.g., Goolsby, 1966). There are at least two reasons why evidence of a test's empirical unidimensionality may be misleading as to the utility of identifying subsets of items. The first reason stems from the definition of empirical unidimensionality; the second is a function of the ways in which unidimensionality is estimated.

The empirical definition of unidimensionality calls for a large first factor on the item intercorrelation matrix. Thus, empirical unidimensionality is a static concept specific to the time of test administration and the population of respondents. Consider a population of respondents and set of items that yield an item intercorrelation matrix with equal off-diagonal elements. Suppose half the items require division with remainder, half the items require multiplication of three-digit numbers, and the population of respondents is beginning fourth grade students. If experiencing an intervention were to uniformly reduce the difficulty of half the items -- for example, the intervention focused on multiplication of three-digit numbers and did not consider division -- the only effect on the item intercorrelation matrix would be to create a difficulty factor. The difficulty factor could be avoided by use of tetrachoric coefficients (Carroll, 1961). Yet, despite empirical unidimensionality (both prior to and after the intervention), there is clearly a useful distinction between the two subsets of items. It is of interest, therefore, to ask whether a test is unidimensional relative to an intervention, i.e., does an intervention affect all item difficulties equally? Searching for differential effects across items is analogous to searching for aptitude-by-treatment interactions (ATI's) and might be called the search for item-by-treatment interactions (ITI's).

Most test data, however, are not confined to individuals receiving a single intervention. In education, different students receive different educational experiences, and these experiences may have different effects across items. If a test is comprised of sets of items defined by concepts such that the effect of an intervention is constant within each set, and if the effects of interventions vary with less-than-perfect correlation across sets of items, the sets should be reflected in the pattern of item intercorrelations. This effect on item intercorrelations occurs because the intervention effects contribute to both the covariance and variance of items within a set but not to the covariance of items between sets. Since data from norm groups of standardized tests would seem to be a case in point, the fact that they are reported to be internally consistent still seems to challenge the importance of ITI's. The apparent unidimensionality of standardized tests, however, may only be evidence for the existence of a strong single dimension, not for the absence of content factors. If, in the situation just described, items were arranged by concepts, the item intercorrelation matrix would be a super matrix with submatrices on the main diagonal representing within-concept correlations. If ITI's are present, the diagonal submatrices will have higher correlations than the off-diagonal submatrices, thus yielding a factor for each concept. (The off-diagonal submatrices could all be equal except for the effects of varying item difficulties.) The off-diagonal submatrices will also tend to have positive correlations, however, because of individual differences in aptitude and the likelihood of positive correlations between intervention effects across sets of items, due to the hierarchical nature of most subject matter. The positive off-diagonal submatrices contribute to a single common factor. Using the Spearman-Brown prophecy formula, the more concepts included, the stronger the general factor. Furthermore, the fewer items per concept, the less clearly defined the second order concept factors. Thus, evidence of an internally consistent test should not be misconstrued as indicating the uselessness in searching for ITI's in evaluations using that test.

When defining practical significance, then, concern for describing test content validity for an intervention and the possibility of ITI's are both important. Those who have been interested in the possibility of item-by-treatment interactions have, for the most part, been relatively unconcerned about constructing achievement tests to reflect the content of interventions (Mandeville, 1972; Moonan, 1955). Recently, the most visible interest in ITI's has been in the area of detecting bias in existing tests (e.g., Cleary, 1968; Jensen, 1976). In this context, few interactions have been found, though Gupta (1969) reported a sex by item interaction for Step-Math 2A. Likewise, those who have called for careful test construction to reflect the content of interventions have not seemed particularly concerned with detecting item-by-treatment interactions (with the exception, maybe, of Hastings, 1966).

As usual, actual practice has lagged well behind recommended practice. The call for program-valid achievement testing in evaluation (e.g., Bloom, Hastings, & Madaus, 1971; Nunnally & Wilson, 1975; Shoemaker, 1975) remains largely ignored. The goals of educational interventions are typically vague, making difficult the selection of content-valid dependent variables. Even when program implementation is given explicit attention, the content goals of the program are inadequately considered. In the discussion of curriculum change by Fullan and Pomfret (1977, p. 361), for example, there was little analysis of content goals (just one of five dimensions considered).

In a few notable exceptions, evaluations have included carefully constructed program-valid achievement measures. Hively, Maxwell, Rabehl, Sension, and Lundin (1973) provided a detailed account of their domain-referenced evaluation of the MINNEMAST Project -- a modern mathematics and science curriculum for elementary school. In that evaluation, the authors make extensive use of item forms to represent domains. The chapters in Part II of Bloom et al. (1971) also provide illustrations of content analyses on which program-valid achievement tests might be constructed. Finally, objectives-referenced test systems which are commercially available make it possible to construct tailor-made achievement tests for a few subject matter areas (e.g., SOBAR Field Manual I, 1972). In more general terms, the concept of "universe-defined" (Osburn, 1968; Hively, Patterson, Page, 1968) or domain-referenced (Hively et al., 1973) tests -- where content is made clear through rules of construction -- holds promise for meeting both ITI and reporting concerns.

An Illustration Using Standardized Test Content

The mainstream of educational evaluation continues to rely on standardized tests of achievement as opposed to tests constructed to fit a particular need. These standardized tests are designed to evaluate individual student differences on a rather amorphous national curriculum. The market is dominated by the Stanford Achievement Tests, Iowa Tests of Basic Skills, California Test of Basic Skills, and the Metropolitan Achievement Tests. The methods for selecting one test over another in any particular situation are not documented and remain unclear. For the most part, it appears that these tests are used interchangeably, with frequent references made to the high intercorrelation among corresponding subtests. As noted previously, however, evidence of intercorrelation (internal consistency) can be misleading in terms of interchanging tests.

Despite frequent attacks on the use of standardized tests for program evaluation (Airasian & Madaus, 1976; Cox & Sterrett, 1970; Shoemaker, 1975), the criticism remains on an abstract level. Far too few careful analyses of standardized test content have been completed --

analyses which could demonstrate the link between test and program to be evaluated and on which searches for ITI's could be based. Jenkins and Pany (1976) analyzed five standardized reading achievement tests of word recognition at grades one and two and seven commercial reading series. After observing differences among both tests and curricula and an interaction between the two they concluded, "It appears doubtful that conventional achievement tests can serve as unbiased estimates of a curriculum's effect, at least at ea:ly grade levels" (p. 12). While some questions can be raised about the construction of their word lists, the possibility of item-by-treatment interactions and misleading aggregate effects is supported. An analysis of standardized tests and curricula for reading comprehension by Armbruster, Steven, and Rosenshine (1977) also yielded differences between skills taught and skills tested. The categories used for this analysis, however, seemed to be more a function of the way in which test questions were asked than the content of the text to be read. These categories did not isolate vocabulary, sentence construction, sentence length, or complexity of concepts, all of which are known to affect comprehension.

Developing a Taxonomy to Measure Content

As part of our work on teacher decisions about the content of instruction, it was necessary to develop a method for describing the variety of content taught in fourth grade mathematics. On the assumption that the items in standardized achievement tests of mathematics at the fourth grade level should reflect that variety, an iterative process of analysis and classification of items on the

Stanford Achievement Test was conducted. The result of that content analysis and classification was the development of a taxonomy (shown in Figure 1 at the end of this paper). The taxonomy is a vehicle for illustrating test content analyses and their usefulness for defining practical significance and investigating ITI's.

The taxonomy provides an explicit description of the match between the content of a program and the content of a test used to evaluate that program. If an intervention addresses the content implied by a subset of the taxonomy cells, then those cells identify item domains that should be included on a test of effects. If there are hypotheses about transfer or concern for unanticipated negative effects, item domains identified by other cells in the taxonomy might also be included. Again, the taxonomy would help to make such interests explicit and so increase the precision with which they are addressed in the evaluation.

Reporting the distribution of items across cells in the taxonomy should also be an effective and efficient way to provide information necessary to support value judgments about size of effect. Further, the taxonomy should be useful in searching for item-by-treatment interactions which, if present, make interpretations of aggregate effects difficult. To facilitate the estimation of such interactions, each item domain should be represented by a set of items. The number of items in a set need not be as large as suggested for reliability in individual assessment, since item-by-treatment interactions are defined on group means rather than individual scores. The standard error of a group mean is directly related to groupsize, and group size counters the impact of low reliability due to few items defining individual scores.

Our taxonomy is defined by the intersections of three factors: (1) mode of presentation (3 levels), (2) nature of the material (13 levels), and (3) operations (12 levels). The intersections of these three factors results in 468 cells. In some respects the taxonomy may appear to be unrealistically detailed, while in others it may appear to gloss over important distinctions. Our goal was to provide a level of detail sufficient for describing teacher decisions about content of instruction. Clearly, the extent to which there are similarities or differences in content between a program and a test is a function of the detail level of the description provided. It is important to assume that our taxonomy is at a level of detail such that instruction can be directed to some cells and not others. The taxonomy has been reviewed by several teachers involved in mathematics instruction in the elementary grades, and those reviews were generally supportive of the assumption.

The first taxonomy factor -- Mode of Presentation -- distinguishes between items which present essential information in graphs, figures, tables, and those which do not. For those items which do not present essential information in graphs, figures, or tables, a further distinction is made between items which specify the operation required for solution and those which do not (e.g., the typical story problem).

The second factor -- Nature of the Material -- has several levels which are not mutually exclusive but which are ordered in complexity. In using the taxonomy, an item is classified at the highest appropriate level of complexity. In using the taxonomy, an item is classified at the highest appropriate level of complexit . In ascending order of complexity, the levels are: (1) single digits, (2) single and multiple digits, (3) multiple digits, (4) single fractions, (5) multiple fractions, (6) decimals, (7) percents, (8) alternative number systems (e.g., Roman numerals, clock arithmetic), (9) place value, (10) number sentences, (11) algeoraic sentences, (unknown quantities not isolated by an equal sign), (12) conversion from one scale of measurement to another, and (13) geometric figures.

The third factor -- Operations -- also includes levels which are not mutually exclusive and again items are classified at the highest level of complexity appropriate. Starting with the least complex, the levels are (1) add, (2) subtract without borrowing, (3) subtract with borrowing, (4) add or subtract fractions without a common denominator, (5) multiply, (6) divide without remainder, (7) divide with remainder, (8) combination (more than one of the basic arithmetic operations), (9) grouping (use of parentheses), (10) identify equivalents (e.g., select the figure with a fourth of its area shaded), (11) identify rule (e.g., number series problems), (12) identify terms (essentially vocabulary).

Classifying the Content of Standardized Tests

The popularity of standardized tests for use in program evaluation makes knowledge of their content important. To that end and to further illustrate the possibility of treatment-by-item interactions on presumably unidimensional tests, our taxonomy has been used to classify fourth grade mathematics content on the four most widely used standardized tests: the Stanford Achievement Test (SAT), the Iowa Test of Basic Skills (lowa), the Metropolitan Achievement Test (MAT), and the California Test of Basic Skills (CTBS)*

The items in the mathematics subtests at the fourth grade level for all four standardized test batteries were independently classified by three of the authors. Assuming that agreement between two out of three raters makes an item classifiable, 98% of all the items could be classified. Inter-rater reliabilities are reported in Table 1 by test battery, subtest, and dimension of the taxonomy. Only those items on the Study Skills subtests pertaining to mathematics were classified. The cell entries represent percent of possible pairs of raters agreeing; for each item, all three raters agreeing counted as three out of three possible pairs and two raters agreeing counted as one out of three. Entries in the columns labeled C of Table 1 represent agreement as to the exact cell in the matrix. As might be expected, the computation subtests were described with the greatest accuracy -- 90% or more agreement at the exact cell level. The concepts subtests contained items most difficult to describe using the taxonomy, with exact cell agreements near 60%. The four tests were nearly equal in

* Iowa Tests of Basic Skills (1971); Level 10; Tests M-1, M-2, and appropriate items on W-2.

Metropolitan Achievement Tests (1970); Elementary Level; Tests 5, 6, & 7.

Stanford Achievement Tests (1973), Primary Level III (3rd Grade), Intermediate Level I (4th grade), and Intermediate Level II (5th grade); Tests 4, 5, & 6.

California Tests of Basic Skills (1968); Level II; Tests 6 & 7.

the extent to which they could be accurately described; the IOWA was a slight exception, particularly since it did not contain a subtest devoted to computation.

The percentages of items in each test battery at all levels of every dimension are presented in Table 2. For these data, an item was classified by reviewing the independent decisions of the three raters and resolving disagreements to the raters' mutual satisfaction. The reliabilities reported in Table 1 represent, therefore, a strong lower bound to that for data in Table 2. In one sense, the data in Table 2 may be misleading in that the percentages reported for the marginals of the taxonomy could be in agreement and still there would be no overlap in classification of items from the different tests at the cell level. To the extent that differences occur on the marginals, however, the tests do differ in content and at a rather low level of detail.

For mode of presentation, three of the four tests appeared quite similar, but the Iowa had a substantially larger proportion of items where essential information was presented in the form of graphs, figures, and tables. This difference was due, in part, to the absence of a computation subtest on the Iowa but not entirely, since the raw number of such items was considereably greater as well. With the exception of the Iowa, roughly 20% of the items involved graphs, figures, or tables, and a little less than a third of the items required the respondent to figure out the necessary operation (for the most part, story problems).

On the nature of materials, there were more similarities than differences among the four test batteries. Still, some important differences existed. For example, the subtotals for the three levels involving whole numbers varied from 39 - 66% with the SAT having the highest percentage. Other frequently-represented levels were algebraic sentences, at roughly 10%, and essential units of measurement, which ranged from a low of 7% on the SAT to a high of 15% on the MAT. Percents, alternative number systems, and geometric figures were not emphasized on any of the tests. (To provide a better understanding of these differences it must be pointed out that for the SAT, a percent is about .9 of an item. Further, an item is equivalent to approximately .2 of a grade equivalent near the middle of the norm distribution on the SAT math subtests.

On the operations factor the tests were quite similar in the percentages of items involving subtract without borrowing (6% - 8%), add or subtract fractions without a common denominator (0% - 2%), divide with remainder (1%), and combinations (6% - 8%). For the remaining levels there were modest to strong differences among the tests. The MAT, for example, had 21% addition items, which was about eight percentage points more than the other tests. The Iowa had at least five percentage points fewer multiplication items than did the other tests. Grouping was tested by the SAT but not at all by either the MAT or the CTBS.

To provide some sense of how the tests varied in content across grade levels, the third and fifth grade levels on the SAT were also analyzed. The results are reported in Table 3 and are based on resolution of any disagreements between two independent raters, both of whom were also raters for the data in Table 2. The percentage distributions of items across mode of presentation levels remained nearly identical from the third grade level to the fifth grade level: Under nature of material, the percentages for items classified as single digits and place value decreased while the percentages increased for items classified as fractions, decimals, and percents. Surprisingly, the percent of items classified as algebraic sentences held quite constant at approximately 10%.

The data in Tables 2 and 3 represent descriptions of mathematics content across all subtests using only the marginals of the taxonomy. The data in Figure 2 represent item distributions across the cells of the taxonomy for the Concepts subtest of the SAT and the MAT. The X's in the upper half of each cell represent items on the SAT, and the O's in the bottom half of each cell represent items on the MAT. Across the two subtests, items fell into 47 different cells. Of those 47 cells, however, only 7 - 15% were common to both tests. While the cell level analysis was most dramatic, sizable differences were reflected in comparisons on the marginals. For example, 12% of the MAT items were classified Operation Not Specified, while there were no such items on the SAT. Twenty-three percent of the MAT items involved essential units of measure while only 6% of the SAT items were classified at that level. The SAT had larger percentages of items classified as grouping, (6% compared to 0%), identify rule (19% compared to 7%) and and identify term (22% compared to 12%).

It is clear from these content analyses that a total score on any one of the subtests considered represents an aggregate across content areas; these aggregates might well vary in their sensitivity to any given mathematics intervention. It seems reasonable that a similar content analysis of a program to be evaluated would yield hypotheses about potential item-by-treatment interactions. Furthermore, there is sufficient variance in content across tests so that some are more likely relevant than others for assessing the effects of a given intervention. Finally, a taxonomy is efficient method for communicating to those reading evaluation reports the information prerequisite to deciding what constitutes practical significance.

Summary

Defining practical significance in program evaluations is a difficult measurement problem which can only be solved by an intimate familiarity with the measures on which effects are estimated and their content relationship with the goals of the program being evaluted. Past attempts to provide general solutions to the size of effect problem have relied on standardized indices which can be estimated and reported without any knowledge of what was measured. For this reason, these efforts are viewed here as steps in the wrong direction. Instead, what is called for is a procedure whereby the content goals of the program, the content implied by a test, and the interrelationship between the two are made explicit. The procedure should investigate treatment-by-item interactions and at the same time, describe the measures used so that persons other than the evaluator can reach their own decisions about practical significance. The taxonomy of fourth grade mathematics illustrated the possibility of obtaining better knowledge about variables on which program effects are estimated. A detailed description of the mathematics sections of the four major standardized tests -- obtained with the taxonomy -- indicated rather substantial differences in content tested. From the analyses, it was clear that the standardized tests are not well suited to the task of estimating item domain by treatment interactions, as most cells in the taxonomy were represented by only one or two items.

References

Airasian, P.W., & Madaus, G.F. A study of the sensitivity of school and program effectiveness measures. A report submitted to the Carnegie Corporation of New York, 1976.

Anderson, R.C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Armbruster, B.B., Steven, R.O., & Rosenshine, B. Analyzing content coverage and emphasis: A study of three curricula and two tests (Tech. Rep. No. 26). Center for the Study of Reading, University of Illinois, Urbana-Champaign, 1977.

Bloom, B.S., Hastings, T.M., & Madaus, G.F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Boruch, R.F., & Gomez, H. Measurement in impact evaluation. Some humble theory on sensitivity, bias, and use. In R. Perloff, & E. Perloff (Eds.), Professional Psychology, 1977.

Brewer, J.R. On the power of statistical tests. American Educational Research Journal, 1972, 9, 391-401.

Carroll, J.B. The nature of data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Cleary, T.A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Cox, R.C., & Sterrett, B.G. A model for increasing the meaning of standardized test scores. Journal of Educational Measurement, 1970, 7 (4), 227-228.

Cronbach, L.J., & Snow, R.E. Aptitude and instructional methods. Irvington, Mass: Irvington Publisher, 1977.

Dalton, H. The measurement of inequality in unions. Economic Journal, 1920, 30, 348-361.

Friedman, H. Magnitude of experimental effect and a table for its rapid estimation. Psychological Bulletin, 1968, 70, 245-251.

Fullan, M., & Pomfret, A. Research on curriculum and instruction implementation. Review of Educational Research, 1972, 47, 335-397.

Glass, G.V., & Hakstian, A.R. Measures of association in comparative experiments: Their development and interpretation. American Educational Journal, 1969, 6, 403-414.

Gold, D. Statistical tests and substantive significance. The American Sociologist, 1969, 4, 42-49 (as reprinted, Morrison, D.E. and Henkel, R.E. no. 21).

Goolsby, T.M. Differentiating between measures of different outcomes in the social studies. Journal of Educational Measurement, 1966, 3, 219-222.

Gupta, R.K. Treatment comparisons: Item responses in multi-factor repeated measures design. Journal of Experimental Education, 1969, 37, 26-29.

Hastings, J.T. Curriculum evaluation: The why of the outcomes. Journal of Educational Measurement, 1966, 3, 27-32.

Hays, W.L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project (CSE Monograph Series in Evaluation, No. 1). Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W., Patterson, H.L., & Page, S.A. "Universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Jenkins, J.R. & Pany, D. Curriculum biases in reading achievement tests (Tech. Rep. No. 16). Urbana, Illinois: University of Illinois, Center for the Study of Reading, November, 1976.

Jensen, A. Test bias and construct validity. Phi Delta Kappan, December, 1976, 58(4), 340-346.

Kempthorne, O., & Folks, L. Probability statistics and data analysis. Ames, Iowa: Iowa State University Press, 1971.

Kennedy, J.J. The eta coefficient in ANOVA designs. Educational and Psychological Measurement, 1970, 30, 885-889.

Mandeville, G.K. A new look at treatment differences. American Educational Research Journal, 1972, 9(2), 311-321.

Moonan, W.J. Simultaneous examination and method analysis of variance algebra. Journal of Experimental Education, 1955, 23, 253-257.

Morrison, D.E., & Henkel, R.E., (Eds.). The significance test controversy. Chicago: Aldine Publishing Company, 1970.

Nunnally, J.C., & Wilson, W.H. Method and theory for developing measures in evaluation research. In E.L. Struening & M.Guttentag (Eds.) Handbook of Evaluation Research. Beverly Hills: Sage Publication, Inc. 1975.

Osburn, H.G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Shoemaker, D.M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

SOBAR Field Manual I. Prepared by the staff of the Program for Research on Objectives-Based Evaluation. Los Angeles: Center for the Study of Evaluation, University of California, 1972 (field test edition).

Subkoviak, M.J., & Levin, J.R. Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 1977, 14, 47-52.

Westinghouse Learning Corporation. The impact of Headstart: An evaluation of the effects of Headstart on children's cognitive and affective development. Blodensburg, Maryland: Westinghouse Learning Corporation and Ohio University, U.S. Office of Economic Opportunity, Contract No. B84-4536, 1969.

Classification of ____________________

By ____________________ Date ____________

IRT/OUTCOMES
10/11/77

MODE OF PRESENTATION

Nature of the Material/Operation		Graphs, Figures, Tables or Physical Objects												Operation(s) Specified												Operation(s) Not Specified (Story Problems)											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Whole Numbers	single digits																																				
	single digit and multiple digit																																				
	multiple digits																																				
Fractions	single																																				
	multiple																																				
Decimals																																					
Percents																																					
[a] Alternate Number Systems																																					
Place Value																																					
Sentences	Number																																				
	[b] Algebraic																																				
[a] Essential Units of Measurement																																					
[a] Geometric Figures																																					
other																																					

Operations
1. Add
2. Subtract w/o borrowing
3. Subtract with borrowing
4. Add or Subtract Fractions
5. Multiply
6. Divide w/o remainder
7. Divide with remainder
8. Combination
9. Grouping
10. Identify Equivalents
11. Identify Rule (Order)
12. Identify Terms *

* Be sure to identify specifics on attached page.

Figure 1

Classification of X - SAT + O - MAT

By ______________________ Date Concepts

IRT/OUTCOMES
10/11/77

MODE OF PRESENTATION

Nature of the Material / Operation		Graphs, Figures, Tables or Physical Objects												Operation(s) Specified												Operation(s) Not Specified (Story Problems)											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Whole Numbers	single digits										oo	o						oo				X	X	X	X												
	single digit and multiple digit						o						X	o				oo				X		XX	X												
	multiple digits										o					oo								XX o						o							
Fractions	single										X oo							X						X													
	multiple																																				
Decimals																																					
Percents											X																										
[a]Alternate Number Systems											o			o									X														
Place Value											X												XXXX ooo	o	X												
Sentences	Number										XX o																										
	[b]Algebraic													X				X			o		oo		X o												
[a]Essential Units of Measurement		o									X o				X			o			o		o					o			o				oo		
[a]Geometric Figures							X						XX ooo																								
other																									o												

Operations
1. Add
2. Subtract w/o borrowing
3. Subtract with borrowing
4. Add or Subtract Fractions
5. Multiply
6. Divide w/o remainder
7. Divide with remainder
8. Combination
9. Grouping
10. Identify Equivalents
11. Identify Rule (Order)
12. Identify Terms *.

* Be sure to identify specifics on attached page.

2: Matrix Distribution of Items on Concepts Subtests of the SAT and MAT.

Inter-rater Agreement *

	Computation				Concepts				Problem Solving				Study Skills				
	D_1	D_2	D_3	C	D_1	D_2	D_3	C	I_1	D_2	D_3	C	D_1	D_2	D_3	C	cell average
SAT	100	98	93	92	94	79	83	63	97	86	93	78	100	83	83	83	79
MAT	100	92	98	90	85	88	73	64	100	82	94	78					77.3
IOWA					81	84	70	54	83	90	98	78	100	82	58	50	60.7
CTBS	100	99	99	99	91	87	89	73	100	80	93	75	100	82	71	69	79.0
cell average				93.7				63.5				77.3				67.3	

* Entries are percent of possible pairs of three raters agreeing

D_1 mode of presentation

D_2 nature of material

D_3 operations

C cell of the matrix

Table 1

Table 2

ITEM DISTRIBUTIONS FOR EACH FACTOR ACROSS TESTS*
FOURTH GRADE LEVEL

	IOWA	MAT	SAT	CTBS
I. Mode of Presentation				
- graphs, figures, tables, etc.	43	15	21	19
- operation(s) specified	29	52	53	59
- operation(s) not specified	29	32	27	22
	(N=84)	(N=115)	(N=116)	(N=113)
II. Nature of Material				
- single digits	12	15	20	2
- single and multiple digits	12	20	23	18
- multiple digits	24	19	22	19
- total -- whole numbers	47	54	66	39
- single fraction	6	4	5	7
- multiple fractions	5	3	-	7
- decimals	6	5	4	10
- percents	-	-	1	6
- alter. number systems	-	2	1	-
- place value	8	3	5	4
- number sentences	6	1	2	-
- algebraic sentences	8	10	8	12
- essen. units meas.	10	15	7	11
- geometric figures	2	3	3	2
** - other	1	1	-	2
III. Operations				
- add	12	21	13	14
- subtract w/o borrowing	8	8	6	8
- subtract with borrowing	11	11	6	5
- add or subtract fractions w/o common denominator	1	-	-	2
- multiply	11	19	16	17
- divide w/o remainder	6	9	15	14
- divide with remainder	1	1	1	1
- combination	8	6	7	7
- grouping	2	-	5	-
- identify equivalents	20	18	16	15
- identify rule (order)	11	3	9	12
- identify terms	8	5	6	4

* entries are percents

** This does not represent a level of Nature of Material, but rather the percent of items on each test that could not be fit into the taxonomy.

Table 3

ITEM DISTRIBUTIONS FOR EACH FACTOR ACROSS GRADES*
STANFORD ACHIEVEMENT TEST

	3rd	4th	5th
I. Mode of Presentation			
- graphs, figures, tables, etc.	18	21	18
- operation(s) specified	59	53	55
- operation(s) not specified	23	27	28
	(N=96)	(N=116)	(N=120)
II. Nature of Material			
- single digits	26	20	13
- single and multiple digits	22	23	21
- multiple digits	13	22	15
- total -- whole numbers	60	66	49
- single fraction)	4	5	8
- multiple fractions	-	-	7
- decimals	-	4	3
- percents	-	1	3
- alter. number systems	2	1	3
- place value	9	5	5
- number sentences	2	2	2
- algebraic sentences	10	8	10
- essen. units meas.	[illegible]	7	6
- geometric figures	2	3	4
- other	3	-	2
III. Operations			
- add	13.	13	10
- subtract w/o borrowing	10	6	2
- subtract with borrowing	9	6	8
- add or subtract fractions w/o common denominator	-	-	1
- multiply	17	16	13
- divide w/o remainder	10	15	8
- divide with remainder	-	1	6
- combination	10	7	13
- grouping	1	5	8
- identify equivalents	18	16	20
- identify rule (order)	7	9	2
- identify terms	4	6	12

* entries are percents