DOCUMENT RESUME

ED 236 250 TM 830 780

AUTHOR	Searls, Donald T., Ed.
TITLE	National Assessment Analysis Procedures.
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
REPORT NO	NAEP-AY-AP-35
PUB DATE	Aug 83
GRANT	NIE-G-80-0003
NOTE	68p.; For related documents, see ED 194 605 and ED 223 679.
AVAILABLE FROM	National Assessment of Educational Progress, Box 2923, Princeton, NJ 08541.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Data Analysis; Databases; Data Collection; *Educational Assessment; Error of Measurement; Mathematical Formulas; *National Programs; Sampling; Scoring; Test Construction; Testing Programs
IDENTIFIERS	*National Assessment of Educational Progress; Secondary Analysis

ABSTRACT

The purpose of this paper is to provide an overview of the analysis of data collected by the National Assessment of Educational Progress (NAEP). In simplest terms, the analysis can be characterized as establishing baseline estimates of the percentages of young Americans possessing certain skills, knowledge, understandings, and attitudes and producing estimates of changes in these percentages over time. The baseline estimates permit comparisons of various subgroups. This paper begins with brief descriptions of key activities. The first sections generally describe the methods used to develop objectives and exercises, select the assessment sample, prepare material for the administration of an assessment, administer the booklets, and score the items. The later sections contain discussions about the NAEP analysis including computations used and potential secondary analyses. Appendices cover a variety of topics including adjustment procedures used in the analysis such as balancing and weight smoothing, methods for equating scores across booklets, and an approach for studying response patterns and bias. Primary type of information provided by report: Procedures (Analysis) (Data Processing). (BW)

NATIONAL ASSESSMENT

ANALYSIS PROCEDURES

AY-AP-35

Education Commission of the States
Suite 700, 1860 Lincoln Street
Denver, Colorado 80295

Edited by
Donald T. Searls

August 1983

TABLE OF CONTENTS

INTRODUCTION

The purpose of this paper is to provide an overview of the analysis of data collected by the National Assessment of Educational Progress (NAEP). In simplest terms, the analysis can be characterized as establishing baseline estimates of the percentages of young Americans possessing certain skills, knowledge, understandings and attitudes and producing estimates of changes in these percentages over time. The baseline estimates permit comparisons of various subgroups.

It should be emphasized that National Assessment is not designed to monitor changes or growth in achievement of individual students. Its purpose is to report on the current educational status of young Americans and to monitor any changes in achievement over time. The way National Assessment has chosen to do this is to monitor three in-school age groups--9-, 13- and 17-year-olds--to see if they are gaining or losing ground in comparison to predecessors of the same age groups in prior assessments. When resources permit, the Assessment also includes a sample of young adults aged 26 to 35 and a sample of 17-year-olds who are not enrolled in a secondary school. National Assessment has conducted major assessments in art, career and occupational development, citizenship, literature, mathematics, music, reading, science, social studies, writing and in several other learning areas on a smaller scale. Nine of these major areas have been reassessed one or more times. Learning areas and ages assessed since 1969 are shown in Exhibit 1.

Learning area assessments evolve from a consensus process. Each assessment is the product of many months of work by a great many educators, scholars and lay persons from all over the nation. After assessment materials have been developed, field tested and reviewed, exercises (items) and background questions are assembled into booklets that can be completed by respondents in about 45 minutes. Each booklet contains a unique set of from four to thirty or more exercises from a specific learning area and a common set of background questions. Exercises may have only one part that requires either a multiple choice or a written response, or they may include several parts, such as multiple questions about a reading passage or several questions about the same topic. Exercises are included from previous assessments to measure changes in achievement, and some exercises are also administered to two or more age groups.

Exhibit 1. Learning Areas and Ages Assessed from 1969 to 1982

Assessment Year/Learning Areas			Ages Assessed*		
	9	13	17IS	17OS	Adult
1—1969-70					
Science	X	X	X	X	X
Writing	X	X	X	X	X
Citizenship	X	X	X	X	X
2—1970-71					
Reading	X	X	X	X	X
Literature	X	X	X	X	X
3—1971-72					
Music	X	X	X	X	X
Social Studies	X	X	X	X	X
4—1972-73					
Science (2)	X	X	X	X	X
Mathematics	X	X	X	X	X
5—1973-74					
Career & Occupational Development	X	X	X	X	X
Writing (2)	X	X	X	X	X
6—1974-75					
Reading (2)	X	X	X	X	
Art	X	X	X	X	
7—1975-76					
Citizenship/Social Studies (2)	X	X	X	X	
Mathematics**	X	X	X		
8—1976-77					
Science (3)	X	X	X		
Basic Life Skills**				X	
Health**				X	
Energy**				X	
Reading** (2)				X	
Science** (3)				X	
9—1977-78					
Mathematics (2)	X	X	X		
Consumer Skills**			X		
10—1978-79					
Art (2)	X	X	X		
Music (2)	X	X	X		
Writing (3)	X	X	X		
11—1979-80					
Reading (3)/Literature (2)	X	X	X	X	
Art (2)		X			

12—1980-81			
No data collection***			
13—1981-82			
Mathematics (3)	X	X	X
Citizenship/Social Studies (3)	X	X	X
Science** (4)	X	X	X
14—1982-83			
No data collection			

NOTES:

*17IS denotes 17-year-olds enrolled in public or private schools; 17OS denotes 17-year-olds who dropped out of school or graduated prior to the time of the assessment.

**Indicates small, special-interest probe assessments conducted on limited samples at specific ages.

***First year of every other year data collection due to budgeting constraints.

( )Shows second and subsequent assessments of an area.

National Assessment uses a deeply stratified, three-stage probability sample in all school assessments. The first stage consists of geographic areas (typically counties). Second, within each geographic area both public and private schools are sampled. Third, within each sampled school a separate random sample of students is drawn, and each student is randomly assigned to respond to one of the exercise booklets. Seventeen-year-olds who are no longer in school are typically drawn from lists of dropouts and early graduates from sampled schools. Young adults are drawn from randomly selected households within the sampled geographic areas. Out-of-school 17-year-olds and young adults have been asked to answer up to four exercise booklets.

For a particular learning area, between three and fifteen different booklets of exercises are developed for each age group. Since each student in school responds to only one booklet, this means that there are multiple student samples for each age group. A specific school is likely to have more than one booklet administration; however, booklet administrations are randomly allocated to schools.

In each assessment, 13-year-olds are assessed in October through December, 9-year-olds in January and February, and 17-year-olds in March and April. Thus, the amount of school experience in terms of time spent in school is approximately the same in each assessment for each age group. (Young adults and out-of-school 17-year-olds are typically assessed during the summer months.)

The exercises for each assessment are administered by a professional data collection staff to minimize the burden on participating schools and to maximize uniformity of assessment conditions. Instructions and items are recorded on a paced audio tape and played back to students to reduce the potential effect of reading difficulties and to insure that all students move through the booklets at the same speed.

Multiple-choice items are scored by an optical scanning machine; open-ended items are hand-scored by trained scorers using scoring guides that define categories of acceptable and unacceptable responses. These scoring guides are developed following field testing of the items and then revised and refined during receipt of initial assessment data.

In addition to reporting national results, National Assessment provides data on the performance of various population subgroups within the national population: sex, race, region of the country, size and type of community lived in, grade and level of parental education. National Assessment aggregates percentages of success on various sets of items to provide data on changes in performance between assessments and on the differential performance of population subgroups.

This paper begins with brief descriptions of key activities. The first sections generally describe the methods used to develop objectives and exercises, select the assessment sample, prepare material for the administration of an assessment, administer the booklets and score the items. The latter sections contain discussions about the NAEP analysis including computations used and potential secondary analyses. Appendices cover a variety of topics including adjustment procedures used in the analysis such as balancing and weight smoothing, methods for equating scores across booklets and an approach for studying response patterns and bias.

Additional detail is contained in several NAEP publications which are listed at appropriate places in the text. Four key publications are:

AY-SA-50 Exploring National Assessment Data through Secondary Analysis, 1982.

SY-DT-50 Introduction to the National Assessment of Educational Progress Public Use Data Tapes, 1981.

11-RL-40 Procedural Handbook: 1970-80 Reading and Literature Assessment, 1981 ED 210 300;

12-IP-57 Issues in the Analysis and Analysis of Change of National Assessment Data, 1980.

CHAPTER 1

KEY ACTIVITIES LEADING TO ANALYSIS

Assessment Planning and Objective Development

The primary goal of National Assessment is to report on the current educational status of young Americans and to monitor any changes in achievement over time.

Planning for a future assessment begins almost three years before the start of data collection. It is a two-phase approach that involves 1) preplanning and 2) detailed background research into all aspects of a given learning area. The planning phase lasts about six months.

Preplanning involves a review and update of National Assessment long-range plans and includes factors such as length of time since the last assessment of a learning area, budget projections and constraints, and importance of the learning area to state and local educational groups, federal agencies and congressional groups.

After the Assessment Policy Committee (APC) has determined that a learning area should be assessed, staff begins detailed background research on the area. This detailed background research includes a review of all aspects of the previous assessment(s) to identify problem areas; a review of recent curricular innovations to ensure the most relevant assessment materials; and a review of existing measurement capabilities so that the latest measurement research and technology is included in or accounted for in the redevelopment process. Curriculum and research specialists in the learning area are also identified during this phase.

The goal of the assessment planning process is a comprehensive plan that includes a rationale for assessing the learning area, describes operational and budget constraints, describes development needs, outlines personnel assignments and schedules and sets out a tentative assessment design, including analysis and reporting possibilities. For each learning area to be assessed, NAEP asks consultants to develop objectives that define the subject area. In addition, they are asked to create guidelines for exercise writers by specifying examples of the knowledge, skills and attitudes to be assessed at each age level.

To develop an assessment that is truly national in scope and takes into account the diversity of curricula, values and goals across the country, National Assessment employs a consensus process for developing objectives, involving representation of many different groups of people.

Several types of consultants help either to develop new objectives or to review and revise existing NAEP objectives developed for prior assessments. Educators and scholars, including college and university specialists, classroom teachers, curriculum supervisors and persons involved in teacher education, make sure that the objectives include concepts, skills and attitudes that the schools should be teaching and those that they presently are teaching. Concerned citizens, parents and other interested lay persons have to agree that the objectives are important for young people to achieve, are free of education jargon and are not biased or offensive to any groups. Consultants are representative of the different regions of the country, minority groups, various types of communities, age levels, education philosophies, and so on.

As an example an outline of the Reading objectives developed for the 1979-80 assessment follows:

I. Values reading and literature
   A. Values the benefits of reading for the individual
   B. Appreciates the cultural role of written discourse as a way of transmitting, sustaining and changing the values of a society

II. Comprehends written works
   A. Comprehends words and lexical relationships
   B. Comprehends propositional relationships
   C. Comprehends textual relationships

III. Responds to written works in interpretive and evaluative ways
   A. Extends understanding of written works through interpretation
   B. Evaluates written works

IV. Applies study skills in reading
   A. Obtains information from nonprose reading facilitators
   B. Uses the various parts of a book
   C. Obtains information from materials commonly found in libraries or resource centers
   D. Uses various study techniques

Development of Exercises

Exercises are developed to provide information about achievement levels for objectives and subobjectives or cells of a content matrix. Each exercise is designed so that its results either can be used alone, as an indicator of performance on a specific task, or used in conjunction with results from other exercises to give a more general picture of achievement levels. Exercises are developed to provide information suitable for analyzing changes in performance over time. In this sense they need to be replicable and a valid measure of achievement over time.

Most item writing is done by groups of people knowledgeable in the subject area. Individuals generate items, which are then reviewed and revised by other professionals in the subject area. Reviewers consider age-level appropriateness, accuracy of content, how well the item measures a question or objective, and readability. Exercises passing the group review are edited by the National Assessment staff to fit NAEP format and technical requirements.

The exercises produced by the writing groups are field tested in schools across the country to discover potential problems in wording, directions or administration procedures and to collect item statistics, timing information and scoring information. "Tryout" schools are selected to represent high- and low-income communities as well as more typical communities.

After the initial exercise pool is developed and field tested, consultants review the exercises and accompanying tryout data to insure that content areas are adequately covered. These people edit the existing exercises and generate new ones, which again, are field tested.

Finally, all items considered appropriate for inclusion in an assessment are reviewed in a series of conferences by numerous consultants. Exercises for each age group are reviewed by a variety of subject-matter specialists including classroom teachers. Lay citizens, representing a variety of occupations and interests, also review the exercises, checking for any type of bias and considering the general importance of each exercise. For more detailed information, see The National Assessment Approach to Objectives and Exercise Development (1980).

Preparation of Assessment Materials

Following the selection of exercises to be included in an assessment, National Assessment staff group and sequence items into exercise booklets. Since students at different ages receive somewhat different sets of exercises, booklets are constructed separately for each age level.

The following constraints are observed in the preparation of exercise booklets:

— Each booklet contains exercises of varying difficulty so that students will not become bored by many easy exercises or discouraged by many difficult exercises.

— Booklets are designed to be as parallel as possible with respect to the number of different objectives measured. Exercises measuring a particular objective are scattered throughout the booklets so that many different students will respond to questions related to a particular objective.

— Exercises cannot cue other exercises. In other words, the answer to one exercise cannot be contained in another exercise in the same booklet.

— Each booklet is timed so that it will take no more than 45 minutes—the length of a typical class period—of a student's time. Booklets contain approximately 30 to 35 minutes of exercise time and an additional 10 to 15 minutes of introductory material, instructions and background questions.

National Assessment makes every effort to minimize difficulties connected with the testing situation so that results will be, as nearly as possible, an accurate reflection of what students know and can do. For example, students have marked their answers directly in the assessment booklets, not on separate answer sheets. It is felt that this procedure reduces the possibility of errors in marking answer sheets, especially for the younger students. To minimize guessing, students are encouraged to write "I don't know" on the answer line for open-ended questions or to select the "I don't know" response option included with each multiple-choice exercise if they feel they do not know the answer to a question.

Paced audio tapes have been used with each exercise booklet to minimize the effect of any reading difficulties and to insure that all students move through the booklets at the same speed. In addition, the use of tapes helps to insure uniform assessment conditions across the country. The following is a typical introduction used with students being assessed.

> "You have been chosen to take part in the National Assessment of Educational Progress. More than 80,000 people from all parts of the United States participate in this program each year. The purpose of National Assessment is to find out the things people know and can do as a result of their education. Although many of the tests you take are given to find out how well you do compared to other students, National Assessment is different. It is interested in finding out <u>how many</u> students know or can do certain things. The results are used to make improvements in American Education.
>
> Many of the questions you will be answering look like the usual kind of test questions. Because it is a survey, you may find that some questions seem easy to answer, and others seem hard. It is important to <u>answer every</u> question carefully.
>
> Your answers will <u>not</u> be shown to anyone in your school and your name will <u>not</u> appear on any materials leaving school.
>
> Now open the booklet to page 2."

Sampling

The target populations for each assessment consist of 9-, 13- and 17-year-olds enrolled in either public or private schools at the time of the assessment who are not functionally handicapped to the extent that they cannot participate in an assessment. Specific groups excluded are: non-English-speaking persons, those identified as nonreaders, persons physically or mentally unable to respond, and persons in institutions or attending schools established for the physically or mentally handicapped. See Table 1 for percent of students excluded.

TABLE 1

PERCENT OF STUDENTS EXCLUDED FROM AN ASSESSMENT

	Age		
Groups Excluded	9	13	17
Non-English-speaking	1.4	1..	0.9
EMR	2.0	2.1	1.5
Functionally disabled and other	1.7	1.8	1.1
Non-readers	*	*	*
Included	94.9	94.8	96.5
TOTAL	100.0	100.0	100.0

*Less than one-half of one-tenth of one percent.

National Assessment does not follow up specific individuals from one assessment to the next. However, in each assessment year, participants are carefully selected to represent each age level. The definitions of the target populations are identical in each assessment. However, the sample design used to obtain representative samples of the target populations may be modified somewhat between assessments. The sampling approach used is planned to approach optimal economic efficiency.

National Assessment uses a deeply stratified, clustered three-stage, national probability sample design with oversampling of low-income and rural areas. In the first stage, the United States is divided into geographical units of counties or groups of contiguous counties meeting a minimum population size requirement of 50,000. These units, called primary sampling units (PSUs), are stratified by region and size of community (See Appendix 1). From the list of PSUs, a sample of PSUs is drawn (without replacement) with probability proportional to population size measures, representing all regions and sizes of communities. Oversampling of low-income and extreme-rural areas is first performed at this stage by adjusting the estimated population size measures of those areas to increase sampling rates. Within PSUs, Census Employment Survey Data are used to further delineate and oversample low-income areas. Counties with high proportions of rural families are also oversampled. Oversampling is a deliberate sampling of a portion of the population at a higher rate than the remainder of the population in order to insure adequate representation of that sub-population.

In the second stage, all public and private schools within each PSU selected in the first stage are listed. Schools within each PSU are selected without replacement with approximately equal probabilities such that the number of booklets assigned to a school are proportional to the number of age-eligibles in the school.

The third stage of sampling occurs during the data collection. A list of all age-eligible students within each selected school is prepared. A simple random selection of eligible students (without replacement) is obtained, and exercise booklets are administered to selected students by specially trained personnel.

Each respondent in the sample does not have the same probability of selection, primarily because some subpopulations are oversampled at twice the rate and adjustments are made to compensate for some schools' refusal to participate and for student nonresponse. The selection probability for each individual is computed, and its reciprocal is used to weight each response in any statistical calculation to compensate for unequal rates of sampling and to insure proper representation in the population structure.

The number of PSUs, schools within PSUs and students within schools is determined by optimum sampling principles. That is, a sample design is selected that will achieve the maximum precision for a given level of resources. The current design uses about 75 PSUs in each assessment and 1,700 schools. The number of students assessed has varied from 60 to over 100 thousand depending upon resources available.

Data Collection

Participation in the National Assessment is voluntary. NAEP makes every effort to encourage the schools initially selected in the sample to participate in the assessment, and National Assessment and Research Triangle Institute staffs have obtained high rates of school cooperation (over 90 percent). Student cooperation rates are also high, especially for 9- and 13-year-olds (90 percent, 85 percent, respectively). Special follow-up procedures employed in recent years secure over 80 percent cooperation at age 17. For more information, see Access to School Districts, Schools and Nonstudents (1980).

A professional data collection staff from the Research Triangle Institute, Raleigh, North Carolina, has been used so that the burden on participating schools would be minimized and to assure uniform administration conditions.

National Assessment protects the anonymity of each respondent. Students' names are listed with their booklet identification number to enable verification if necessary. However, these lists do not leave the schools and are destroyed six months following the assessment.

School officials are asked to respond to a questionnaire asking about the enrollment in various grades, the types of communities in which the students live and the general occupational levels of the people in the community. The assessment administrator codes each student's birth date, sex, grade, racial/ethnic classification and identification number on his or her booklet. Six different racial classifications are used: white, black, Spanish heritage, American Indian or Alaskan native, Pacific Islander or Asian, and unclassified.

Each age group is assessed at approximately the same time of the school year in each assessment. As noted previously, 13-year-olds are assessed in October through December, 9-year-olds in January and February and 17-year-olds in March and April.

Following data collection, assessment administrators send completed booklets to the scoring contractor, Westinghouse DataScore Systems, Iowa City, Iowa. Booklets are counted and quality-checked to verify that correct administrative procedures were followed. Coded identification information is also checked for accuracy.

Scoring

Scoring and conversion of the data into machine-readable form have been contracted to Westinghouse DataScore Systems. Responses to multiple-choice exercises are read directly by optical scanning machines. A special staff has scored responses to open-ended exercises by hand. Scorers used well defined guides to categorize responses and code the information into ovals that can be read by the optical scanning machine.

Scorers are carefully trained in the use of the scoring guides by scoring sample responses until they feel comfortable using the guides and categorizing the data reliably. To further ensure the quality and consistency of scoring open-ended exercises, quality-control checks are conducted at regular intervals. If discrepancies in scoring became apparent, scorers were retrained and, on some occasions, responses were rescored.

To measure changes in performance accurately, all responses to open-ended items collected in successive assessments of the same subject area either were mixed together and categorized at the same time by the same scorers, or calibration samples were used.

Data Analysis

Measures of Achievement

National Assessment reports the performance of groups of students, not individuals. The basic measure of achievement is the percentages responding acceptably to an item. This percentage is an estimate of the percentage of 9-, 13- or 17-year-olds who would give acceptable responses to a given item if every 9-, 13- or 17-year-old in the country were assessed.

To present a general picture of comparisons between subgroups and changes in achievement, National Assessment summarizes the performance for each assessment (either for the entire learning area or for some appropriate set of exercises) by using the mean, or arithmetic average of percentages of acceptable responses to the exercises.

Percentage of acceptable responses are used because each item is designed as a separate measure of some aspect of an objective or subobjective. The purpose of National Assessment is to discover if more or fewer people are able to answer these items correctly.

In addition to providing national results, National Assessment reports on the achievement of various subpopulations of interest. Groups are defined by region of the country, sex, race, size and type of community lived in, grade and level of parents' education. Results for some additional variables are also analyzed. The definition of reporting groups is found in Appendix 1.

Procedures for estimating percentages of acceptable responses to exercises are dependent on the sample design. Each response by an individual is weighted. An estimate of the percentage of a particular age group that would have responded to an exercise acceptably if the entire age group were assessed is defined as the weighted number of all the responses. A similar ratio of weights is used to estimate percentages of acceptable responses for reporting groups or subpopulations of interest. See appendixes 3-8 for weighting adjustments.

Estimating Variability in Achievement Measures. National Assessment uses a national probability sample at each age level to estimate the proportion of people who would sucessfully complete an exercise. The particular sample selected is one of a large number of all possible samples of the same size that could have been selected with the same sample design. Since an achievement measure computed from each of the possible samples would differ from one sample to another, the standard deviation of this achievement measure is used as a measure of the sampling variability among achievement measures from all possible samples. A standard error, based on one particular sample, which estimate this standard deviation serves to estimate that sampling variability.

In the interest of sampling and cost efficiencies, National Assessment uses a complex, stratified, multistage probability sample design. Typically, complex designs do not provide for unbiased or simple computation of sampling errors. A reasonably good approximation of standard error estimates of acceptable response percentages is obtained by applying the jackknife procedure (Miller 1964, pp. 1594-1705; Miller 1968, pp. 567-582; Mosteller and Tukey 1968) to first-stage sampling units within strata. Standard errors for achievement measures such as group differences, mean percentage or mean group differences for a particular assessment year are estimated directly, taking advantage of features of the jackknife procedure that are common to all of these statistics. Since samples for different assessments are independent, the standard errors of the differences in achievement measures between assessments can be estimated simply by the square root of the sum of squared standard errors from each of the assessments.

For exploratory and/or special studies such as the public-private tabulations standard errors are estimated by an equation obtained from an empirical fit of the standard errors actually computed for the regular reports (see Chapter 2).

Controlling Systematic Errors. Systematic errors can be introduced at any stage of an assessment--exercise development, preparation of exercise booklets, design or administration procedures, field administration, scoring or analysis. These nonsampling, nonrandom errors rarely can be quantified, nor can the magnitude of the bias they introduce into our estimates be evaluated directly.

Systematic errors can be controlled in large part by employing uniform administration and scoring procedures and by requiring rigorous quality control in all phases of an assessment. If the systematic errors are nearly the same from age to age or group to group, then the differences in percentages or mean percentages are measured with reduced bias because subtraction will tend to cancel the effect of the systematic errors.

Similarly, the effect of systematic errors in different assessment years can be controlled by carefully replicating in the second assessment the procedures carried out in the first. Differences in achievement across assessment years will also be measured with reduced bias since subtraction will again tend to cancel systematic errors.

Although it is not possible for every condition or procedure to remain exactly the same between assessments conducted several years apart, National Assessment makes every effort to keep conditions as nearly the same as possible and to document any changes that are introduced.

CHAPTER 2

ANALYSIS COMPUTATIONS

Computation of Measures of Achievement, Changes in Achievement and Standard Errors

Measures of achievement are obtained by weighting individual responses appropriately. Reasonably good approximation of standard error estimates of these achievement measures can be obtained by applying the jackknife procedure to first-stage sampling units (replicates) within strata, using the method of successive differences and accumulating across strata.

In this section, the measures of achievement are first defined in algebraic form, followed by a description of the jackknife method used by NAEP to estimate standard errors.

Measures of Achievement

Based on the sample design, a weight is assigned to every individual who responds to an exercise administered in an assessment. The weight is the reciprocal of the probability, with adjustment for nonresponse, of selecting a particular individual who then takes a particular exercise. Since the probabilities of selection are based on an estimated number of people in the target age population, the weight for an individual estimates the number of similar people that individual represents in the age population.

A sum of the weights for all individuals at an age level responding to an exercise is an estimate of the total number of people in that age population. A sum of weights for all individuals at an age responding correctly to an exercise is an estimate of the number of people who would be able to respond correctly in the age population if the entire population were assessed. These concepts also apply to any reporting group (e.g., defined by region, sex and so on) and category of response (e.g., correct, incorrect and "I don't know").

Let W^{e}_{ihk} = sum of weights for respondents to exercise e who are in reporting subgroup i and who are in the kth replicate of the hth sampling stratum, and

C^{ej}_{ihk} = sum of weights for respondents to exercise e who are in reporting subgroup i, who are in the kth replicate of the hth sampling stratum, and who selected response category j (e.g., correct foil or a particular distractor) for the exercise.

Note that $W^{e}_{ihk} = \underset{j}{\text{Sum}}\ C^{ej}_{ihk}$

Then, summing k over the n_h sample replicates in the stratum h, and summing over the H sampling strata,

$$W^e_{i++} = \underset{h=1}{\overset{H}{\text{Sum}}}\ \underset{k=1}{\overset{n_h}{\text{Sum}}}\ W^e_{ihk}$$

estimates the number of eligibles in the population who are in subgroup i.

Similarly, $C^{ej}_{i++} = \underset{h=1}{\overset{H}{\text{Sum}}}\ \underset{k=1}{\overset{n_h}{\text{Sum}}}\ C^{ej}_{ihk}$

estimates the number of eligibles in the population who are in subgroup i and who would select response category j for exercise e.

An estimate of the proportion of the eligibles in the age population in group i who would select response category j on exercise e is:

$$P^{ej}_i = C^{ej}_{i++}/W^e_{i++} \ . \tag{1}$$

In the special case where the proportion of all age eligibles who would select response category j on exercise e is estimated, the index A (for ALL) will be used in place of i as follows:

$$P^{ej}_A = C^{ej}_{A++}/W^e_{A++} \ . \tag{2}$$

In National Assessment reports, the proportion in (1) multiplied by 100 is called the group percentage, and the proportion in (2) multiplied by 100 is called the national percentage. The difference between the proportion in subgroup i who would select category j on exercise e and the proportion in the nation is denoted by:

$$dP^{ej}_i = P^{ej}_i - P^{ej}_A \quad \text{and is called delta P.} \tag{3}$$

National Assessment also reports the arithmetic mean of the percentage of correct responses over sets of exercises corresponding to the measures in (1), (2) and (3). These means are taken over the set of all exercises or a subset of exercises classified by a reporting topic or content objective. The mean proportion of correct responses taken over m exercises in some set of exercises corresponding to measures (1), (2) and (3) are, respectively:

$$\overline{P}_i = \frac{1}{m} \text{Sum } C^e_{i++}/W^e_{i++} \ . \tag{4}$$

$$\overline{P}_A = \frac{1}{m} \underset{e}{\text{Sum}}\ C^e_{A++}/W^e_{i++} \ \text{and} \tag{5}$$

$$\overline{dP}_i = \overline{P}_i - \overline{P}_A \ . \tag{6}$$

Note that the response category subscript j has been suppressed since the means are understood to be taken over the correct response category for each exercise.

Each of these six achievement measures is computed and used in describing achievement data for any assessment. The simple difference in these measures between two assessments of the same exercise (or sets of exercises) provides six measures of change in achievement.

The next section describes how standard errors are estimated for the twelve statistics.

Computation of Standard Errors

In order to obtain an approximate measure of the sampling variability in the statistics (1) through (6), a jackknife replication procedure for estimating the sampling variance of nonlinear statistics from complex, multistage samples was tailored to National Assessment's sample design. Miller (1968, 1974) and Mosteller and Tukey (1977) provide information about the jackknife technique, while Folsom (1977) describes how the procedure is used in estimating standard errors for National Assessment's sample designs.

To demonstrate the computational aspects of this technique, consider estimating the variance of the statistic in (1)—the proportion of age-eligibles in subgroup 1 who would select response category j on exercise e.

This statistic is based on the data from all the n_h replicates in the H strata. Let p^{e}_{i-hk} be defined as a replication estimate of p^{ej}_{i} and constructed from all the replicates excluding the data from replicate k in stratum h. These replication estimates are computed as if the excluded replicate had not responded and a reasonable nonresponse adjustment is used to replace the data in replicate hk in estimating p^{ej}_{i}. Several choices for replacing the data in replicate hk are available. In order to obtain a convenient and computationally efficient algorithm for approximating standard errors, National Assessment replaces C^{ej}_{ihk} and W^{e}_{ihk} from the hkth replicate with corresponding sums from another paired replicate in the same stratum. The replicate estimate is then computed. The replicate estimates to be used in the calculations are determined by arranging all the replicates in each stratum into pairs. That is, replicate 1 is paired with replicate 2, replicate replicate 2 with replicate 3, 3 with 4, ... (n_h-1) with n_h and replicate nh with replicate 1.

The contribution to the variance of P^{ej}_{i} by each pair of replicates is the change in the value of the statistic incurred by replacing the data from each replicate in the pair with the data from the other replicate in the pair and recomputing p in the usual way. This produces two replicate estimates. Squaring the difference between these replicate estimates and then dividing by eight, produces a measure of the contribution of this pair of replicates to the total variance. The sum of these contributions over all n_h successive pairs in the stratum contributions is the estimate of the standard error of p_i^{ej}.

Algebraically, the two replicate estimates for the pair k, k+1 (where) k=1, ... n_h and n_h+1=1) are:

$$P_{i-hk}^{ej} = \frac{C_{i++}^{ej} - C_{ihk}^{ej} + C_{ih(k+1)}^{ej}}{W_{i++}^{ej} - W_{ihk}^{ej} + W_{ih(k+1)}^{ej}} \quad (7)$$

and

$$P_{i-h(k+1)}^{ej} = \frac{C_{i++}^{ej} - C_{ih(k+1)}^{ej} + C_{ihk}^{ej}}{W_{i++}^{ej} - W_{ih(k+1)}^{ej} + W_{ihk}^{ej}} \quad (8)$$

The contribution to the total variance from stratum h is:

$$\text{var}\ (P_{ih}^{ej}) = \frac{1}{8} \underset{k}{\overset{n_h}{\text{Sum}}} \left[P_{i-hk}^{ej} - P_{i-h(k+1)}^{ej} \right]^2 \quad (9)$$

And, finally, an estimate of the standard error of P_i^{ej} is:

$$\text{SE}\ (P_i^{ej}) = (\underset{h}{\overset{H}{\text{Sum}}}\ \text{var}\ P_{ih}^{ej})^{1/2} \quad (10)$$

Multiplying P_i^{ej} by 100 yields the percentage of response to category j. Multiplying SE(P_i^{ej}) by 100 yields the corresponding estimated standard error of the percentage.

In general, the jackknifed standard errors of the proportion estimates will be larger than the simple random sampling formula $(pq/n)^{1/2}$, where $p=p_i^{ej}$, q=1-p and n is the number of sampled respondents in subgroup i who took the exercise. The larger size of SE (p_i^{ej}) reflects mainly the loss of precision due to cluster-sampling of schools and students. The ratio of jacknifed standard errors to the simple random sampling standard error is defined to be the square root of the design effect. Design effects are often in the vicinity of 2.

The standard errors for the achievement measures (2) through (6) are computed through a series of steps analogous to those followed in computing SE (p_i^{ej}). The most complicated step in computing standard errors occurs in forming the paired replicate estimates analogous to (7) and (8) for each successive pair of replicates. Once this bookkeeping chore is done, the computations (9) and (10) follow in a straightforward manner.

The standard errors for the differences between two assessments for any of the achievement measures (1) through (6) are computed as the square root of the sum of the squared standard errors from each of the separate assessments.

The size of the standard errors depends primarily on the number of schools included in the sample, and on the number of respondents in each of the reporting groups.

The size of the standard errors of the means of the achievement measures for sets of exercises is also influenced by the number of exercises in the exercise set and the number of booklets over which the items in the set are spread. Our suggestions for simpler approximations for standard errors are summarized below.

Design Effect Adjustments. Virtually all common statistical packages assume simple random sampling in computing statistics and their associated variances. The estimated variances are systematically too small (we gain some precision through our stratification, but lose more through multistage sampling). At least for linear statistics—means, percentages, etc.—the variances tend to be too small by a factor of about 2, called a design effect. That is, our effective sample size is about one-half (1/design effect) of what it would have been with simple random sampling (SRS). Multiplying the SRS variance estimates by 2 (or the SRS standard error estimates by the square root of 2) yields estimates of about the appropriate magnitude.

For standard errors of proportions, the design effect correction works well when the proportions are between 0.30 and 0.70. Estimates for proportions outside that range tend to be systematically too small and a further correction is required:

For simple random sampling:

$$se_p = \sqrt{(P(1-P)/n)} \; .$$

For National Assessment data:

$$\text{if } .30 \leq p \leq .70$$

$$se_p = \sqrt{(2\,P(1-P)/n)} \text{ (design effect correction).}$$

If $P < .30$ or $p > .70$,

the above approach provides estimates that are too small. Averaging the above estimate with the estimate at the point .30 or .70 provides:

$$se_p = .324/\sqrt{(n)} + 1/2\sqrt{(2\ P(1-P)/n)}$$

$$= .324/\sqrt{(n)} + \sqrt{(P(1-P)/2n)}$$

Where

P = weighted proportion

se_p = estimated standard error of a proportion

n = sample size.

For standard errors of exercise percentages, substitute percentages for proportions in the above equations (note (1-P) becomes 100(1-P)) and change .324 to 32.4 in the last equation.

Estimated Standard Errors for Means of Exercise Percentages. Our most common summary measure is the average percentage of correct responses to a set of exercises. Examples include the average percentage of correct responses on all reading exercises at age 13 or the average percentage of correct responses to political knowledge exercises at age 17. The mean percentages are relatively easy to compute. The computation of their estimated standard errors are expensive and time consuming for estimates involving multiple booklets.

Over the past several years, we have studied the effect of several factors on the estimated standard errors. To date, the most important factors appear to be the total number of respondents and exercises, as well as whether the respondents are drawn from all sample schools (nation, sex, whites, etc.) or from subsets of sample schools (minorities, region, community size, etc.). Smoothed estimates of standard errors are presented in tabular form in Exhibit 2 and in graphic form in Exhibits 3 and 4.

The estimates were obtained from the 1977-78 mathematics assessment. Large numbers of jackknifed standard errors of means were regressed on functions of the number of exercises (NEX) and sample sizes (n). The best fitting models were:

a) For variables appearing in all schools

$$s.e. = 0.3633 + 0.4985/\sqrt{(NEX)} + 0.1280/NEX + 12.3807/\sqrt{(n)} + 119.0076/n$$

b) for variables appearing in subsets of schools

$$s.e. = 0.6797 + 0.1970/\sqrt{(NEX)} + 0.9964/NEX + 19.6587/\sqrt{(n)} - 7.7244/n.$$

The estimates in Exhibits 2 through 4 were generated with these equations.

The standard errors of mean percentages thus estimated are slightly too large for assessments conducted before 1975-76. Those assessments utilized more schools and fewer students per school.

The tables are easiest to use when results are expressed in percentage units, but they can be used in other situations. Consider for example an exercise which has three parts. We computed the percentage of correct responses to each part. The weighted percentages are:

	Percentage Correct
Part A	85.2
Part B	92.7
Part C	83.1

The average percentage of correct responses, 87.0, is for the nation, so we use the top table in Exhibit 12 and find the entry for three exercises and a sample size of 2500. The estimated standard error is 0.99 percent.

We also computed the mean number of correct responses. The mean is 2.61. We can use the tables to estimate the standard error for the mean number of correct responses in two equivalent ways. The average number of correct responses can be converted to an average percentage by dividing by the number of exercises (3) and multiplying by 100 to get 87.0 percent. Alternatively, the estimated standard error, 0.99, can be multiplied by 3 and divided by 100 to get a standard error of 0.0297 in "number correct" units.

This specific example is for three exercise (or exercise parts) in one booklet, a situation where design effect estimates are also feasible (but generally less precise). Had the three exercises (or exercise parts) come from two or more booklets, the table look-up would be the same, but the sample size would equal the sum of sample sizes for the two or more booklets.

EXHIBIT 2

Smoothed Estimates of Standard Errors of Means for Differing Numbers of Exercises and Sample Sizes for Variable Categories (A) That Appear in Most Schools and (B) That Do Not Appear in Most Schools

Variable Categories A. Nation, Sex, White, Parental Education, Grade, etc.

Sample Size

Number of Exercises	250	500	750	1000	1250	1500	2000	2500	5000	7500	10000	25000
3	1.95	1.49	1.30	1.20	1.14	1.09	1.03	0.99	0.89	0.85	0.83	0.78
5	1.87	1.40	1.22	1.12	1.06	1.01	0.95	0.91	0.81	0.77	0.75	0.69
10	1.79	1.33	1.14	1.04	0.98	0.93	0.87	0.83	0.73	0.69	0.67	0.62
25	1.73	1.26	1.08	0.98	0.91	0.87	0.80	0.76	0.67	0.63	0.60	0.55
50	1.70	1.23	1.05	0.95	0.88	0.84	0.77	0.73	0.64	0.60	0.57	0.52
100	1.67	1.21	1.03	0.92	0.86	0.81	0.75	0.71	0.61	0.57	0.55	0.50
250	1.65	1.19	1.01	0.91	0.84	0.79	0.73	0.69	0.59	0.55	0.53	0.48

Variable Categories B. Region, Type of Community, Size of Community, Minorities, etc.

Sample Size

Number of Exercises	250	500	750	1000	1250	1500	2000	2500	5000	7500	10000	25000
3	2.34	1.99	1.83	1.74	1.68	1.63	1.56	1.52	1.40	1.35	1.32	1.25
5	2.18	1.83	1.67	1.58	1.52	1.47	1.40	1.36	1.24	1.19	1.16	1.09
10	2.05	1.71	1.55	1.46	1.39	1.34	1.28	1.23	1.12	1.07	1.04	0.97
25	1.97	1.62	1.47	1.37	1.31	1.26	1.19	1.15	1.04	0.98	0.95	0.88
50	1.94	1.59	1.44	1.34	1.28	1.23	1.16	1.12	1.00	0.95	0.92	0.85
100	1.92	1.57	1.42	1.32	1.26	1.21	1.15	1.10	0.99	0.94	0.91	0.83
250	1.91	1.56	1.40	1.31	1.25	1.20	1.13	1.09	0.97	0.92	0.89	0.82

EXHIBIT 3

Smoothed Estimates of Standard Errors of Means for Differing Numbers of Exercises and Sample Sizes for Variable Categories That Appear in Most Schools, Such as Nation, Sex, Whites, Parental Education, Grade, etc.

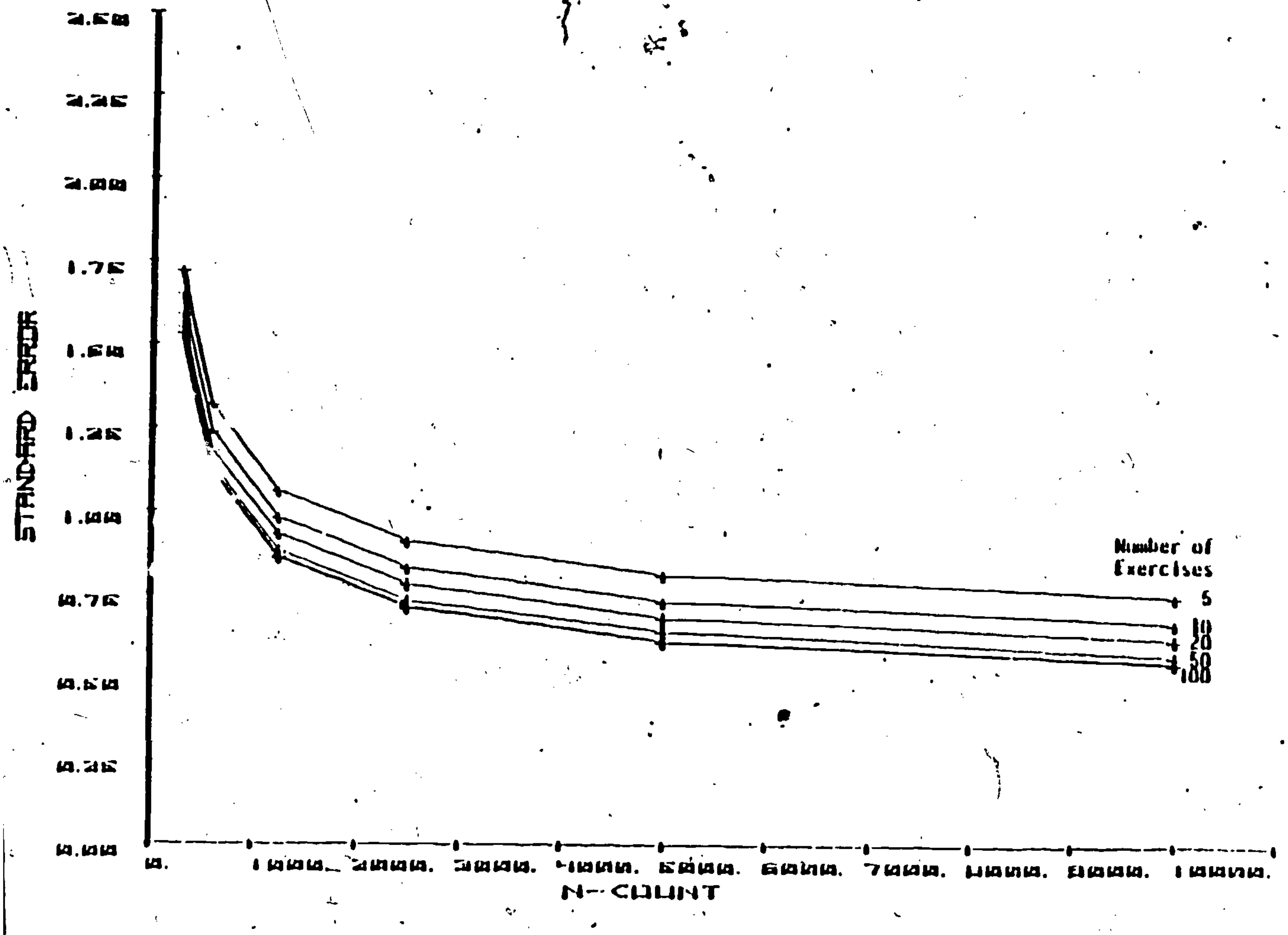
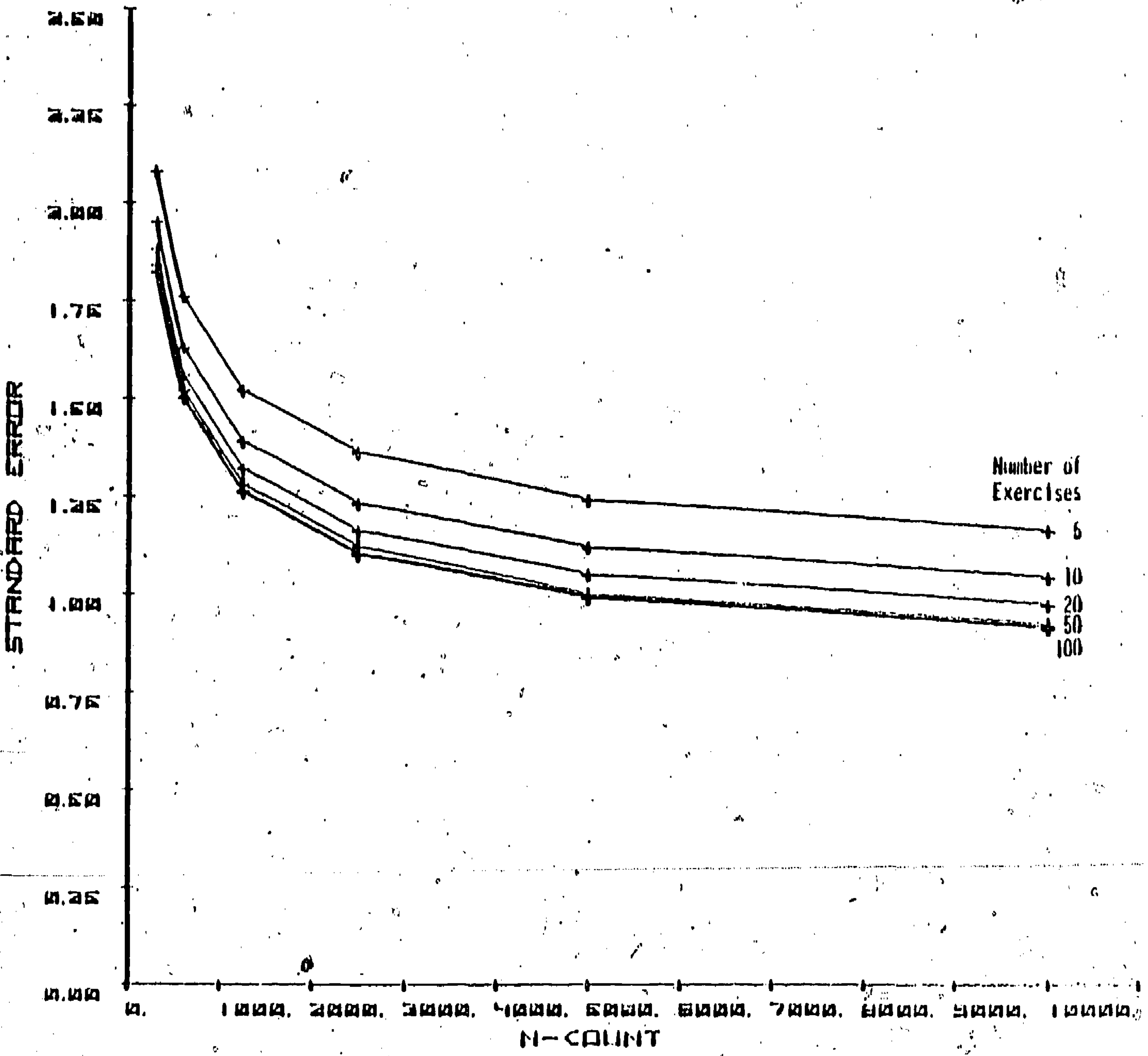


EXHIBIT 4

Smoothed Estimates of Standard Errors of Means for Differing Numbers of Exercises and Sample Sizes for Variable Categories That Do Not Appear In Most Schools, Such as Region, Type of Community, Minorities, etc.



CHAPTER 3

SECONDARY ANALYSIS POTENTIAL

Public-Use Data Tapes

Due to the broad coverage of learning areas and nationally representative samples of respondents, National Assessment's data base provides an unparalleled source of information for researchers. The complex sampling design and regulated testing conditions provide results that are generalizable to the nation as well as to many subgroups. Since 1979, National Asesssment has made a major effort, with National Institute of Education and National Science Foundation funding, to create public-use data tapes. Over 400 national probability samples for approximately 2500 respondents each are available, with each sample providing data on 150-250 variables per respondent.

The tapes contain complete respondent information except for identifiers (such as school, district, county, state, etc.) that violate Privacy Act provisions and/or confidentiality agreements.

Documentation is provided in machine-readable form on the tapes. It includes procedural and technical documentation, cross-reference lists and detailed codebooks. OSIRIS-IV, SPSS and SAS-readable documentation files are also included. The SPSS and SAS files contain all information required to create labeled systems files and to begin analysis. Copies of exercise booklets, exercise scoring guides and other data collection forms are provided on 24x microfiche. Researchers are required to sign a nondisclosure agreement stating that they will not publish the exact contents of secure items.

In general, a separate public-use data tape is provided for each age level in an assessment year. Each tape contains several data files, one for each item booklet; there are two to sixteen booklets per age group per assessment. Each file typically contains responses to 25 to 30 attitudinal and achievement questions for a separate national probability sample of about 2500 respondents from public and private schools (ages 9, 13 and 17) or households (young adults).

Additionally, each file contains numerous background variables at the school and respondent level. Background variables common to all data files are listed below.

- School-level variables. Included are region, census division, type and size of community, occupational mix of attendance area, grade range, racial composition. Total enrollment and ESEA Title I eligibility.

- Respondent-level variables. Included are age, sex, race/ethnicity, grade, parents' education and reading materials in the home. From 1972-73 on, regional migration variables are included for the older age groups. From 1975-76 on, 17-year-olds were asked a number of additional background questions, including: homework and TV viewing habits, languages spoken at home and self-identified racial/ethnic heritage.

Methods of data analysis that adjust survey data for disproportionate representation of one subgroup within another are presented in Appendix 4, Weighting Class Adjustments, when interest is in one comparison and in Appendix 5, Balancing, when interest is in multiple comparisons.

Interested researchers are encouraged to obtain the brochure "Public-Use Data Tapes" (NAEP publication SY-DT-36), which includes more detailed information on tape contents, file contents, learning area specifics, tape characteristics and how to order data tapes. An other recommended document is the "Introduction to the National Assessment of Educational Progress Public-Use Data Tapes," which users receive with any data tape purchased. This document describes variable documentation and technical considerations and presents examples of analyses using SPSS and SAS files.

Analyses with Public-Use Data Tapes

The public-use data tapes allow access to information on American educational achievement in various content areas for selected age groups. Research may focus on methodological issues, hypothesis-testing or model-testing, descriptive studies or policy relevant studies. While these categories are not mutually exclusive, the following list provides some examples of the breadth of topics that may be explored.

- Methodological. Design effects, item response theories, item characteristics (such as format and readability), bias, effects of guessing, response patterns (see Appendix 9).

- Hypothesis-Testing and Model-Testing. Relationship of achievement to attitudes and experiences, relationship of achievement to school characteristics, exploration of education and psychological models.

- Descriptive and Policy Relevant. Trends in achievement over time, analysis of groups with special needs (such as bilingual, minority and low achievers), analysis of special topics (such as television watching habits, reading habits and use of calculators and computers).

A number of published studies have already been conducted using public-use data tapes, including nine that were supported by small contracts from the National Institute of Education. See "Exploring National Assessment Data through Secondary Analysis" (NAEP publication AY-SA-50.)

Analysis Complexities Due to Sample Design

The purpose of National Assessment is to survey the educational attainments of population groups instead of testing and ranking individuals. Multi-stage probability samples with disproportionate sampling rates are used to collect high quality data at an affordable cost. Further cost effectiveness is achieved by partitioning each age-specific sample (9, 13, 17 or young adults) into several randomly equivalent national probability samples. Each sample is given a different booklet of exercises (items), and its data are contained in a separate data file. These design features create numerous analysis complexities, and not all of them are intuitively obvious. The following describes some of the major consequences:

- Analyzing all exercises for a particular topic (such as mathematical computation skills) means analyzing multiple data files because the exercises have typically been spread across as many different booklets/samples/data files as possible. This practice has many sound reasons; however, it is not common practice in educational surveys.

- As in most large-scale surveys, respondents have differential selection probabilities. Consequently, most analyses require the use of sampling weights, even for exploratory analyses.

- Variance estimation is not straightforward for data from complex survey designs. Running the data through standard routines that assume simple random sampling procedures produces erroneous estimates of standard errors, confidence intervals and probabilities of Type I errors. National Assessment's approach to variance estimation is described previously.

- Each sample data file contains data for a separate, national probability sample, but the samples are not strictly independent. All student samples in a particular assessment are drawn from a common set of primary sampling units and, within any one age, students are sampled from intersecting sets of schools. For purposes like cross-validation, it is better to split one sample in half by randomly sampling half of the primary sampling units. See "Introduction to the National Assessment of Educational Progress Public-Use Data Tapes" (NAEP report SY-DT-50).

- Analysis at the student level across booklets requires the use of equating procedures. See Appendix 8 for a simple procedure.

Additional discussion is contained in Appendix 2.

Data Base Linkages

Relating NAEP Data to Other Data Bases and Studies. Measuring educational achievement with the resources allocated to NAEP can provide accurate estimates of achievement. However, data collected by other programs can be used to paint a more complete picture of educational progress. For example, internal state, district and zip code information used for sampling can link NAEP data with socioeconomic information provided by Census summary tapes. NAEP can provide results relating achievement to many home and community variables. These analyses would require confidentiality guarantees from the researcher.

In addition, the National Center for Educational Statistics (NCES) periodically collects information that could be related to NAEP data at the school district level. Various fifty-state data bases can be linked to NAEP data. Data on school finance information at the local level, for example, could be used to cluster communities according to various financial indicators such as relative wealth, tax effort, source of revenues, expenditure per pupil, number and type of students in a district, and number of schools in a district. These profiles of similar districts could permit reporting on achievement levels for different district configurations.

NAEP might also ask questions in assessments to relate its data to data obtained in other large-scale testing projects. Finally, indirect relationships can be established using meta-analysis techniques to link NAEP data with data from various large-scale educational studies (Glass, McGaw and Smith 1981).

Examples of economic, education programs and other policy issues that could be related to achievement are listed in Appendix 7.

APPENDIX 1

DEFINITIONS OF NATIONAL ASSESSMENT REPORTING GROUPS

In addition to reporting change results for all 9-, 13- and 17-year-old students in the United States, National Assessment reports results for a number of population subgroups.

Definitions of the key subgroups follow.

Region

The country has been divided into the four office of Business Economic regions: Northeast, Southeast, Central and West.

Sex

Results are reported for males and females.

Race

Results are presented for blacks, whites and Hispanos. (Data for Hispanos are reported only for sets of exercises, not for individual items, because of small sample sizes.)

Level of Parental Education

Three categories of parental-education levels are defined by National Assessment, based on students' reports. These categories are: 1) those whose parents did not graduate from high school, 2) those who have at least one parent who graduated from high school, and 3) those who have at least one parent who has had some post-high-school education.

Type of Community

Communities in this category are defined by an occupational profile of the area served by a school as well as by the size of the community in which the school is located.

Advantaged-urban Communities

Students in this group attend schools in or around cities having a population greater than 200,000 where a high proportion of the residents are in professional or managerial positions.

Disadvantaged-urban Communities

Students in this group attend schools where a relatively high proportion of the residents are on welfare or are not regularly employed which are in or around cities having a population greater than 200,000.

Extreme-rural Communities

Students in this group attend schools in areas with a population under 19,000 where many of the residents are farmers or farm workers.

Size of Community

Big Cities

Students in this group attend schools within the city limits of cities having a 1970 census population over 200,000.

Fringes Around Big Cities

Students in this group attend schools within metropolitan areas (1970 U.S. Bureau of the Census urbanized areas) served by cities having a population greater than 200,000 but outside the city limits.

Medium Cities

Students in this group attend schools in cities having a population between 25,000 and 200,000, not classified in the fringes-around-big-cities category.

Small Places

Students in this group attend schools in communities having a population less than 25,000, not classified in the fringes-around-big-cities category.

Grade in School

Results are categorized for 9-year-olds in the third or fourth grade; 13-year-olds in the seventh or eighth grade; and 17-year-olds in the tenth, eleventh or twelfth grade.

Modal Grade by Region

Results are categorized for 9-, 13- and 17-year-old respondents in grades four, eight and eleven, respectively, who live in the Northeastern, Southeastern, Central or Western regions of the country.

Modal Grade by Community Size

Results are categorized for 9-, 13- and 17-year-old respondents in grades four, eight and eleven, respectively, who live in big cities, fringes around big cities, medium cities and small places.

Modal Grade by Sex

Results are categorized for 9-, 13- and 17-year-old males and females in grades four, eight and eleven, respectively.

Achievement Class

The achievement class variable classifies each respondent for a given age and year into one of a number of categories (achievement classes) based on the respondent's estimated standing in the population in terms of achievement. The classes form a partitioning of the population by achievement and are defined as follows:

Class 1

The lowest fourth (students in the lowest 25 percent of the population on achievement).

Class 2

The next-to-lowest quarter (achieve higher than 25 percent and lower than 50 percent of the population).

Class 3

The next-to-highest quarter (achieve higher than 50 percent and lower than 25 percent of the population).

Class 4

The highest fourth (students in the upper 25 percent of the population on achievement).

For our purposes, the measure of achievement for an individual will (generally) be the person's package score (mean percent correct) over all items from a particular subject area, such as reading, science, mathematics and so on. The package score for an individual is the ratio of the number of correct responses over the number of items (including items where the "I don't know" foil was selected).

APPENDIX 2

INFERENCE FROM SURVEY DATA AND ANALYSIS OF DATA FROM COMPLEX SAMPLE SURVEYS

NAEP is a sample survey that collects data by using a multistage design with unequal probabilities of selection of elements. Such a survey design is commonly referred to as "complex" sampling. It permits the collection of representative data for the population and many subgroups in an extremely cost effective manner. NAEP data are suited for various descriptive purposes.

1. Estimation of achievement by exercise and summary level for the age populations as a whole and for a variety of subgroups of the populations.

2. Estimation of changes in achievement on a variety of tasks (and summaries) over time. Change can be measured for populations and for subgroups.

3. Explorations of observed associations between variables of interest.

The various descriptive statistics should be computed taking the structure of the sample into account.

The NAEP data present certain difficulties in analysis which are shared by all complex surveys. These difficulties include problems in inference due to the type of data and complications in analysis due to the sample design. These difficulties are detailed in the following sections.

Inference from Survey Data[1]

In education, as in many other fields "causation" has to be given a simple pragmatic meaning:

- 'If we change this, in this direction, will that change in that direction (and can we judge by how much?)

Generally no single variable's change will "cause" changes in the behavior that interests us to the exclusion of effects from changes in other variables.

[1] This section written by John Tukey.

Indeed, changing almost any variable probably changes any behavior that interests us, though the amount of change may be miniscule, unimportant and undetectable.

The data collected from any survey including those collected by NAEP, are observational and not experimental--the values of the factors studied having arisen by a relatively complex natural process, rather than having been chosen by an experimenter with the aid of randomization. As a result, survey data can directly demonstrate only association--that under these circumstances we see more (or less) of this behavior. While association may suggest causation, it is rare indeed that we can be sure about causation because of any simple argument mainly based on association. (A classic example is that fires attended by more fire engines tend to involve larger financial losses. The association is well-established and wide-spread, but who would believe that more fire engines cause more loss?). This need for care in interpreting association has nothing to do with whether or not statistical procedures have been applied. (In the fire engine case, the existence of 12 studies, in each of which the association was significant beyond 1%, would do nothing to make causation less implausible. All it could do is to give formal documented evidence that this particular association that we all believe in actually exists.)

Having established, at least roughly, how much association is present, it is then our responsibility to consider a variety of ways in which it might have arisen (thus an association between better student performance and higher teacher's salaries could have arisen (a) because better schools encourage communities to support their schools better or (b) because communities with more interest in schools develop more interest in schools among their children, hence better performance, and also pay their teachers better, as well as because (c) better pay for teachers attracts better teachers whose teaching leads to better performances. Then we can test each of these possible patterns of causation against our general knowledge and other kinds of information. As a result, we may

- feel moderately secure about some one pattern of causation, but more likely
- come to see what kinds of studies might help elucidate what is happening.

Complications in Analysis Due to Sample Design

The multistage design of the NAEP sample causes complications in analysis. Many standard statistical procedures assume that data are acquired by means of a simple random sample of the population and that individuals are independent. These assumptions are not met since the NAEP sample employs stratification, clustering and unequal probabilities of selection.

Certain subgroups of the population are sampled at a higher rate than the remainder of the population in order to ensure adequate representation. Consequently, those subgroups, which tend to have different characteristics (including achievement) than the remainder of the population, are overrepresented in the sample. Analyses that ignore this are apt to produce biased and misleading information, since those groups may have unwarranted impact. This difficulty is avoided by conducting weighted analyses, in which the weight assigned to an individual is related to the reciprocal of his or her probability of selection.

Because of cost and administrative efficiency considerations, NAEP data are obtained by selecting a number of schools and then selecting a number of students within each of the schools. Since the students are selected in clusters, observations from various students are not independent. Student responses within a school tend to be relatively more homogeneous than student responses from different schools.

Ignoring the effect of stratification and clustering in analysis tends to produce severe underestimates of the variability of statistics. Many studies (such as Ross, 1976; Kish and Frankel, 1974; Frankel, 1971) have demonstrated large influences of complex survey designs on sampling errors of various statistics, such as regression and correlation coefficients. It has been demonstrated (Shah, Holt and Folsom, 1977) that regression analyses of data from complex survey samples produce tests of significance that are generally too liberal when the structure of the sample is not taken into account. Additionally, as noted by Fellegi (1979) in his consideration of goodness of fit tests, the distribution of certain statistics, as well as their dispersions, can be affected by the sample design.

Because of the nonlinearity of many of the statistics of interest, it is not possible to estimate standard errors in closed form. However, several procedures do this approximately. Among these are:

1. jackknifing--the procedure used by NAEP (detailed by Folsom, 1977);

2. balanced repeated replications (detailed by McCarthy, 1969); and

3. Taylor series approximation (detailed by Folsom, 1977).

Shah, Holt and Folsom (1977) give procedures for estimation and hypothesis testing of regression models for data from a complex sample survey using Taylor series approximations. A general procedure for obtaining approximate variances by Taylor series approximations is given by Woodruff and Causey (1976).

Fellegi (1979) and McCarthy (1969) give procedures for using balanced repeated replications to conduct goodness of fit tests.

Folsom (1977) gives the procedures used by NAEP in its analyses using jackknifing methodology.

A comparison of the performance of the three procedures for means, variances, correlations and regression coefficients is given by Frankel (1971) and Kish and Frankel (1974).

APPENDIX 3

ADJUSTMENT OF RESPONDENT WEIGHTS BY SMOOTHING

Background

A weight is assigned to every individual who responds to an exercise administered in an assessment. The weight is the reciprocal of the probability of selection of the individual with adjustment for nonresponse, and the probabilities of selection are based on the estimated number of people in the target age populations; therefore, the weight for an individual estimates the number of similar people that that individual represents in the age population. The sum of the weights of all individuals at an age level responding to an exercise is an estimate of the total number of people in that age population in the year that the exercise was assessed. Similarly, the sum of weights for all individuals who took the exercise and who also are members of some demographic category (such as blacks) gives an estimate of the number of people in the age population, for the year, who are also members of the category. The ratio of the two totals estimates the proportional representation of the demographic category in the age population for the given year.

For each of the assessed age populations in each assessment year, separate estimates of the proportional representation of the various demographic subgroups are provided by each booklet administered to that age group in that year. Due to random sampling variability, the estimates of population proportions for a given year based on single booklets administered in the year will vary. In addition to any trends in population proportions over time, there is also random sampling variation in these proportions from year to year.

It is desirable to reduce the random variability of population proportions as much as possible, since this variability has an effect on performance estimates. For example, the percentage of acceptable responses for an age group is a function of the relative proportion of high-performing and low-performing groups. If the relative proportions of these groups are very different in different assessments due to sampling variability, then a portion of the change in percentage of acceptable responses for an age group might be attributable to yearly sampling difference in the relative proportions of high- and low-achieving groups.

In addition to reporting performance estimates for an age group as a whole, National Assessment also reports performance for various subpopulations, such as whites or blacks. Because variability of subgroups sizes within these subpopulations (such as males and females within the white subpopulation) influences the performance estimates for the subpopulations, it is desirable that fluctuations of proportions of all subgroups of each subpopulation be reduced as much as possible.

For each age and year, each of the various booklets administered will provide estimates of a given population proportion. Since these estimates are subject to booklet-to-booklet variability, a better estimate of the population proportion, which will have reduced variability, is obtained by combining the information from all booklets. However, these proportions vary from year to year due to random sampling variability or systematic differences in sampling procedures. An even better estimate of population proportions for any single year can be obtained by smoothing the proportions over several assessment years. The word "smoothing" is used here in the sense of fitting a smooth curve to a sequence of numbers by robust/resistant procedures. Smoothing estimates of population proportions reduces a large portion of the sampling variability while preserving, as far as possible, actual trends occurring in the age population.

After the population proportions have been smoothed, adjusted weights are derived for the assessed individuals so that the population proportions computed using the adjusted weights are equal to the smoothed proportions. The adjusted weights are then used for all analyses.

Smoothing Procedures Used by National Assessment

The most direct way to smooth proportions is to classify people into mutually exclusive multiway cells on the basis of their membership in categories of various important variables and then to smooth the proportions within each of the resulting multiway cells across years. Unfortunately, this procedure tends to produce a large number of cells with few people and, consequently, quite unstable estimates of smoothed proportions.

To circumvent this difficulty, National Asessment has utilized various smoothing procedures since the 1976-77 assessment. Each of these procedures, which are all basically weighting-class adjustments applied independently to each age, is designed to control, to varying degrees, fluctuations in certain key subgroups while avoiding, as much as possible, instabilities due to small cells.

The procedure used for the 1976-77 assessment was a weighting-class adjustment applied independently to each age and reporting variable (nation, region, sex, and so on). The details of the procedure are given in Appendix B of technical report 08-S-21: Three Assessments of Science, 1969-77: Technical Summary (1979). While this procedure performs well, it is complicated and requires large amounts of time and computer resources to implement. By independently smoothing proportions within each reporting variable, it was possible to produce good estimates of the marginal proportions of people within each category of the variable while disturbing as little as possible the relationships between other reporting variables within the adjusted variable. However, this means that each individual had a different adjusted weight for each reporting variable under consideration. While this presents no problem for the estimation of performance within a reporting variable, the multiplicity of weights definitely complicates any analyses, such as regression, that involve several variables.

Because of the complexity of the procedure used in 1976-77, a different and simpler procedure was adopted in 1977-78. This procedure is detailed in Appendix F of Report 90-MA-40, Procedural Handbook: 1977-78 Mathematics Assessment (1980). The 1977-78 smoothing procedure produced a single adjusted weight for each individual, and hence greatly reduced the complexity of subsequent analyses of performance data. The 1977-78 procedure involved applying a weighting-class adjustment independently to each age. The weighting classes, which were different at each age, consisted of individuals who were alike on certain demographic characteristics and who would be expected to have similar educational achievement characteristics. There were around seventy adjustment cells used for each age.

Although the 1977-78 procedure produced acceptable results, in 1978-79 National Assessment adopted yet another procedure that we believe has the best characteristics of the three procedures used. The 1978-79 procedure, which is detailed below, has several advantages.

1. It produces a single adjusted weight for each individual.

2. It affords good control on the distribution of proportions of certain key variables.

3. It produces the greatest stability of performance estimates.

4. It is the easiest to implement.

The Current Smoothing Procedure

The first step in the smoothing procedure involved the partitioning of the population of age class eligibles into the six smoothing cells given in Exhibit 1. The same cells were used for all ages.

Exhibit 1. Smoothing Cells

Cell	Race	Region	Community Size	(CS)
1	White	All	Big City +Fringe	(BC +FR)
2	White	All	Medium City	(MC)
3	White	All	Small Places	(SP)
4	Black	SE	All	
5	Black	Not SE	All	
6	Other	All	All	

Then, for each age and every year, the proportion of the population in each of the cells was estimated. For a given age and year, the proportion of the population in a particular cell was computed as the sum of weights of all respondents assessed in the given year who were of the specified age and who belonged in the cell, divided by the total of the weight of all respondents of the given age assessed in that year.

Each of the six cells was comprised of a sequence of estimated population proportions corresponding to the various years of assessment. Each such sequence of proportions was then smoothed by fitting robust/resistant lines (See Tukey, 1977.) Using data from the U.S. Census and Current Population Surveys, trends in enrollment by age and race and by age and region were obtained. The data from these surveys were adjusted to correspond with NAEP definitions as much as possible. The resistant lines within the smoothing cells were constrained to satisfy the trends from the U.S. Census and Current Population Surveys data.

The final step in the smoothing procedure was to adjust the respondents' weights to be consistent with the smoothed proportions. Since each respondent takes only one booklet, the weight adjustments were done independently for each booklet. For a given age, year and booklet, population proportions using the original weights were obtained for each of the smoothing cells. Then the weights of all respondents within a given cell were multiplied by the ratio of the smoothed cell proportion to the proportion using the original weights. This produced the adjusted weights that are used in all analyses.

To adjust respondent weights to be consistent with the smoothed proportions, the following procedure was employed:

1. For each booklet, classify the respondents according to smoothing cell and obtain the raw population proportions for each cell. For example, the raw proportion for a booklet of 9-year-olds in smoothing cell four is the total of the weights of all 9-year-olds in the booklet who are black and in the Southeastern region, divided by the total of the weights of all respondents to the booklet.

2. For each booklet and smoothing cell, obtain a weight adjustment factor as the ratio of the smoothed population proportion (for the appropriate age, year and smoothing cell) over the raw population proportion.

3. The adjusted weight for an individual is the product of that individual's original weight and the appropriate adjustment factor.

Changes in Smoothed Proportions as New Assessments Are Completed

Every time an assessment is completed, a new time point is added to each of the sequences of population proportions within the smoothing cells. This means that, even though robust/resistant procedures are used, the addition of a new point may somewhat change the values of smoothed proportions for prior years. Additionally, any changes in methodology will have an impact on the estimates.

This means that the smoothed proportions, obtained after the addition of the next assessment data, are apt to differ somewhat from the corresponding smoothed proportions without the new data.

APPENDIX 4

WEIGHTING CLASS ADJUSTMENTS

In order to compare achievement effects for two subgroups of interest when proportions of favorable or unfavorable background characteristics differ considerably, it is frequently helpful to adjust the effective distribution of weights in the two subgroups in such a way as to more closely balance these background characteristics.

The most direct way to do this is to employ what is commonly referred to as a weighting class adjustment. First, classify students into mutually exclusive multiway cells on the basis of their membership in categories of important variables. Small sized cells can be combined with larger ones if necessary, zero cells are to are to be avoided at all costs.

Corresponding multiway cells for the subgroups being compared can then be matched (adjusted) as follows:

Let W_{1is} be the weight for the sth student in the ith multiway cell for subgroup 1 and W_{2is} be the weight for the sth student in subgroup 2. The sum of weights over all students in the ith multiway cells is designated as $W_{1i.}$ and $W_{2i.}$ respectively. $W_{1..}$ and $W_{2..}$ represent summations of weights over all cells.

What is needed is an adjustment factor K_i for the ith cell in subgroup 1 such that

$$K_i W_{1i.}/W_{1..} = W_{2i.}/W_{2..}$$

This will equate the proportionate representation of the ith cell of subgroup 1 to the proportionate representation of the corresponding cell in subgroup 2 so that the composition of the first subgroup will "look" like the composition of the second subgroup in terms of the characteristics used to develop the multiway cells. In effect we standardize by creating a sample from a hypothetical population that had the same composition as subgroup 2.

In order to find the K_i solve

$$K_i = W_{2i.} W_{1..}/W_{1i}\ W_{2..}$$

and adjust the weights of the student data in the ith cell as:

$$W'_{1is} = K_i W_{1is}$$

Computation of exercise level and mean performance across exercises can then be completed with the new weights. Subgroup comparisons are then made for the hypothetical situation that the two groups have the same composition in terms of other background characteristics.

APPENDIX 5

BALANCING

The reporting categories National Assessment uses were selected for their interest and because they reflect differences in achievement occurring in the population: the age levels mark the end of primary, intermediate and secondary education; regions and sex groups have traditionally shown differences in educational attainments; school districts are thought to vary with the size and type of community (STOC) they serve; and level of parental education (PED) and RACE are believed to differentiate socioeconomic and home and family environments.

The percentage of any one group from one of these five categories responding acceptably to an exercise and the difference between the group percentage and that of the whole age group (group effect) are estimates of performance as it exists in the population. These percentages are estimates of the proportion of people who can respond correctly to an exercise and are facts about the population. Balancing is an adjustment procedure intended to add meaning to these facts by obtaining modified numbers, not to alter these facts.

Interpretations based on a comparison of one group effect with another or adding together group effects from different categories may be misleading if unadjusted percentages are used. The fact that observed group effects reflect Northeast of Southeast regional performances does not mean that these performances occur solely because the respondents live in the Northeast or Southeast. For example, a larger fraction of respondents in large cities live in the Northeast than in the Southeast. Similarly, a larger fraction of respondents in rural areas live in the Southeast than in the Northeast. Consequently, effects associated with size and type of community may be masquerading as part of an unadjusted regional effect. Simiarly, persons whose parents went beyond high school are more frequent in high metro communities than in the country as a whole. If persons having parents with no high school education do poorly, then we expect persons in the extreme rural areas will do poorly also. Adding together unadjusted group effects would "double count" those who do well in the first example and those who do poorly in the second.

Balanced Group Effects

The purpose of data adjustment procedures is to reveal information that cannot be seen in the unadjusted form. Balancing, in particular, is intended to remove the masquerading of one group effect as another and to avoid "double counting" of individuals. The procedure determines "balanced group effects" which can be used to compare one group with another in the same category and to add together group effects from different categories. Both masquerading and double counting result from persons of one group being disproportionately represented in other groups. This disproportionality existing in our nation's population is accurately reflected in our weighted data. Ignoring the specific details of our sampling and estimation procedures, note that the total weight of all individuals in a particular sample is an estimate of the number of individuals belonging to the corresponding group in the population. An unadjusted weighted percentage of success for an exercise is computed by dividing the sum of weights of the individuals in the group responding acceptably to the exercise by the sum of weights of all individuals attempting the exercise. The unadjusted percentages directly reflect the disproportionality existing in the population.

Balancing is an adjustment which simultaneously "balances," for each group and category, the disproportionate representation of the other groups that exists in the population. Interpretations of balanced group effects can be guided by thinking of a "conceptual" balanced population where the fraction of each group of one category occurring in each group of the other categories is the same as the fraction of each group occurring in the whole age population. For example, approximately 87 percent of all 9-year-olds in the nation are non-black and about 13 percent are black. In the Southeast, however, only 74 percent are non-black and about 26 percent are black. In the "conceptual" balanced population, the Southeast region would have effectively the same proportion of non-black and black as the nation.

Consider the effect of balancing. If persons with parents with post-high-school education do well and are more frequent in one region (than in others), while persons with parents having no high school education do poorly and are more frequent in another region we would expect the balanced effect of the first region to be less than its unadjusted effect and the balanced effect of the second region to be less negative than its unadjusted effect. If the magnitude of a group effect is substantially reduced by the balancing adjustment, one might conclude that the group itself may not be what causes the unadjusted differences but that substantial parts of these differences come from the unbalanced representation of the other variables.

Limitations of Balanced Group Effects

There are three basic kinds of limitations of balanced results. The first concerns interpretation. The group names NAEP uses for data analysis are labels standing for that factor indicated by its name and for a variety of other factors National Assessment did not (or could not) measure--factors associated with the named factor. As a survey, and not a controlled experimental study, National Assessment produces unadjusted group effects which cannot attribute cause to any particular factor named. Like unadjusted results, balanced group effects do not show what is caused by the labelled factor. They show only what part of the unadjusted effect can be conveniently named and attached to a group for bookkeeping purposes. They can show classes of individuals who perform differently, free of masquerading by other measured factors and related double counting. Once these differences are identified, additional information, or perhaps expert judgment, is required to find the cause or causes of these differences.

Balancing of National Assessment data has been limited to the groups: Region, Sex, STOC, RACE and PED. The interrelation of the wide variety of factors associated with or determining educational achievement is only partly known. Some important factors may not be represented in any clear way in our factors and others may not be represented at all. Factors may exist which are more sensitive or have smaller "proxy" bundles of other factors. Clearly, balancing becomes more useful as we are able to identify potentially "more important" factors and are able to measure them better.

The third type of limitation is concerned with the balancing model we have used. The balancing procedure utilizes an additive model which emphasizes balancing of marginal group effects and ignores balancing on combinations of groups. For example, the fraction of blacks living in rural areas in the Southeast is greater than the fraction of blacks living in rural areas in the Northeast. If rural Blacks living in the Southeast do poorly/ compared to all blacks living in rural areas, then we would expect the balanced Southeast region (and, of course, the unadjusted region) to do poorly. Thus, problems of masquerading and "double counting" also exist for balanced marginal effects, resulting from disproportionality of combinations of groups. Similar disproportionate representation exists for the other two-, three- and four-way group combinations.

The Balancing Procedure

The final algebraic form for the "conditions for balance" as stated by Tukey (1970) can be written for the 21 balanced group effects--region (4), sex (2), STOC (7), Race (3) and PED (5)--as follows:

$$\sum_{\substack{\text{except}\\ \text{one}}} n_{ijk\ell m}(P_{nat} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_k + \hat{\theta}_\ell + \hat{\phi}_m) = \sum_{\substack{\text{except}\\ \text{one}}} C_{ijk\ell m},$$

Where the sum is taken over all indices except one and Greek letters denote the balanced group effects corresponding to the groups denoted by the indices $i=1,\ldots,4$; $j=1,2,$; $k=1,\ldots,7$; $\ell=1,2,3$; and $m=1,\ldots,5$ belonging to Region, Sex, STOC, Race and PED respectively. P_{nat} is the overall national percent correct for the age group, $n_{ijk\ell m}$ is the weighted number of observations in each cell, and $C_{ijk\ell m}$ is the weighted number in each cell responding correctly to the exercise. The balancing condition above generates a single equation for each value of the omitted subscript in the summation. Thus, a solution to the simultaneous set of 22 equations (4+2+7+3+5 = 21 plus one equation for the sum over all indices) gives a set of fitted balanced group effects. This set of equations is, however, not of full rank and cannot yield unique balanced effects directly. The number of linearly independent equations is 17, but the set can be built up to full rank by appropriately replacing five of these equations with usual side conditions conventionally imposed when an additive linear model is fitted to a multiway crossed classification of the data, namely

$$\sum_i n_{i\cdots\cdot}\hat{\alpha}_i = \sum_j n_{\cdot j\cdots}\hat{\beta}_j = \sum_k n_{\cdot\cdot k\cdot\cdot}\hat{\gamma}_k = \sum_\ell n_{\cdot\cdot\cdot\ell\cdot}\hat{\theta}_\ell = \sum_m n_{\cdot\cdot\cdot\cdot m}\hat{\phi} = 0,$$

where the "dot" denotes the sum over the replaced subscript. Thus, a solution to the independent set of equations results in a unique set of balanced group effects.

The balancing equations can be shown to be algebraically equivalent to the usual set of normal equations resulting from minimization of the error sum of squares for the five-factor additive linear model

$$y_{ijk\ell mr} = P_{nat} + \alpha_i + \beta_j + \gamma_k + \theta_\ell + \phi_m + e_{ijk\ell mr}$$

where $y_{ijk\ell mr}$ is the weighted response for the rth person in the $ijk\ell m$th cell and $e_{ijk\ell mr}$ is the error associated with the rth person. The solutions to the normal equations (the balanced group effects) are, at least, simple least squares estimates. If the variability of the percentages in the five-way cell combinations happens to be proportional to the reciprocal of the corresponding cell weights, then the balanced group effects are minimum variance estimates (Tukey 1970 memo).

The condition for balance follows directly from the normal equations since each line shows that a linear combination of the balanced effects equals the number of estimated group successes in the population. Thus, the fitted number of successes is equal to the observed weighted number of successes.

Adding new variables to the balancing model (or replacing existing variables in the balancing set) offers no problems: if the additional groups are proportionately represented across the original groups, they will have no effect on the original balanced group effects. If they are disproportionately represented in the original groups, their addition to the balancing model can increase or decrease original balanced group effects. Masquerading of one group effect as another is uncovered as new variables are added to the balancing model.

Thus substantial masquerading of the effect of a newly added variable as an effect of a previously included variable is likley to be revealed by a substantial decrease of the balanced effect of that previously included variable.

Balanced Interactions

The normal equations indicate the potential use in balancing interactive effects. It is theoretically possible to solve the set of normal equations corresponding to the full interactive model. For this model, there are 2,880 normal equations with rank 840. It is not an easy task, however, to find the necessary side conditions to build full rank or to solve a set of equations of that order.

This number of equations is unmanageable and costly since a balanced solution must be obtained for each exercise. A reasonable compromise might be to select a manageable number of interactive group effects to include in the balancing equations recognizing that some important interactive effects might be left out.

Another difficulty occurs because we do not have any observations in our sample for some of the multi-group combinations. This is important since the condition for balancing requires that the observed number of successes is equal to the balanced (fitted) number of successes for each group. To extend balancing to include interactions requires that the condition hold for each multi-group combination, thus, some of the balanced (fitted) number of successes for a multi-group combination are zero, not because everyone failed but because we did not observe anyone.

Thus, balancing with the full interactive model has not been done both because of little evidence of interactions and for practical reasons. Note, however, that the normal equations for the full model give the new condition for balancing on all interaction effects. That is, the fitted (balanced) number of successes in each subgroup must equal the estimated (observed) number of successes in each subgroup in the population. This suggests that the condition for balancing on some manageable subset of the interaction effects can be obtained from corresponding normal equations. The condition for balancing any marginal, two-way, three-way, four-way or five-way effect is that the fitted (balanced) number of successes entering into each effect equal the observed number of successes. The normal equations insure this condition will hold for a unique solution.

Modified Balanced Interactions

As mentioned before, there exist some real practical limitations on the estimation of balanced interaction effects. These limitations are more stringent for the three-way, four-way and five-way interactions, but also apply in the case of the simplest two-way interactions.

In particular, the balancing condition requires that the observed number (weighted) of successes equals the balanced (fitted) number of successes. These numbers are smaller than the numbers in the margins and are subject to more random sampling variability. In addition, the estimates of two-way effects depend on the number (weighted) of cases in each three-way cell. Many of the three-way cell sizes are zero or quite small.

For practical reasons two-way balanced interactions are the simplest to estimate, but for other practical reasons it is difficult to obtain reliable estimates. For these reasons we have not attempted to estimate the two-way balanced interactions precisely. Instead, we have chosen to estimate a modified form of the two-way balanced interactions which is an approximation and can be shown to be algebraically different.

An explanation of how the balanced estimates differ from modified balanced estimates of two-way interaction effects requires an understanding of the notion of "transfers" (see notes below) and the algebraic forms of the normal equations corresponding to a linear model consisting of all main effects and two-way interactions. The notion of modified balanced interactions is briefly summarized in what follows. The algebraic expression of a two-way balanced interaction for group i and group j can be written

$$\widehat{\alpha\beta}_{ij} = \text{Unadjusted } P_{ij} - (\hat{P}_{nat} + \hat{\alpha}_i + \hat{\beta}_j)$$

$$- \Sigma(\text{transfers from group margins to ijth cell})$$
$$- \Sigma(\text{transfer from two-way cells to ijth cell}).$$

This is basically a simplified version of the normal equation. Since the transfers from the two-way cells to the ijth cell depend on the weighted number of cases in the other three-way cells, and these numbers are often small or zero in our sample, these transfers are often unreliable or impossible to estimate.

For our exploratory analysis of two-way interaction effects we have been using an approximation which drops these transfers from two-way cells out of the computation. The approximation is called a modified balanced two-way interaction and can be written

$$\widehat{\alpha\beta}_{ij} = \text{Unadjusted } P_{ij} - (\hat{P}_{nat} + \hat{\alpha}_i + \hat{\beta}_j)$$

$$- \Sigma(\text{transfers from group margins to ijth cell}).$$

Summary

NAEP's best estimate of a group's performance relative to the nation is given by the unadjusted group effect. The percentage of people in a group responding correctly to an exercise is a fact about the age populaton. Balancing does not change that fact. Balancing provides a better comparison of group effects by avoiding the masquerading and "double counting" present in unadjusted effects.

Balancing, though limited, adds meaning to the marginal group effects. Balancing will be more useful as one can balance on other important variables and interactions. The question is not whether to balance (or adjust) or not, but, what are the useful ways? What variables do we need to adjust for? What do the variables mean? and How shall we interpret the results? The balanced group and modified interaction effects are probably better guides to the mechanisms involving the complex set of factors affecting education, while the unadjusted results show more clearly the magnitude of the problems (and measure of success) faced by the people and schools in this country.

Notes on transfers of balanced effects to unadjusted effects.

The 5 factor addivitive linear model we use represents the relationship among group effects in a conceputal balanced population

$$Y_{ijk\ell mr} = P_{nat} + \alpha_i + \beta_j + \gamma_k + \theta_\ell + \phi_m + \varepsilon_{ijk\ell mr}$$

where the Greek letters denote main effects of the 5 factors region, sex, color, STOC and PED. The weighted number of observations for each cell is $n_{ijk\ell m}$ and $Y_{ijk\ell mr} = 1,0$ corresponding to a correct or incorrect response.

Illustration of transfers of balanced effects to unadjusted group effects

Consider the equation for estimating P_{nat} from the usual normal equations.

$$1)\quad n_{....}\hat{P}_{nat} + \sum_i n_{i....}\hat{\alpha}_i + \sum_j n_{j....}\hat{\beta}_j + \sum_k n_{..k..}\hat{\gamma}_k + \sum_\ell n_{...\ell.}\hat{\theta}_\ell + \sum_m n_{....m}\hat{\phi}_m$$

$$= \sum_{ijk\ell m} n_{ijk\ell m}P_{ijk\ell m}$$

Note that the RHS of (1) is equal to the total weighted number of successes for an age group. Divide both sides by n.... and rewrite: (Dot denotes sum over omitted subscript.)

$$2)\quad P_{nat} + \sum_i \frac{n_{i....}}{n_{.....}}\hat{\alpha}_i + \sum_j \frac{n_{.j...}}{n_{.....}}\hat{\beta}_j + \sum_k \frac{n_{..k..}}{n_{.....}}\hat{\gamma}_k + \sum_\ell \frac{n_{...\ell.}}{n_{.....}}\hat{\theta}_\ell \quad \sum_m \frac{n_{....m}}{n_{.....}}\hat{\phi}$$

$$= \frac{\text{wted \# successes}}{\text{wted \# cases}}$$

The usual side conditions, terms 2-6 of the LHS, are all equal to zero. So as expected $P_{nat} = \frac{\text{WTED * SUCCESSES}}{\text{WTED \# CASES}}$. But note for the limited conceptual model, the ratio of group weights to total weight define the proportions in the conceptual model. Now, consider another normal equation for estimating one of the marginal group effects, α_i say

$$3)\quad n_{i....}\hat{P}_{nat} + n_{i....}\hat{\alpha}_i + \sum_j n_{ij...}\hat{\beta}_j + \sum_k n_{i.k..}\hat{\gamma}_k + \sum_\ell n_{i..\ell.}\hat{\theta}_\ell + \sum_m n_{i...m}\hat{\phi}_m$$

$$= \sum_{jk\ell m} n_{ijk\ell m}P_{ijk\ell m}$$

The RHS is equal to the weighted number of successes in group i and $n_{i....}$ is the weighted number of cases in group i. Divide both sides by $n_{i....}$ and subtract P_{nat}, then

$$4)\quad \hat{\alpha}_i + \sum_j \frac{n_{ij...}}{n_{i....}}\hat{\beta} + \sum_k \frac{n_{i.k...}}{n_{i.....}}\hat{\gamma}_k + \sum_\ell \frac{n_{i..\ell.}}{n_{i....}}\hat{\theta}_\ell + \sum_m \frac{n_{i...m}}{n_{i....}}\hat{\phi}_m$$

$$= \frac{\text{wted \# successes group } i}{\text{wted \# cases group } i} - p_{nat}$$

Note that the RHS is the unadjusted group i effect, and $\hat{\alpha}_i$ is the balanced group i effect. In a balanced model the following equalities exist:

$$(5) \quad \frac{n_{ij\cdots}}{n_{i\cdots\cdot}} = \frac{n_{ij\cdots}}{n_{\cdots\cdots}}, \quad \frac{n_{i\cdot k\cdot\cdot}}{n_{i\cdots\cdot}} = \frac{n_{\cdot\cdot k\cdot\cdot}}{n_{\cdots\cdots}}, \quad \frac{n_{i\cdot\cdot \ell\cdot}}{n_{i\cdots\cdot}}, \quad \frac{n_{i\cdots m}}{n_{i\cdots\cdot}} = \frac{n_{\cdots\cdot m}}{n_{\cdots\cdots}}$$

That is, the proportion of group j in subpopulation i is the same as the proportion of group j in the total population. Then terms 2 - 5 on the LHS of (4) are zero and the balanced group effect $\hat{\alpha}_i$ equals the unadjusted group effect, the RHS of (4).

To the extent that the proportions in (5) differ, the balanced effect will differ from the unadjusted group effect.

The 2nd, 3rd, 4th and 5th terms of the LHS of (4) are, respectively, the transfers to the unadjusted group effect (RHS of 4), from the balanced effects of the variable corresponding to the other subscript in each term.

e.g. $\hat{\alpha}_i$ = unadjusted group i effect
- (transfer from j + transfer from k
+ transfer from ℓ + transfer from m)

Note that although balancing marginal effects requires that the observed number (weighted) of successes equals the balanced (fitted) number of successes, the transfers depend on the observed number (weighted) of cases in two-way cells. Since the number (weighted) of cases in the two-way cells are smaller than the number in the margins, they are subject to more random variability in the sampling process. Thus, reliable estimates of the transfers to marginals depends on reliable estimates of two-way cell weights.

APPENDIX 6

WEIGHT TRIMMING

Each student has a weight assigned to his or her responses that is dependent upon the probabilities of selection into the sample. Due to oversampling, nonresponse adjustments and surprises in the schools (larger number of eligibles than expected for example) the weight can vary considerably. Weight trimming as used by NAEP is a procedure for reducing very large weights. It was suggested by John Tukey for the purpose of lessening the effect from any particular school of potentially extreme contributions in the estimatior of p-values.

It is expected that application of the procedure while introducing some bias, will still result in smaller mean-squared errors for the estimates or at least a minimizing of maximum errors.

All students within a particular school taking the same booklet receive the same weight. Since exercises are usually given only in a single booklet, the trimming is performed on each booklet separately. For an arbitrary booklet let:

M = total number of schools receiving the package

N_i = total number of students in school i who responded to the package

W_i = the weight assigned to each of the N students

P_{ik} = the p-value (0 or 1) for an arbitrary exercise for student k of school i.

The national p-value for the arbitrary exercise is

$$P = \left(\sum_{i=1,k=1}^{M} \sum^{N_i} \left[W_i \, P_{ik} \right] \right) / \left(\sum_{i=1}^{M} N_i \, W_i \right) .$$

Assuming

$$\mathrm{Var}\,(P_{ik}) = \sigma^2 \qquad \text{for all i and k}$$

$$\mathrm{Corr.}\,(P_{ik}, P_{je}) = \begin{cases} \rho & i=j,\ k\neq e \\ 0 & i\neq j \end{cases}$$

and treating the weights, W_i, and counts, N_i, as fixed, the variance of P is

$$\mathrm{Var}\,(P) = \left(\sum_{i=1}^{M} N_i \left[1+(N_i-1)\rho \right] W_i^{\,2} \sigma^2 \right) / \left(\sum_i N_i W_i \right)^2 = K \sum_i V_i \, .$$

In the second representation of Var (P),

$$K = \sigma^2/(\sum_i N_i W_i)^2$$

is a constant, and

$$V_i = N_i(1+N_i-\rho)\ W_i^2$$

is proportional to the contribution of school i to the variance of P.

The proportion of the variance of P due to school i is

$$V_i/V_+, \text{ where } V_+ = \sum_{j=1}^{M} V_j .$$

Requiring that school i contribute no more than θ proportion of the total variance is equivalent to requiring

$$V_i \leq \theta\ V_+$$

which is the form of the trimming criterion used. The value of ρ has been set equal to .25.

A rationale of the selection of θ can be obtained by requiring that the contribution to the variance of P by any single school be no more than ten times the average contribution by any school. That is,

$$V_i \leq 10\ \overline{V} = (10/M)\ V_+ .$$

so that θ = 10/M.

Values for θ are shown in the following table:

Booklet Sample Size	Average Group Size	Number of Schools per Booklet	θ
2400	12	200	.050
2500	16	156	.064
2000	16	125	.080
1200	16	75	.133

APPENDIX 7

Examples of economic, education programs and other policy issues that could be related to achievement.

Economic/Fiscal

1. General State Fiscal Structure

   a. General Expenditures
      1) Total General Expenditure 1981
      2) Elementary/Secondary Expenditure 1981
      3) Postsecondary Expenditure 1981

   b. Expenditures--Percent Distribution by Source

      1) Percent Elementary/Secondary of Total General Expenditure 1981
      2) Percent Postsecondary of Total General Expenditure 1981

   c. Expenditures--Percent Change Over Time

      1) Total General Expenditure Percent Change 1976-1981
      2) Elementary/Secondary Expenditure Percent Change 1976-1981
      3) Postsecondary Expenditure Percent Change 1976-1981

2. K-12 Education Finance

a. Public Enrollment

1) Total Enrollment 1982
2) Total Enrollment Percent Change 1978-1982
3) Elementary Enrollment 1982
4) Elementary Enrollment Percent Change 1978-1982
5) Secondary Enrollment 1982
6) Secondary Enrollment Percent Change 1978-1982
7) Average Daily Membership (ADM) 1982
8) ADM Percent Change 1978-1982
9) Average Daily Attendance (ADA) 1982
10) ADA Percent Change 1978-1982

b. Instructional Staff

1) Total Classroom Teachers 1982
2) Total Classroom Teachers Percent Change 1978-1982
3) Elementary Classroom Teachers 1982
4) Elementary Classroom Teachers Percent Change 1978-1982
5) Secondary Classroom Teachers 1982
6) Secondary Classroom Teachers Percent Change 1978-1982
7) Nonsupervisory Instruction Staff 1982
8) Nonsupervisory Instruction Staff Percent Change 1978-1982
9) Principals and Supervisors 1982
10) Principals and Supervisors Percent Change 1978-1982

c. Teacher Salaries

1) Average Salary Elementary Teachers 1982
2) Average Salary Secondary Teachers 1982
3) Average Salary Elementary Teachers Percent Change 1978-1982
4) Average Salary Secondary Teachers Percent Change 1978-1982

d. Sources of Revenue--Total and Percent

1) Total Revenue Receipts 1982
2) Total Revenue Receipts Percent Change 1978-1982
3) Percent Federal Revenue of Total Revenue 1978
4) Percent Federal Revenue of Total Revenue 1982
5) Percent State Revenue of Total Revenue 1978
6) Percent State Revenue of Total Revenue 1982
7) Percent Local Revenue of Total Revenue 1978
8) Precent Local Revenue of Total Revenue 1982

e. Current Expenditures

1) Total Current Expenditure 1982
2) Total Current Expenditure Percent Change 1978-1982
3) Capital Outlay Expenditure 1982
4) Interest on School Debt 1982
5) Total Current Expenditure Per Pupil in ADA 1982
6) Total Current Expenditure Per Pupil in ADA Percent Change 1978-1982

f. State Aid Structure

g. Pupil/Teacher Ratio

1) Elementary Pupil/Teacher Ratio 1978
2) Elementary Pupil/Teacher Ratio 1982
3) Secondary Pupil/Teacher Ratio 1978
4) Secondary Pupil/Teacher Ratio 1982
5) Teacher/Administrator Ratio 1978
6) Teacher/Administrator Ratio 1982

K-12 Education Programs

1. Programs for Special Populations

a. Compensatory Education

1) Compensatory Education Students Served 1980-1981
2) State Funds for Compensatory Education 1980-1981

b. Special Education

1) Special Education Students Served 1980-1981
2) State Funds for Special Education 1980-1981

c. Bilingual Education

1) Bilingual Education Students Served 1980-1981
2) State Funds for Bilingual Education 1980-1981

d. Handicapped Enrollment

1) Handicapped Children (Age 3-21) Served 1981

2. State Mandated Testing Programs/Requirements

a. Elementary/Secondary Assessments
b. High School Graduation Requirements
c. Four-year College Entrance Requirements
d. Two-year College Entrance Requirements
e. Vocational/Technical Entrance Requirements

3. State Programs of School Improvement

    a. Curriculum
    b. Planning/Accreditation
    c. School Improvement/Effective Schools Projects
    d. Dissemination/Adoption
    e. Student Testing/Assessment
    f. Parent/Community Involvement

4. Education Economic Development Programs

5. Business, Industry, Education Cooperative Programs

6. Math, Science, Computer Initiatives

7. Textbook Selection

8. Health Education

Quality of Education Workforce

1. General Description of Workforce

    a. Enrollment Pressures
    b. Teacher Shortages

2. Standards

    a. Teacher Preparation Requirements
    b. Teacher Recertification Requirements
    c. Teacher Testing Requirements/Programs

3. State Programs to Improve Teachers/Administrators

    a. State-encouraged local efforts
    b. Teacher/Administrator Training Academies

4. Fiscal Incentives

    a. Loans/Scholarships, Forgiveness
    b. General Increases in Salary
    c. Merit Pay Plans
    d. Salary Differentials

5. Nonfiscal Incentives

    a. Sabbaticals, Workshops, Travel
    b. Alternative Responsibilities

6. Career Opportunities

    a. Relationship with Private Industry
    b. Restructuring

Legal/Constitutional

1. State Guarantees for Education

   a. Education Clause (General)
   b. Equal Protection Clause
   c. Other

2. Special Education

   a. Special Education Statutes
   b. Civil Rights Statute

3. Equity in Education

   a. Constitution

      1) Prohibition Against Special Legislation
      2) Anti-discrimination Clause: Race
      3) Anti-discrimination Clause: Sex
      4) Anti-discrimination Clause: Religion
      5) Conscientious Objection

   b. Statutes Prohibiting Race Discrimination

      1) General Provision (Cite)
      2) Private Cause of Action (Can individual sue?)
      3) State Enforcement Responsibility
      4) Sanctions
      5) Defacto Discrimination Prohibited
      6) Affirmative Action Requirement

   c. Statutes Prohibiting Sex Discrimination

      1) General Provision (Cite)
      2) Private Cause of Action
      3) Enforcement Responsibility
      4) Sanctions
      5) Defacto Discrimination Prohibited
      6) Affirmative Action Requirement

   d. Statute Prohibiting Discrimination Against Native Americans

      1) General Provision (Cite)

   e. Bilingual Education

      1) General Provision (Cite)
      2) Private Cause of Action
      3) Sanctions
      4) Affirmative Action Requirement

4. Choice in Education

a. State Constitutional Provisions
b. Compulsory Education Laws
c. Regulation of Private Schools
d. If Home Instruction Permitted
e. Aid to Private Schools
f. State Liaison Office for Private Schools

APPENDIX 8

A SIMPLE METHOD OF EQUATING STUDENTS' PERFORMANCE ACROSS BOOKLETS

Percentile ranking of students within an exercise booklet provides a mechanism for estimating a student's performance on other booklets if it can be assumed that the student would retain that same relative position on the other booklets. This provides a mechanism for estimating a total score for a student even though he/she was only assessed with a subset of the total number of exercises.

A conceptually comparable but simpler approach can be employed that provides similar results, since NAEP booklets are given to randomly matched samples from the same population. The national p-values provide a scaling of all the exercises on a scale from 0 to 100. By comparing an individual student's performance on the booklet he took with the scaled p-values, it is possible to estimate how many exercises he might have gotten correct had he taken them all. While this estimate is not very reliable for an individual student, it can be used for analysis of subgroups.

The only data needed to implement the approach are the national p-values, the individual's number correct (n_c), and the total number of exrecises in each booklet (n_p).

Arrange the p-values by booklet in order as in Table 1. Count down within the booklet n_c places and interpolate halfway between n_c and $n_c + 1$. This number represents an ability scaling for the student. Next count the number of exercises across all packages with p-values greater than this number. This is an estimate of the number of exercises the student would have gotten correct had he/she taken all exercises.

When $n_c = 0$ interpolation is between 100 and the highest p-value in the package. Similarly when $n_c = n_p$ interpolation is between 0 and the lowest p-value in the package. In Table 1 assume a student taking booklet A got four exercises correct out of eight. Find the average of 72 and 59 = 65.5 and count up the p-values greater than this amount = 8.

This represents the expected number correct for the complete set of exercises or 8/24 = 33.3 percent correct. Lower numbers represent highest ability since the number represents the item level estimate of difficulty that the student can handle.

If all that is required is a scaling of the student's performance across exercises (for regression analysis, etc.) then the level obtained through the interpolation (65.5 in this example) represents an ability scaling that is inversely related to estimated achievement. For convenience 100 minus this estimate might be used.

Table 1
Exercise P-Values by Booklet.

	Booklets			
	A	B	C	All
	96			96
	92			92
	88			88
			84	84
		80		80
		76		76
	72			72
		68		68
		64		64
			60	60
	59			59
	58			58
		57		57
		56		56
			54	54
			52	52
	50			50
	48			48
		46		46
		44		44
			43	43
			42	42
			41	41
			40	40
Means	70.4	61.4	52.0	61.3

*Prepared by Don Searls.

INDICES TO MEASURE ABERRANCY, BIAS AND GUESSING TENDENCIES IN ITEM RESPONSE PATTERNS2

Student performance on a booklet is typically graded simply by counting the number of items correct and using this measure as a total score. Students with similar scores are assumed to be at similar achievement levels. For a given score, however, the patterns of response can differ dramatically and these patterns can provide insights into student, item and subgroup (male, female, minorities, etc.) characteristics.

For example, if the items conform to a Guttman scale (students getting easy items correct and missing harder ones), which is often the intent of item developers, then response patterns that deviate from expectation contain information about students such as tendencies to guess, lack of concentration, test anxiety, unusual learning patterns (they learn the harder material before they have mastered the easier material), or possible bias in items against particular subgroups. If particular items are biased or lend themselves to guessing or do not discriminate well, they will reveal distinctly different patterns of response across students.

Patterns of response can be represented by a matrix of zeroes and ones representing student responses to items. A row is associated with each student and a column with each item. Ones represent correct responses and zeroes incorrect responses. Arranging rows and, columns so that items (columns) are arranged from left to right in ascending order of difficulty and students (rows) are arranged from top to bottom in descending order of achieved total scores results in a matrix that has been called the Student-Problem (S-P) table by SATO (1975) and Tatsuoka (1978). An expected S-P table would be a matrix with mostly ones in the upper left corner of the matrix and mostly zeroes in the lower right corner.

Recently there has been considerable interest in identifying unusual response patterns and several approaches have been suggested. Harnisch and Linn (1981) documented these measures and added improvements.

The existing indices based directly on right/wrong patterns either correlate response patterns to marginal patterns or measure simple displacements from a Guttman Scale. They do not differentiate between patterns that are characterized by ones appearing where mostly zeroes are expected (a tendency to guess) from patterns where zeroes appear where mostly ones are expected (a possible bias if consistent across a subgroup) or for that matter from a general intermixing of zeroes and ones that could have many explanations.

[2] A paper presented at the Spring meeting of the Colorado-Wyoming Chapter of ASA by Don Searls.

There are a variety of other approaches for exploring test item bias. See Shepard, et al. (1980) for a comparison of six procedures and Ortiz and Searls (1982) for a regression approach. These approaches rely on latent trait models or use marginal or summarized results. They tend to require more assumptions than those that derive directly from the response patterns.

This section describes an enhanced index of overall aberrancy (A), an index that tends to scale on a bias continuum (B), an index that tends to scale on a guessing continuum (G) and a combination that scales on a bias-guessing continuum (BG).

For any row or column of an S-P table let $\mathcal{Z}$ represent the sum of the squared number of zeroes to the left of (or above) each one (1).

$$\mathcal{Z} = \sum_{i=2}^{N} e_i \left[\sum_{k=1}^{i=1} (1-e_k)\right]^2$$

Where e_i is the ith element of the S-P matrix row or column and N is the total number of ones (N_1) and zeroes (N_0) in the row or column. Then G can be written as

$$G = \mathcal{Z}/N_0{}^2 N_1 \ .$$

Let I represent the sum of the squared number of ones to the right of (or below) each zero,

$$I = \sum_{i=1}^{N-1} (1-e_1) \left[\sum_{k=i+1}^{N} e_k\right]^2 .$$

B can be written as

$$B = I/N_0 N_1{}^2 \ .$$

An overall aberrancy measure can be constructed as,

$$A = (2N_1N_0)^{-1} \left[2(N_0I + N_1\mathcal{Z})\right]^{1/2} \qquad 0 \leq A \leq 1$$

and the bias-guessing scale is represented as,

$$BG = \pm\ 2(N_1N_0)^{-1} \left[|N_0I - N_1\mathcal{Z}|\right]^{1/2} \qquad -1 \leq BG \leq 1$$

where the sign of BG = sign of ($N_0I - N_1\mathcal{Z}$), negative signs indicate a guessing direction, positive signs a bias direction.

Examples are presented for patterns of ten items with six correct answers. Patterns are arranged in order of increasing aberrancy and represent what a small portion of an S-P table might look like. This would be an unrealistically short booklet but the general principles are apparent in these examples.

Student Order	Item Response Pattern										A	BG	G	B
1	1	1	1	1	1	1	0	0	0	0	.00	.00	.00	.00
2	1	1	1	1	0	1	1	0	0	0	.16	.17	.02	.03
3	1	1	1	1	0	1	0	1	0	0	.21	-.26	.05	.03
4	1	1	0	1	1	1	1	0	0	0	.28	.53	.04	.11
5	1	1	0	1	1	1	0	1	0	0	.31	.42	.07	.12
6	1	1	1	1	1	0	0	0	0	1	.31	-.75	.17	.03
7	0	1	1	1	1	1	1	0	0	0	.40	.87	.06	.25
8	1	1	0	1	1	0	0	1	1	0	.41	-.58	.21	.17
9	1	1	1	1	0	0	0	0	1	1	.47	-.94	.33	.11
10	0	0	1	1	1	1	1	1	0	0	.61	1.00	.25	.50
11	1	1	1	0	0	0	0	1	1	1	.61	-1.00	.50	.25
12	0	0	0	0	1	1	1	1	1	1	1.00	.00	1.00	1.00

REFERENCES

Access to School Districts, Schools and Nonstudents, Report 12-IP-56. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Exploring National Assessment Data through Secondary Analysis, Report AY-SA-50. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1982.

Fellegi, I. P. "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multi-Stage Samples," Survey Methodology, vol. 4, 1979.

Folsom, R. E. National Assessment Approach to Sampling Error Estimation. Sampling Error Monograph 25U-796-5. Research Triangle Park, N.C.: Research Triangle Institute, N.C., 1977.

Frankel, M. R. Inference from Survey Samples. University of Michigan, Ann Arbor, Mich.: Institute for Social Research, 1971.

Glass, G., McGaw, B. and Smith, M. L. (1981) Meta Analysis in Social Research, Beverly Hills, Calif.: Sage Publications.

Gentleman, W. M., Gilbert, J. P., and Tukey, J. W. "The Smear Analysis." In The National Halothane Study, edited by J. P. Bunker, W. H. Forrest, Jr., F. Mosteller, and L. D. Vandam, pp. 287-313. Contract PH43-63-65, Department of Health, Education and Welfare; Public Health Service; National Institute of General Medical Sciences, 1969.

Harnisch, D., and Linn, R. "Analysis of Item Response Patterns." Journal of Educational Measurement, 1981, 18, 133-146.

Introduction to the National Assessment of Educational Progress Public Use Data Tapes, Report SY-DT-50. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1981.

Issues in the Analysis and Analysis of Change of National Assessment Data, Report 12-IP-57. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Kish, L. and Martin R. F. "Inference from Complex Samples," Journal of the Royal Statistical Society Series B, vol. 36, 1974.

McCarthy, P. J. "Pseudo-Replications: Half Samples," Review of the International Statistical Institute, vol. 37, 1969.

Miller, R. G. Jr. "A Trustworthy Jackknife," Annals of Mathematical Statistics, no. 35, 1964.

Miller, R. G. Jr. "Jackknifing Variances," Annals of Mathematical Statistics, no 39, 1968.

Miller, R. G. Jr. "The Jackknife--A Review," Biometrika, no. 61, 1974.

Mosteller, F. and J. W. Tukey. "Data Analysis Including Statistics," in Handbook of Social Psychology, 2nd ed., edited by E. Aronson and G. Lindzey. Reading, Mass: Addison-Wesley, 1968.

Mosteller, F. and J. W. Tukey. Data Analysis and Regression (Chapter 8). Reading, Mass.: Addison-Wesley, 1977.

Ortiz, E., and Searls, D. "Statistical Identification of Biased Items." Proceedings of the American Statistical Association Meetings, 1982.

Procedural Handbook: 1970-80 Reading and Literature Assessment, Report 11-RL-40. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1981. ED 210 300.

Procedural Handbook: 1977-78 Mathematics Assessment, Report 09-MA-40, 1977-78 Assessment. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980. ISBN 0-89398-143-5.

Ross, K. N. Searching for Uncertainty. Occasional Paper no. 9. Hawthorn, Victoria, Australia: Australian Council for Educational Research, 1976.

Sato, T. "The Contruction and Interpretation of S-P Tables," Tokyo: Meiji Tosho, 1975.

Shah, B. V. et al. "Inferences about Regression Models from Sample Survey Data." Paper presented at the International Association of Survey Statisticians Third Annual Meeting, New Delhi, India, December 5-15, 1977.

Shepard, L., Camilli, G., and Averill, M. "Comparison of Six Procedures for Detecting Test Item Bias Using Both Internal and External Ability Criteria." Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980.

Tatsuoka, M. M. "Recent Psychometric Developments in Japan: Engineers Grapple with Educational Measurement Problems." Paper presented at the ONR Contractors Meeting on Individualized Measurement, Columbia, MO, 1978.

The National Assessment Approach to Objectives and Exercise Development, no. 12-IP-55. Denver, Colo.: National Assessment

1980.

Three Assessments of Science, 1969-77: Technical Summary, Report 08-S-21, 1969-70, 1972-73 and 1976-77 Assessments. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1979. ERIC no. ED 168 901. ISBN 0-89398-297-0.

Tukey, J. W. "Techniques for Analysis of Groups." A personal memo to National Assessment staff and the Analysis Advisory Committee, 1970.

Tukey, J. W. Exploratory Data Analysis, Reading, Mass.: Addison-Wesley, 1977.

Woodruff, R. S. and Causey, B. D. "Computerized Method for Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, vol. 71, 1976.