NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS: AN EDUCATIONAL INDICATOR

Robert C. Larson, W. Todd Rogers, Donald T. Searls,
National Assessment of Educational Progress

In the past, the value of the educational system has most often been equated with the number of people educated by that system. However, recent major emphasis has been placed on the outputs of the educational process. Data is now required that indicates the extent to which skills, knowledge, and attitudes are acquired by young people. Both the Organization of Economic Cooperation and Development [1] and The Office of Management and Budget [2] recently identified the National Assessment of Educational Progress as an educational indicator that provides this desired information. Although a relatively young project, these acknowledgements indicate the potential of National Assessment as an educational indicator of output.

A project of the Education Commission of the States, National Assessment has the purpose of providing dependable information describing what young Americans (9-, 13-, 17-year-olds and young adults 26-35) know and can do. More specifically, the assessment is designed 1) to obtain, at regular periodic intervals, census like data on the knowledge, skills, understandings, and attitudes possessed by various subpopulations in the United States and 2) to measure the growth or decline in educational attainments that takes place over time in learning areas of educational concern. Ten subject areas are assessed: Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies and Writing. The assessment timetable through 1980 is summarized in Exhibit 1. Within each subject area, National Assessment determines the proportions of the subpopulations of interest who can acceptably answer a question or successfully perform a task judged important by scholars and lay people and accepted by educators as something that should be taught in our schools. Successive assessments in a subject area will determine changes in these proportions which reflect the magnitude of change in the educational attainment of the various subpopulations for which data has been collected. Thus data will be available that will be of value not only for allocating public and private funds to the educational enterprise, but also to all levels of the educational decision making community and general public.

1. ASSESSMENT TIMETABLE

1969-70	SCIENCE, WRITING, CITIZENSHIP
1970-71	READING, LITERATURE
1971-72	MUSIC, SOCIAL STUDIES
1972-73	MATHEMATICS, SCIENCE*
1973-74	CAREER AND OCCUPATIONAL DEVELOPMENT, WRITING*
1974-75	ART, CITIZENSHIP*
1975-76	READING*, LITERATURE*
1976-77	MUSIC*, SOCIAL STUDIES*
1977-78	MATHEMATICS*, SCIENCE**
1978-79	CAREER AND OCCUPATIONAL DEVELOPMENT*, WRITING**
1979-80	ART*, CITIZENSHIP**

*SECOND ASSESSMENT
**THIRD ASSESSMENT

The intent of this paper is to provide a brief overview of the structure and methodology[1] of National Assessment and to present some of our output measures collected to date and how these might be used as indicators of educational attainment and change.

NATIONAL ASSESSMENT MODEL--METHODOLOGY

Subject matter areas. As mentioned above, ten different subject areas are included in the assessment. The first stage in our procedure is to develop objectives and subobjectives within each subject area. These objectives, developed and reviewed by panels of scholars, educators, and concerned lay people, must be considered important to subject matter specialists, reflect acceptable teaching goals, and be of value and important for young people in a modern society to learn.

Once an appropriate and acceptable set of objectives has been written, several exercises are developed to represent each objective. These exercises are the instruments with which NAEP gathers data to be used to indicate the extent to which the objectives are being met by each of the various subpopulations of people for which results are desired. Unlike more common educational instruments, the exercises used by National Assessment are not designed to discriminate between individuals or groups of individuals. Like objectives, each exercise is developed and thoroughly reviewed by specialists in exercise construction within the area to be assessed and other scholars, educators and concerned lay people.

From the pool of developed exercises

[1]For a more thorough discussion of National Assessment procedures, see references [3].

representing the area to be assessed, a set of exercises is selected and used in the actual assessment so that the number of minutes of administration time at each age level takes between 150 and 200 minutes. No student enrolled in school is allowed to spend more than 45 minutes in the assessment. Therefore, the selected exercises are divided into about 12 different sets of packages. Typically, nine of the packages include those exercises that can be group-administered to twelve respondents at once. The remaining packages are administered to one respondent at a time. These individually administered packages include performance exercises (such as a science experiment or playing a musical instrument) and exercises which require a "mini-interview" of the respondent. (The number of different packages varies from year to year and from one age to another.) In each package there are both exercises which are unique at each age and exercises which overlap across two or more adjacent age levels.

Sampling and administration. National Assessment policy is to provide data at a national level and avoid making comparisons among states, schools and school districts, teachers, or individuals. This policy is reflected in our sampling. A national probability sample of approximately 2500-2600 individuals per group administered package and 2100-2200 individuals per individually administered package are assessed each year. The samples, selected using a multi-stage design, are stratified by region, size of community, and socio-economic status. A school frame is used for the 9-, 13- and about 87 percent of the 17-year-old samples. Young adults are obtained through a household frame; 17-year-olds not enrolled in school are assessed through the adult household frame and a specially designed frame of dropouts and early graduates. The total sample size required for the three younger ages is approximately 29,000. Since young adults are permitted to take up to four packages each, approximately 4800 adults are assessed each year. In the last two assessments, approximately 95% of the selected schools and 82% of the selected eligible adults have taken part.

The packages are administered by specially trained exercise administrators. Group administered exercises are presented on tape to standardize procedures and to facilitate understanding for young people with poor reading ability. Individually administered packages are presented in an interview mode according to well documented instructions for administration. Thirteen year olds are assessed during October and November, 9s during January and February, and 17s in-school during March and April. The supplementary frame for 17s out-of-school is completed during May-July; the adult assessment runs from October through May.

Analyses. The completed packages are carefully edited for errors. Some of the exercises are in multiple choice format in which respondents record their answers by darkening ovals; others require scoring by specially trained people with subject matter background in the areas assessed. This data, weighted for disproportionate sampling, forms the base for all analyses. For each exercise the estimated proportions of acceptable and unacceptable responses within each population of interest is computed.

Approximately fifty percent of the assessed exercises are retained for future assessments and are therefore not released. These exercises provide the basis for measuring change across time and are selected to provide reasonable coverage across all objectives, difficulty levels and populations of interest.

Since no respondent receives all the packages, the concept of a "total score" is inappropriate. Instead, results are reported by individual exercise, as well as by a summary value (e.g., the median) for small subsets of exercises which measure the same objective, or have some other common attribute.

OUTPUT MEASURES AND POTENTIAL INDICATORS

Output measures. Present reporting categories include age, and within age, geographic region, sex, size and type of community, educational level of parents, and color.[2] The age levels correspond to

2. REPORTING CATEGORIES PRESENTLY USED BY NATIONAL ASSESSMENT

CLASSIFICATION	NUMBER OF SUBGROUPS	SUBGROUPS
Age level	4	9, 13, 17, 26-35 years
Sex	2	Male, Female
Geographic region	4	Northeast, Southeast, Central, West
Level of Parental Education		No high school, Some high school, Graduated high school, Post high school
Size and Type of Community	7	Extreme inner city, Extreme affluent suburb Extreme rural, Inner city fringe, Urban fringe, Medium-size city, Small city
Color	2	Black, White, other

the end of primary, intermediate, secondary, and formal post-secondary education. The categories Region and Sex have traditionally shown large differences. School districts are thought to vary most

[2]See the list of references for released reports by subject area.

by Size-and-type of Community, and Color and Parents' Education are thought to do well in differentiating socio-economic levels and home and family environments.

Each output measure for an exercise is an educational indicator which possesses meaning for the factors reported. One should be careful not to overemphasize the effect of the identified factor used in labeling the reporting category. Each label for each group represents not only the factor indicated by its name, but also a variety of other factors which may influence the performance of that group.

Several examples are presented which represent the various output measures we have been reporting over the last two years. The examples are presented in a mixed fashion across subject areas and objectives in order to give a general view of our output measures as well as illustrate various aspects of our data.

3. OUTPUT MEASURES FOR HAWK EXERCISE (SCIENCE). PERCENT OF 17 AND YOUNG ADULT RESPONDENTS.

OBJECTIVE: II. Possess the abilities and skills needed to engage in the processes of science

SUBOBJECTIVE: F. Ability to check the logical consistency of hypotheses with relevant laws, facts, observations, or experiments

IN A PARTICULAR MEADOW THERE ARE MANY RABBITS THAT EAT THE GRASS. THERE ARE ALSO MANY HAWKS THAT EAT THE RABBITS. LAST YEAR A DISEASE BROKE OUT AMONG THE RABBITS AND A GREAT NUMBER OF THEM DIED. WHICH OF THE FOLLOWING PROBABLY THEN OCCURRED?

17	ADULT	RESPONSES
4%	2%	○ THE GRASS DIED AND THE HAWK POPULATION DECREASED.
1%	1%	○ THE GRASS DIED AND THE HAWK POPULATION INCREASED.
68%	52%	●* THE GRASS GREW TALLER AND THE HAWK POPULATION DECREASED.
4%	4%	○ THE GRASS GREW TALLER AND THE HAWK POPULATION INCREASED.
20%	30%	○ NEITHER THE GRASS NOR THE HAWKS WERE AFFECTED BY THE DEATH OF THE RABBITS.
2%	10%	○ I DON'T KNOW.
1%	1%	○ NO RESPONSE.

4. OUTPUT MEASURES FOR GOVERNMENT POWER EXERCISE (CITIZENSHIP). PERCENT OF SUCCESS FOR 9, 13, 17, AND YOUNG ADULT FOR SEVERAL REPORTING CATEGORIES

OBJECTIVE: IV. Know the main structure and functions of our governments.

SUBOBJECTIVE: B. Recognize the main functions and relations of governmental bodies; recognize that governmental powers are limited by the people through the constitution.

SCORING: Examples of acceptable reasons to Response C: people could stop him; elected officials could stop him; checks and balances system of government; laws stop him; country would be a dictatorship; not the democratic way.

A. DOES THE PRESIDENT HAVE THE RIGHT TO DO ANYTHING AFFECTING THE UNITED STATES THAT HE WANTS TO DO?

○ YES
○ NO
○ I DON'T KNOW

B. (IF YES) WHY?

C. (IF NO) WHY NOT?

	Stated that the President does NOT have the right to do anything affecting the United States that he wants (No to A).				Stated that the President does NOT have the right and gave an acceptable reason (on C).			
	9	13	17	AD	9	13	17	AD
National	49.2	72.3	78.4	89.0	18.4	51.7	70.8	80.0
Region								
Northeast	50.2	72.2	79.6	89.6	22.6	57.0	75.2	80.5
Southeast	47.3	70.3	76.9	85.9	16.7	43.9	66.8	74.1
Central	49.2	72.7	79.4	87.6	15.1	51.1	69.5	79.6
West	49.7	73.7	77.2	92.8	19.7	53.8	71.3	84.9
Sex								
Male	49.9	78.9	82.7	91.8	18.4	57.9	76.1	81.7
Female	48.5	65.9	74.6	86.5	18.4	45.7	66.1	78.5
Parents Educ								
No H.S.	45.7	57.2	75.6	84.9	18.7	31.4	70.1	72.3
Some H.S.	32.2	66.5	68.5	92.7	10.8	43.3	58.9	83.5
Grad H.S.	46.5	66.4	76.1	90.5	16.1	47.8	69.5	82.7
Post H.S.	58.7	79.3	84.5	92.0	25.2	57.9	77.2	86.4

The percentages reported for the two exercises above are measures of the level of performance of an entire age group (nation) or reporting category. By themselves these percentages are not useful indicators unless reference points can be defined upon which meaningful judgments about the performance can be based.

Several meaningful reference points can be constructed from the data itself. The national level of performance of an entire age group, for example, provides individual reference points for each exercise. The difference between subgroup performance and the national level of performance is useful for describing how uniformly learning is spread among the subgroups defined on one exercise. (See Exhibit 5.)

For exercises retained for later assessments, a similar reference point is defined for the second cycle. A comparison of the differences between subgroup performance and the national performance from one cycle to the next indicates how the relative distribution of learning across subgroups of interest is changing over time on a single task. Also, the difference between proportions of success from assessment to assessment determines if more or fewer people in an age group or subgroup of interest are able to respond acceptably.

For exercises administered to more than one age group during an assessment year, similar kinds of comparisons between ages can be made in the form of indicators which answer such questions as: At what age have most of the people gained the understandings measured by the exercise? or Is the distribution of learning across subgroups for a given exercise similar at different ages?

All the output measures considered so far have been for single exercises. Often it is more convenient to summarize the output measures for all exercises of one type or classification (such as the set of exercises reflecting one objective, typed as physical science or biological science, classified by goal in citizenship or by theme in reading and literature). The median of the differences between subgroup performance and

national performance for a single classification of exercises is the basic summary statistic we have reported. In a sense, this median value is "typical" of the performance of the respondents in one subgroup relative to all respondents of the same age for exercises of one type included in the assessment. Thus, generalizations about the "tendencies" of a subgroup with respect to certain types of exercises can be formed.

5. OUTPUT MEASURES FOR POEM EXERCISE (READING). DIFFERENCE IN PERCENT CORRECT BETWEEN RESPONDENTS IN A REPORTING CATEGORY AND ALL RESPONDENTS.

Read the poem and answer the questions which follow it.

THE WAYFARER

The wayfarer,
Perceiving the pathway to truth,
Was struck with astonishment.
It was thickly grown with weeds.
"Ha," he said,
"I see that no one has passed here
In a long time."
Later he saw that each weed
Was a singular knife.
"Well," he mumbled at last,
"Doubtless there are other roads."

A. What do the following lines from the poem mean?

"I see that no one has passed here
In a long time."

⇨ The way of truth is popular.
⇨ People are fascinated by the truth.
⇨ Truth comes and goes like the wind.
⇨ The truth is difficult to recognize.
➡ Few people are searching for the truth.

⇨ I don't know.

THEME: VII. Receding and Drawing Inferences.
A. Drawing Inferences from Information Given

OBJECTIVES: IB3a: Comprehend what is read. Interpret non-literal and figurative language. Recognize the connotations as well as the denotations of the words read.

IIA2. Analyze what is read and be able to trace sequences. Follow the development of an author's idea.

	13	17	Adult
National	61.5	69.2	66.9
Region			
Northeast	1.2	-0.8	4.9
Southeast	-2.7	-4.1	-13.8
Central	0.6	1.8	-1.8
West	0.9	2.2	6.0
Sex			
Male	-4.9	-3.6	-0.5
Female	5.2	3.8	0.5
Parents' Educ			
No H.S.	-15.0	-12.7	-9.3
Some H.S.	-6.0	-10.7	-2.3
Grad H.S.	0.0	1.5	1.0
Post H.S.	8.3	7.5	16.1

Potential Indicators. The output measures and indicators discussed in the previous section are ones which National Assessment has used in their reports and are derived from the data. Different reference points, independent of the data, can be defined, each reflecting concerns and interests of special groups, such as teachers, curriculum people, policy makers or lay groups. For example, the estimation of expected levels of performance establishes a meaningful and relevant base for different interest groups to judge the performance on an exercise. The task of setting expectations has been undertaken by a group of prominent science educators representing the National Science Teachers Association. They are interested in making National Assessment data on Science exercises more meaningful to science educators. Thus, part of their effort has been to make subjective (expert) judgments of an expected level of performance as a basis for evaluating the assessed performance. National Assessment will continue to study procedures for establishing expectation levels of different groups of people. Of particular concern is the "public's" expectation of how our nation should perform and to provide information which makes comparison of performance with public expectation convenient.

Several states have recently conducted state assessments in which National Assessment exercises were used. Thus, they are able to compare directly their performance with the performance of similarly defined reporting categories in the nation--e.g., it is important to find out if those in the rural areas of a state perform as well as all rural areas in the nation.

National Assessment is currently investigating interactions between subgroups as an additional output measure and potential indicator. It is informative to know, for example, if Blacks in the Southeast do as well as those in the Southeast in general, and if Blacks in the Northeast do worse than those in the Northeast in general. These pieces of information give different implications for allocation of resources to the schools in those regions. The first, if true, would imply that we allocate resources to all schools in the Southeast, regardless of color composition; the second implies that we allocate resources to schools in the Northeast which are predominantly Black.

Aggregate indicators are useful to relieve the problem of too much information which obscures important results. However, there are certain sizeable problems inherent in the development of aggregate indicators from National Assessment data which make such indicators hard to visualize. These problems include differential weighting of exercises, determining units of measure, and the location of meaningful reference points and dimensions. As a possible solution, we have been investigating the use of a singular value decomposition of a non-square array of exercises by reporting categories (cells contain output measures). This could lead to various meaningful linear combinations of subgroups or exercises falling along single interpretable orthogonal dimensions. These linear combinations of exercises could be formed at the objective, theme, goal, subject, between subject levels or even between subjects assessed in different years. Each

combination providing different, useful summary measures on "sets" of exercises. Similarly, linear combinations of reporting categories weight exercises along orthogonal "cultural", "psychological", "socio-economic", and "behavioral" dimensions.

National Assessment has a relatively short research history. Much of our time has involved "doing the business of assessing." We have just begun to generate ideas for educational indicators. The task is not small, and the challenge is there. We invite your participation.

REFERENCES

[1] A Goal-Oriented Framework of Educational Statistics, Office of Economic Co-operation and Development, 1971.

[2] Social Indicators: Education, Office of Management and Budget, Working Paper, March, 1972.

[3] Report 02-GIY Reading and Literature--General Information Yearbook (Preliminary report). Available from National Assessment, 1860 Lincoln Street, Denver, Colorado 80203.

[4] Report 1 Science National Results; Report 2 Citizenship National Results; Report 3 Writing National Results; Report 4 Science Results (by region, sex, size of community); Report 5 Writing Results (by region, sex, size of community); Citizenship Results (by region, sex, size of community); Report 8 Writing Mechanics. Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.
Report 7 Science Results (raw and balanced by region, sex, size and type of community, color, parental education); Report 02-R-00 Reading Summary; Report 02-R-20 Reading (Selected Exercises). Available from National Assessment, 1860 Lincoln Street, Denver, Colorado 80203.
Literature, Music, Social Studies and Writing Reports will be released during the next year.