DOCUMENT RESUME

ED 199 299 TM 810 339

AUTHOR Anderson, Ronald E.; And Others
TITLE Methodological Considerations in the Development of Indicators of Achievement in the NAEP Data.
INSTITUTION Minnesota Univ., Minneapolis. Center for Social Research.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE [81]
GRANT NSF-SED-79-17259
NOTE 29p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Rating; *Data Analysis; Educational Assessment; Elementary Secondary Education; National Surveys; *Research Design; *Research Methodology; *Research Problems
IDENTIFIERS *Data Interpretation; *National Assessment of Educational Progress

ABSTRACT

The organization of data at the National Assessment of Educational Progress (NAEP) is undergoing a significant transition from a system designed only for national assessment purposes to one designed both for assessment and a variety of academic research interests. The advent of NAEP public-use data files opens up many possibilities for those who have the skills, time, and resources to do secondary analysis. An analysis of the mathematics test items is presented which demonstrates alternate procedures for developing indicators of mathematics achievement. This analysis demonstrates that the NAEP item subsets will not always meet conventional psychometric criteria. This failure to meet standard achievement test criteria does not mean that secondary analysis of the data is unwise. It does imply, however, that interpretation of findings, especially those using subtests, must be made cautiously. Limitations of the methodology must be acknowledged. Conventional achievement testing is not item-centered like assessment testing. The measurement priority of assessment is stability across multiple testings, not relative comparisons among persons. Consequently, standards of item discrimination and construct validity have obviously less import. Of far greater importance for assessments are standards of face validity, content validity, internal consistency and the application of rigorous data analysis techniques. (Author/RL)

***********************************************************************
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
***********************************************************************

ERIC

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED199299

# METHODOLOGICAL CONSIDERATIONS IN THE DEVELOPMENT OF INDICATORS OF ACHIEVEMENT IN THE NAEP DATA

Ronald E. Anderson
Wayne W. Welch
Linda J. Harris

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Anderson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Minnesota Center for Social Research
2122 Riverside Avenue
University of Minnesota
Minneapolis, Minnesota 55454

This article was prepared with the support of National Science Foundation Grant No. SED 79-17259. Any opinions, findings, conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation or anyone else.

TM 810 339



ERIC

METHODOLOGICAL CONSIDERATIONS IN THE DEVELOPMENT OF INDICATORS OF ACHIEVEMENT IN THE NAEP DATA

ABSTRACT

The organization of data at the National Assessment of Educational Progress (NAEP) is undergoing a significant transition from a system designed only for national assessment purposes to one designed both for assessment and a variety of academic research interests. Those acquiring NAEP data for secondary analysis must be aware of the organization of the data due to its historical emphasis upon assessment. Researchers should be cognizant of the unique features of NAEP data as well as potential sources of error in using the data. In its ongoing analysis of the thirty files from 1977-78 assessment of mathematics, the research group at the University of Minnesota has identified and summarized some of these positive and negative features. In this report we review alternate procedures for evaluating measures of achievement at the level of the individual data package. Some findings are presented and their methodological implications discussed.

## INTRODUCTION

The National Assessment of Education Progress (NAEP) and their findings have been extersively discussed in both the popular press and in the academic literature. As described by Wright, Larsson, and Remlow (1981), the management of NAEP data has recently en reorganized in order to facilitate a wide variety of secondary analyses. Specifically, NAEP data are now disseminated as a series of public-use data tapes where each data file contains the results of one assess.. booklet (package) for one age group during one year or cycle of asse..sment. As discussed in an earlier paper (Anderson, Welch, Harris, 1980), this organization of the data has some serious implications for the secondary analyst. In this paper we are concerned only with the problems of identifying and developing indicators of achievement.

Tyler (1970), Womer (1973) and others have elaborated on the differences between assessment and testing, especially standardized achievement testing. The main difference is that standardized testing seeks to compare and rank student scores whereas assessment ignores individual scores and reports on the state of the system and its major social groups. Millman (1978) clarifies this distinction further by contrasting three major types of assessment models: (1) item-centered, where many test items are used to assess a large content domain, (2) objective-centered, where each objective is represented by a number, e.g., 5, of test items, and (3) subtest-centered, where the focus is upon several subtests, each of which is measured with a moderate number of items. NAEP has always been item-centered in its approach, but the new public- use data system fosters a subtest-centered approach. With several hundred test exercises (items), NAEP assesses a content domain



ERIC

very extensively but specific subdomains and their associated subtests may not be thoroughly covered. This problem is magnified when the analysis becomes package-centered as with the public-use data tapes. Not only may some subdomains be very underrepresented, but the overall domain may not be well represented by the limited set of items within a single booklet/package.

It is quite reasonable to expect many secondary data analysts to be subtest-centered in their analysis interests. For instance, a researcher might want to identify correlates of "spatial reasoning". Spatial reasoning is a subset of the NAEP "shapes, sizes, and relationships" subdomain, which is a subset of the mathematics domain. NAEP, in their data analysis system, can analyze the relationship between a subtest and standard reporting variables by aggregating across all packages. This is much more difficult for a secondary analyst of public- use tapes; separate processing of each of ten to thirty files is necessary. Further, many interesting analyses are not possible this way because many independent as well as dependent variables are included in only one or two packages. For many potentially interesting analyses, the secondary analyst must settle for analysis of one and only one package. Before proceeding with such analysis, the methodologically cautious researcher must check for the validity and reliability of the package's specific indicators. Thus the secondary analyst can not, as does NAEP, depend upon the large bank of test items utilized in an item-centered assessment. At the package-specific level of analysis, where only a small portion of the content domain is included, the analyst is obligated to substantiate the psychometric soundness of the measures available. One exception to this is the use of single achievement items for analytic purposes. For instance, in error analysis one might look at the frequency of occurrence of specific incorrect responses to a specific item stem.

In the past, NAEP has deemphasized or avoided psychometric techniques such as item analysis, factor analysis and the selection of items on the basis of discrimination. In addition, NAEP generally does not complete estimates of reliability, however a dissertation by Mullis (1978) on the 1976 NAEP supplementary assessments of mathematics and political knowledge, does report reliabilities for achievement exercises. The reliabilities for mathematics averaged 0.91 but for political knowledge averaged 0.71 in Mullis' analysis.

Since our analysis and most future secondary analysis of NAEP data will necessarily be data package oriented, reliability and validity analyses are important for the refinement of indicators. At the data package level one is limited to a relatively small set of achievement items, especially if one needs scale or subtest scores to examine specific subdomains. Since the number of items per content area varies considerably across data packages, the content validity and reliability of indicators must be evaluated separately for each data set.

In recommending psychometric analyses of NAEP items at the package level, we are not implicitly criticizing NAEP methodology. The primary goal of NAEP is to measure changes in educational performance and to report these to the nation for purposes of policy formulation. Consequently, there has been little need to compute total achievement scores. In fact, until recently NAEP has avoided the use of the term "achievement" in its various reports of progress or performance of the nation. In secondary analysis the goals may change, thus the methods must change to meet the corresponding requirements that insure the highest quality of an analysis.

## EVALUATING PACKAGE-SPECIFIC INDICATORS OF ACHIEVEMENT

On the basis of our analysis of major sources of error in NAEP data (Anderson, Welch, Harris, 1980) and using the recommendations of the conventional wisdom in educational testing (cf. Lord and Novick, 1968; Guilford, 1954; Nunnally, 1978), a set of procedures have been outlined. These procedures are listed in Figure 1 and are designed to guide the evaluation of NAEP achievement indicators.

We applied these procedures to three data packages from the 1977-78 NAEP mathematics assessment of 17-year- olds. While the next section is limited to this mathematics data, the problems and procedures are relevant to other content areas as well.

### (1) CONTENT VALIDITY

The planning for the 1977-78 NAEP mathematics assessment was guided by a two dimensional classification of the content domain (Figure 2). One dimension includes five subject matter content areas and four cognitive process levels. These categories are defined in the Mathematics Objectives book (National Assessment of Educational Progress, 1978) as follows:

#### (A.) Numbers and Numeration

This category contains the largest number of exercises because of its importance in the curriculum. Exercises deal with the way numbers are used, processed or written. Knowledge and understanding of numeration and number concepts are assessed for whole numbers, fractions, decimals, integers and percents, with considerable emphasis placed on operations. Number properties and order relations are also included. Problem-solving exercises include routine number problems, nonroutine problems and consumer problems. Nonroutine problems are exercises not normally encountered in the curriculum, but understandable to the age group. Consumer problems deal primarily with the uses of mathematics in commercial situations (for example, buying and selling, interpreting graphs and saving money) and are emphasized more at the 17-year-old level than at the two younger age levels.

(B.) Variables and Relationships

The use of variables and relationships corresponds to an important part of the school mathematics curriculum. The exercises for this content category deal with facts, definitions and symbols of algebra; the use of variables in equations and inequalities, the use of variables to represent elements of a number system; and exponential and trigonometric functions. There are very few exercises appropriate for 9-year-olds in this category, and only a few topics are appropriate for 13-year-olds. Many more are appropriate at the 17-year-old level.

(C.) Shape, Size and Position

The exercises in this content category measure objectives related to school geometry. The emphasis in the assessment is not on geometry as a formal deductive system. The exercises concern plane and solid shapes, congruence, similarity, properties of triangles, properties of quadrilaterals, constructions, sections of solids, other basic theorems and relationships, rotations and symmetry.

(D.) Measurement

A portion e assessment is devoted to measurement, reflecting increased emp on measurement in the school curriculum. The exercises cover appropriate units; equivalence relations; instrument reading; length, weight, capacity, time and temperature; perimeter, area and volume; non-standard units; and precision and interpolation. A substantial number of the measurement exercises require the use of metric units.

(E.) Other Topics

Other mathematical content topics included in this assessment at all age levels are probability and statistics; graphs, tables and charts; and logic. Special assessment exercises and procedures have been developed to assess attitudes related to mathematics, computer literacy and the use of the hand calculator.

The four cognitive processes identified in the objectives plan were:

(I.) Mathematical Knowledge

Mathematical knowledge refers to the recall and recognition of mathematical ideas expressed in words, symbols or figures. Mathematical knowledge relies, for the most part, on memory processes. It does not ordinarily require other more complex mental processes.

Exercises that assess mathematical knowledge require that a person recall or recognize one or more items of information. An example of an exercise involving recall would be one that asks for a multiplication fact such as the product of five and two.

(II.) Mathematical Skill

Mathematical skill refers to the routine manipulation of mathematical ideas. Mathematical skill relies on algorithmic processes. An algorithm is a standard procedure that always leads to an answer. Mathematical skill requires the recollection of how to use the algorithm.

(III.) Mathematical Understanding

Mathematical understanding refers to the explanation and interpretation of mathematical knowledge. Mathematical understanding relies primarily on translation processes. The mathematical knowledge can be expressed in words, symbols or figures; and the translation may be within or between any of these modes of expression. Mathematical understanding involves memory processes as well as processes of associating one item of knowledge with another.

(IV.) Mathematical Application

Mathematical application refers to the use of mathematical knowledge, skill and understanding. Mathematical application relies on memory, algorithmic, translation and judgment processes.

Exercises that assess mathematical application require a sequence of processes that relate to the formulation, solution and interpretation of problems. The processes may include recalling and recoding knowledge, selecting and carrying out algorithms, making and testing conjectures and evaluating arguments and proportion, or it might require the demonstration that two geometric figures are congruent.

The foregoing taxonomy of goals for mathematics education is basically consistent with others who have attempted to classify the goals of mathematics education. Begle (1979) reviews some of these discussions and elaborates upon the classification developed for the National Longitudinal Study of Mathematics. Their categorization of objectives by cognitive level and content is quite similar to that of NAEP's second assessment (Figure 2). In planning for the 1977-1978 assessment NAEP developed a "blueprint" defining the relative emphasis of each category of objectives for each age group (Figure 3). Each cell of the blueprint tables specified the recommended number of exercises for the corresponding category or subdomain for a particular age group. Although the blueprint was used as a guide for the design of the tests, the actual distribution of exercises departs somewhat, as can be seen in Table 1, which gives the booklet distributions for age 17. The proportion of items for a specific category varies considerably across booklets. For instance, the blueprint specified that 24% of the items should be "understanding" items but one package contained no such items and the highest contained 23% with an average across packages of 16%. The main reason for the overall departure from the blueprint was that items at the "understanding" level presumably required longer than average completion times. NAEP weights the items by timing factors before determining how many items should be selected for a given category.

Except for specific circumstances where a booklet is designed to be used in conjunction with a calculator or "handout", e.g., ruler, the items are randomly selected for a booklet. While this is advantageous from a matrix sampling point of view, it means that the content validity must be checked for each data set an investigator contemplates using. Some booklets are strong in one content domain and other booklets in another. With the exception of booklet 12C, which was designed for a calculator study, the booklets have at least a few items for each content and process category. This contributes to the content validity of the measurement of overall mathematics achievement. It also implies that researchers must be selective if their interests require measurement of specific subsets of this domain.

(2) FACE VALIDITY

The second step recommended is to review the individual items for face validity. We found some questions to be potentially ambiguous in meaning, for instance:

> The Thompson's dinner bill totaled $28.75. Mr. Thompson wants to leave a tip of about 15%. About how much should he leave for the tip?

If the student interprets the phrase "About how much" literally, then a rough guess is an acceptable answer. Apparently the NAEP scorers interpreted the question more restrictively as only 23% of the 17-year-old students got this question "right".

Another exercise had semantic problems:

> An advertisement for a sale indicates that all merchandise has been reduced by 40 percent of its regular price. If the sale price of a washer is advertised at $144, what was its regular price before the sale?

The phrase "reduced by 40 percent" is not commonly used in retailing; instead the common terminology is "40% off" or "reduced 40%." While not everyone may agree with our interpretation, the item was either very confusing or very difficult because only 4% got the correct answer.

The seriousness of this problem may be less than is suggested here. As described by Holmes (1980), NAEP has followed unusually careful procedures to guard against any kind of hidden bias or ambiguity in the language of their test items. Never-the-less, we recommend that secondary analysts examine the specific items they incorporate into subtests. If any items appear questionable, they are condidates for deletion.

(3) BAD CASE ANALYSIS

The third step calls for an additional quality review of the data. Specifically, cases suspected to have considerable missing information should be examined. In checking these problematic cases we found only one or two cases with seriously large blocks of missing data. For purposes of completeness we left these cases in the working data set.

(4) CONSTRUCT VALIDITY

The construct properties of mathematics achievement were explored by first pooling the three data sets and then computing the subtest intercorrelations (Table 2). In general the intercorrelations are very high; most are greater than 0.70. This structure implies that the subdomains represented by the subtests are not highly distinct from the global construct, mathematics achievement. It may be that the cognitive capacities required for one subdomain are quite similar to those required by another, or by mathematics in general.

To explore this question further we performed a series of factor analyses on the MATS10 data set. Both the alpha and the principle component methods of factor extraction were applied to the complete set of items. The results were not particularly useful because the resulting factors appeared to be mainly a function of the multi-part exercise structure in NAEP booklets. A number of exercises, particularly in mathematics have several subquestions which NAEP calls "parts". For instance one page might have 3 simple division problems requiring three different answers, but they are identified as a single exercise. Because the parts tend to be highly similar to each other and in some cases build upon the preceding part, the intercorrelations tend to be unusually high. These clusters of correlations produce a sufficiently large amount of common variance and often are extracted as a single factor. This resulted in a great many factors and no meaningful structures. In order to avoid the multi-part exercise problem, we randomly selected one part from each of the multi-part questions. This reduced the item set to 31 items. A principle components, varimax-rotated factor matrix of these items is presented in Table 3. Four factors explaining 37% of the total variance were produced, although the first factor accounted for most of that variance. Table 4 was assembled to provide assistance in the interpretation of the factor solution. Whenever an item had a factor loading greater than 0.30, it was listed with that factor and the symbols for its content and process categories were entered adjacent to it in the table. The symbols are simply the first letter of each category label, e.g., N for Numbers; A for Application; S for Shape; V for Variables; M for Measurement; U for Understanding; and K for Knowledge and Skills. What is surprising about this is that both before and after reducing the items, no discernable structure is evident. The only clustering of items roughly representing a single cell in the objectives matrix is factor 4, where half of the items are a combination of Numbers and Knowledge/Skill. The other factors have an even less defined conceptual character.

When we ignored the prior assignment of items to categories of objectives, we still did not find clearly meaningful clusters of items associated with specific factors. The results suggest that learning in mathematics is probably not as segmented and multi-faceted as the definition of the content domain suggests.

They also suggest that secondary analysts must proceed with great caution in constructing subtests from NAEP data packages. Even if the subtest reliabilities are adequate, the items may not constitute a homogenou, independent cluster in all respects.

(5) ITEM ANALYSIS

The reliability of a test indicates how free it is from random error and certain types of indicator bias. In our analysis of reliability we applied the widely used coefficient of internal consistency, Cronbach's alpha. Reliability coefficients were computed for all the scales corresponding to all nine of the content/process categories. These results, along with the number of items for each scale and the average proportion correct are given in Tables 5-7 for each of the three data sets. In these tables the "complete" tests include all of the available items whereas the "refined" tests have a few items deleted according to criteria which will be explained later. The reliability levels of the composite or total test are above 0.93 in every instance. This level of reliability is quite satisfactory and is consistent with the findings of Mullis (1978) from the 1976 supplementary mathematics assessment. The shorter scales or subtests not surprisingly have lower reliability levels and these vary from package to package. They are most sensitive to differences in the number of relevant items in a package. For example, data MATS01 has only 4 shape (geometry) items, which have an alpha of only 0.36, whereas MATS03 has 14 shape items and they have a reliability of 0.84. Nunnally (1978) recommends that reliabilities of tests be at least 0.70 for "exploratory research". Using this criterion we find an acceptable reliability for each test and subtest in at least one data set with the exception of the "measurement" scale and the "other topics" scale. There are typically only a few items for each of these two areas and perhaps most importantly, their respective domains are not defined as homogeneous, unified categories.

The deletion of items for the "refined" tests were based upon traditional item analsis techniques. Specifically, items were dropped if they had either (1) a point-biserial correlation of less than 0.30, or (2) an extreme p-value, i.e., less than 10% or more than 90% answered the item correctly. As can be seen in Tables 5, 6 and 7, the reliabilities of the "refined" tests are roughly the same as the "complete" tests, even though the number of items is generally reduced. While it might seem desirable to use the reduced or refined version of the tests, there are strong arguments in favor of using the "complete" tests. The philosophy of assessment argues that the full range of ability, not just the average, should be tested (Tyler, 1970). If criteria of discrimination, i.e., high biserial correlation, are applied, then items which are extreme tend to be eliminated. Since the objective of assessment is to measure total performance of all students, very easy and very difficult items should be included as long as they reside within the definition of the domain of content. Thus, while the "refined" test is more efficient, the "complete" test is probably more valid in that it represents the domain more fully. Our recommendation is that "complete" tests or subtests be utilized in secondary analysis except when the reliability is low.

(6) EXAMINE SHAPE OF DISTRIBUTION

The final step in our recommended procedures specifies the construction of a total score for achievement and a visual examination of its distribution within the sample. Nunnally(1978) stresses the importance of this activity for the purpose of identifying whether or not the distribution is symmetrical and/or skewed. A sample distribution from one data set is displayed in Figure 4. It reveals



ERIC

that the distribution is quite symmetrical and only has a small upward skew. The average percent correct on this set of items was .6.

## SUMMARY

The advent of NAEP public-use data files opens up a wealth of possibilities for those who have the skills, time, and resources to do secondary analysis. Our analysis of the mathematics test items has demonstrated the potential for developing indicators of mathematics achievement. This analysis has also demonstrated that the NAEP item subsets will not always meet conventional psychometric criteria. This failure to meet standard achievement test criteria does not mean that secondary analysis of the data is unwise. But it does imply that interpretation of findings, especially those using subtests, must be made cautiously. Limitations of the methodology must be acknowledged.

Conventional achievement testing is not item-centered like assessment testing. The measurement priority of assessment is stability across multiple testings, not relative comparisons among persons. Consequently, standards of item discrimination and construct validity have obviously less import. Of far greater importance for assessments are standards of face validity, content validity, internal consistency and the application of rigorous data analysis techniques. Our recommended procedures should be followed any time a secondary analyst seeks to utilize a package-level test or subtest of performance or achievement. Prudent methodology will insure the discovery of the substantive potential buried in the NAEP data base.

## BIBLIOGRAPHY

American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association. (1966)

Anderson, R.E., Welch, W., and L. Harris. "Methodological Problems in Secondary Analysis of Data from the National Assessment for Educational Progress". Minnesota Center for Social Research, University of Minnesota, Minneapolis, MN. (1980)

Begle, E.G. Critical Variables in Mathematics Education. Mathematical Association of America and the National Council of Teachers of Mathematics, Washington, D.C. (1979)

Guilford, J.P. Psychometric Methods. New York: McGraw-Hill Book Company, Inc. (1954)

Holmes, B.J. "Bias: Psychometric and Social Implications for the National Assessment for Educational Progress". Denver, Colorado: Education Commission of the States. (1980)

Lord, F.M. and M.R. Novick. Statistical Theories of Mental Test Scores. Reading, Mass: Addison- Wesley. (1968)

Millman, Jason. "Strategies for Constructing Criterion-Referenced Assessment Instruments", Cornell University. (1978)

Mullis, I. Effects of Home and School on Learning Mathematics and Political Knowledge and Attitudes. University of Colorado, PhD Dissertation, Boulder, CO. (1978).

National Assessment of Educational Progress. Mathematics Objectives, Second Assessment. Denver, CO: Education Commission of the States (1978).

National Assessment of Educational Progress. Procedural Handbook: 1977-78 Mathematics Assessment, Education Commission of the States, Suite 700, 1860 Lincoln Street, Denver, Colorado, 80295. (1980).

Nunnally, J.C. Psychometric Theory, Second Edition. New York: McGraw-Hill. (1978)

Tyler, Ralph W. "National Assessment: A History and Sociology," School and Society, December, 1970, pp. 471-77.

Womer, F.B. Developing a Large Scale Assessment Program. Denver, CO: Cooperative Accountability Project, 1973.

Wright, D.J., Larsson, I.E., Ramlow, C.E., "National Assessment Public-Use Tapes," Denver, CO: Education Commission of the States, 1981.



ERIC

FIGURE 1

PROCEDURES FOR EVALUATING ACHIEVEMENT INDICATORS

1. Perform content validity analyses by examining domain objectives and number of data package items falling into each cell of the objectives categorization to determine if all subdomains are adequately represented.

2. Perform face validity analysis on all exercises and items by examining the questions and the response distributions to determine if any items are seriously questionable.

3. Check for bad cases by listing those cases
   with a composite score of zero, or
   those possessing a problematic value on the package condition code (PKGCON), which generally indicates partial completion. If the listing reveals either total nonresponse on math exercises or total nonresponse on background items, then reject that particular case.

4. Perform factor analysis of items to evaluate construct validity of items.

5. Perform item analysis to obtain reliability estimates and to identify contributions of individual items.

6. Produce and examine histograms of achievement scores for the complete set of exercises to evaluate properties of the distribution including number of extreme scores.

FIGURE 2

FRAMEWORK FOR OBJECTIVES, 1977-78 MATHEMATICS ASSESSMENT

CONTENT

| PROCESS | | A. Numbers and Numeration | B. Variables and Relationships | C. Shape, Size and Position | D. Measurement | E. Other Topics |
|---|---|---|---|---|---|---|
| I. | Mathematical knowledge | | | | | |
| II. | Mathematical skill | | | | | |
| III. | Mathematical understanding | | | | | |
| IV. | Mathematical application | | | | | |

FIGURE 3

# BLUEPRINT DEFINING RELATIVE SIZE OF CONTENT AND PROCESS CATEGORIES OF MATHEMATICS OBJECTIVES*

**Approximate Number of Exercises by Age and Content**

| CONTENT | Age 9 | Age 13 | Age 17 |
|---|---|---|---|
| A. Numbers and numeration | 110 | 150 | 100 |
| B. Variables and relationships | 20 | 40 | 90 |
| C. Shape, size and position | 30 | 50 | 70 |
| D. Measurement | 40 | 50 | 60 |
| E. Other topics | 30 | 60 | 50 |

**Approximate Number of Exercises by Age and Process**

| | Age 9 | Age 13 | Age 17 |
|---|---|---|---|
| Mathematical knowledge | 45 | 45 | 55 |
| Mathematical skill | 65 | 85 | 110 |
| Mathematical understanding | 60 | 105 | 105 |
| Mathematical application | 60 | 115 | 160 |

Figure 4

DISTRIBUTION OF MATHEMATICS ACHIEVEMENT TOTAL SCORE (COMPLETE TEST OF 58 ITEMS) FOR 17 YEAR OLDS IN 1977-78 NAEP DATA SET MATS10 (N=2294)

| SCORE | COUNT |
|---|---|
| 57 | 8 |
| 56 | 7 |
| 55 | 9 |
| 54 | 13 |
| 53 | 19 |
| 52 | 20 |
| 51 | 21 |
| 50 | 30 |
| 49 | 24 |
| 48 | 41 |
| 47 | 27 |
| 46 | 29 |
| 45 | 26 |
| 44 | 38 |
| 43 | 26 |
| 42 | 34 |
| 41 | 35 |
| 40 | 32 |
| 39 | 26 |
| 38 | 41 |
| 37 | 34 |
| 36 | 35 |
| 35 | 38 |
| 34 | 34 |
| 33 | 36 |
| 32 | 31 |
| 31 | 25 |
| 30 | 30 |
| 29 | 39 |
| 28 | 25 |
| 27 | 20 |
| 26 | 30 |
| 25 | 17 |
| 24 | 21 |
| 23 | 22 |
| 22 | 23 |
| 21 | 15 |
| 20 | 13 |
| 19 | 12 |
| 18 | 7 |
| 17 | 8 |
| 16 | 8 |
| 15 | 11 |
| 14 | 6 |
| 13 | 7 |
| 12 | 4 |
| 11 | 1 |
| 10 | 2 |
| 9 | 2 |
| 8 | 1 |
| 7 | 2 |
| 6 | 3 |
| 5 | 0 |
| 4 | 1 |
| 3 | 0 |
| 2 | 0 |
| 1 | 1 |
| 0 | 4 |

TABLE 1

PERCENT OF TOTAL MATHEMATICS ACHIEVEMENT ITEMS IN EACH CONTENT AND PROCESS CATEGORY FOR 1977-78 AGE 17 DATA

| | Booklet Identification No. | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONTENT CLASSES | 1 | 2 | 3H | 4 | 5 | 6 | 7 | 8 | 9H | 11 | 12C | 10 | Blue-Print | All Books |
| A. NUMBERS | 58 | 41 | 38 | 46 | 38 | 37 | 56 | 47 | 36 | 50 | 40 | 53 | 37% | 45% |
| B. VARIABLES | 13 | 18 | 16 | 22 | 20 | 18 | 11 | 14 | 19 | 15 | 4 | 14 | 21% | 17% |
| C. SHAPE | 7 | 20 | 24 | 18 | 22 | 11 | 11 | 14 | 20 | 15 | 0 | 19 | 16% | 17% |
| D. MEASUREMENT | 11 | 11 | 10 | 5 | 7 | 16 | 6 | 8 | 14 | 10 | 2 | 3 | 14% | 9% |
| E. OTHER TOPICS | 11 | 10 | 12 | 9 | 13 | 19 | 17 | 18 | 11 | 11 | 54 | 10 | 12% | 12% |
| | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100% | 100% |

| PROCESS CLASSES | 1 | 2 | 3H | 4 | 5 | 6 | 7 | 8 | 9H | 11 | 12C | 10 | Blue-Print | All Books |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I.& II. Knowledge & skills | 60 | 56 | 64 | 72 | 67 | 56 | 70 | 54 | 67 | 66 | 100 | 58 | 38% | 63% |
| III. Understanding | 18 | 23 | 21 | 11 | 7 | 19 | 15 | 20 | 17 | 8 | 0 | 18 | 24% | 16% |
| IV. Applications | 22 | 21 | 16 | 17 | 25 | 25 | 14 | 26 | 16 | 26 | 0 | 25 | 37% | 21% |
| | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100% | 100% |

| | 1 | 2 | 3H | 4 | 5 | 6 | 7 | 8 | 9H | 11 | 12C | 10 | Blue-Print | All Books |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N OF ITEMS | 55 | 61 | 58 | 65 | 55 | 57 | 72 | 51 | 64 | 62 | 48 | 58 | 430 | 654 |

TABLE 2

INTERCORRELATIONS AMONG MATHEMATICS REFINED SUBTESTS AND COMPOSITE TEST, 1977 NAEP DATA SETS MATS01, MATS03, AND MATS10 POOLED (N=6,782 )

| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|---|
| Numbers (arithmetic) | (1) | | | | | | | |
| Variables (algebra) | (2) | .72 | | | | | | |
| Shapes (geometry) | (3) | .63 | .60 | | | | | |
| Knowledge/Skills | (4) | .93 | .81 | .74 | | | | |
| Understanding | (5) | .80 | .74 | .72 | .77 | | | |
| Application | (6) | .76 | .67 | .68 | .72 | .67 | | |
| Total | (7) | .94 | .85 | .78 | .97 | .87 | .84 | |



ERIC Full Text Provided by ERIC

TABLE 3

FACTOR MATRIX (ROTATED) OF 31 SELECTED MATHEMATICS ITEMS FROM 1977-78 DATA SET MATS10

| ITEMS | FACTORS 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| E04b | .35 | .25 | .06 | .32 |
| E05 | .38 | .21 | .16 | .19 |
| E06b | .22 | -.02 | .09 | .44 |
| E07 | .24 | .11 | .16 | .47 |
| E08d | .14 | .12 | .09 | .44 |
| E09a | .21 | .24 | .04 | .54 |
| E10 | .14 | .11 | .26 | .21 |
| E12 | .43 | .09 | .22 | .18 |
| E13a | .23 | .14 | .43 | .09 |
| E14 | .37 | .21 | .34 | .10 |
| E15 | .24 | .14 | .38 | .10 |
| E16 | .24 | .23 | .17 | .10 |
| E18 | .10 | .06 | .40 | .11 |
| E19e | .37 | .32 | .19 | .25 |
| E20 | .20 | .39 | .50 | .05 |
| E21 | .20 | .53 | .42 | .18 |
| E22a | .20 | .38 | .26 | .27 |
| E23 | .22 | .43 | .14 | .19 |
| E24b | .11 | .18 | .24 | .10 |
| E25c | .11 | .31 | .20 | .42 |
| E26 | .15 | .49 | .41 | .24 |
| E27a | .10 | .18 | .17 | .33 |
| E28 | .34 | .03 | .09 | .15 |
| E29 | .33 | .16 | .12 | .06 |
| E30 | .29 | .42 | .21 | .36 |
| E31 | .15 | .46 | .48 | .09 |
| E32 | .42 | .26 | .11 | .23 |
| E33 | .24 | .37 | .34 | .27 |
| E36 | .44 | .08 | .13 | .24 |
| E38 | .31 | .13 | .14 | .13 |
| E39 | .39 | .11 | .23 | .13 |
| Eigenvalues | 7.8 | 1.6 | 1.2 | 1.0 |
| Proportion of Total Variance | .25 | .05 | .04 | .03 |

## TABLE 4

### CONTENT CATEGORIZATION OF ITEMS FROM FACTOR ANALYSIS OF THE 31 REDUCED ITEM SET FOR MATHEMATICS, 1977-78 NAEP DATA SET MATS10

| Factor 1 | | | Factor 2 | | | Factor 3 | | | Factor 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item # | Cont. | Process | Item # | Cont. | Process | Item # | Cont. | Process | Item # | Cont. | Process |
| 04b | N | K | 19e | N | K | 13a | O | A | 4b | N | K |
| 5 | N | A | 20 | O | A | 14 | S | K | 6b | N | U |
| 12 | N | A | 21 | V | K | 15 | N | A | 7 | V | U |
| 14 | S | K | 22a | V | A | 18 | O | U | 8d | N | K |
| 19e | N | K | 23 | N | K | 20 | O | A | 9a | N | K |
| 28 | S | K | 25c | N | K | 21 | V | K | 25g | N | K |
| 29 | O* | U | 26 | V | K | 26 | V | K | 27a | N | K |
| 32 | O | A | 30 | V | U | 31 | V | K | 30 | V | U |
| 36 | N | K | 33 | S | K | 33 | S | K | | | |
| 38 | O | A | | | | | | | | | |
| 39 | S | U | | | | | | | | | |

* "O" represents the "Other topics" content category (including "Measurement").

** Items were included in this table whenever they had factor loadings greater than .30.

TABLE 5

RELIABILITIES OF MATHEMATICS SUBTESTS AND COMPOSITE TEST FOR COMPLETE AND REFINED TESTS, 1977-78 NAEP
DATA SET MATS01 (N=2294)

| | Complete Test | | | Refined Test | | |
|---|---|---|---|---|---|---|
| | N of Items | Reliability (Alpha) | Average Percent Correct | N of Items | Reliability (Alpha) | Average Percent Correct |
| Numbers (arithmetic) | 32 | .88 | .62 | 26 | .88 | .54 |
| Variables (algebra) | 7 | .60 | .57 | 6 | .61 | .61 |
| Shape (geometry) | 4 | .40 | .45 | 3 | .33 | .53 |
| Measurement | 6 | .65 | .64 | 6 | .65 | .64 |
| Other | 6 | .48 | .58 | 4 | .49 | .61 |
| Knowledge/Skills | 33 | .87 | .66 | 22 | .84 | .67 |
| Understanding | 10 | .69 | .56 | 8 | .70 | .56 |
| Applications | 12 | .65 | .47 | 8 | .65 | .46 |
| Total Test | 55 | .92 | .60 | 45 | .92 | .57 |

ERIC

TABLE 6

RELIABILITIES OF MATHEMATICS SUBTESTS AND COMPOSITE TEST FOR COMPLETE AND REFINED TESTS, 1977-78 NAEP
DATA SET MATSO3 (N=2272)

| | Complete Test | | | Refined Test | | |
|---|---|---|---|---|---|---|
| | N of Items | Reliability (Alpha) | Average Percent Correct | N of Items | Reliability (Alpha) | Average Percent Correct |
| Numbers (arithmetic) | 22 | .85 | .62 | 20 | .85 | .60 |
| Variables (algebra) | 9 | .76 | .30 | 8 | .76 | .34 |
| Shape (geometry) | 14 | .84 | .51 | 14 | .84 | .51 |
| Measurement | 6 | .59 | .59 | 4 | .58 | .42 |
| Other | 7 | .53 | .63 | 7 | .53 | .63 |
| Knowledge/Skills | 36 | .90 | .50 | 32 | .90 | .50 |
| Understanding | 12 | .76 | .59 | 12 | .76 | .59 |
| Applications | 9 | .64 | .48 | 8 | .63 | .53 |
| Total Test | 58 | .93 | .54 | 53 | .93 | .53 |

TABLE 7

RELIABILITIES OF MATHEMATICS SUBTESTS AND COMPOSITE TEST FOR COMPLETE AND REFINED TESTS, 1977-78 NAEP
DATA SET MATS10 (N = 2216)

| | Complete Test | | | Refined Test | | |
|---|---|---|---|---|---|---|
| | N of Items | Reliability (Alpha) | Average Percent Correct | N of Items | Reliability (Alpha) | Average Percent Correct |
| Numbers (arithmetic) | 31 | .89 | .68 | 25 | .89 | .71 |
| Variables (algebra) | 8 | .79 | .49 | 7 | .80 | .43 |
| Shape (geometry) | 11 | .72 | .53 | 9 | .72 | .58 |
| Measurement | 2 | .37 | .49 | 2 | .37 | .49 |
| Other | 6 | .52 | .46 | 4 | .48 | .50 |
| Knowledge/Skills | 33 | .91 | .70 | 27 | .90 | .67 |
| Understanding | 10 | .72 | .69 | 9 | .70 | .66 |
| Applications | 14 | .75 | .40 | 10 | .75 | .48 |
| Total Test | 58 | .93 | .62 | 47 | .93 | .61 |



ERIC