Perspectives

on

Background Questions in

The National Assessment of Educational Progress

By

Paul E. Barton

Report to
The National Assessment Governing Board
September 2002

Contents

- - - - - - - - - - - -

The views expressed are entirely those of the author.

INTRODUCTION

The passage of the No child Left Behind legislation, with its resolve to raise the achievement of all students, with its emphasis on disaggregating achievement data, with its special focus on reading, with its mandating the National Assessment of Educational Progress (NAEP) in all states, and with the new responsibilities given to the National Assessment Governing Board, presents an opportune time to review the approach to the use of background questions in NAEP.

Until the 1984 Assessment, a relatively small number of background questions was used to report the achievement results for population subgroups. Beginning with the 1984 assessment, the use of background questions hugely expanded, and this extended set of questions—modified from time to time—became a regular part of the assessment. As will be recounted in Section 4, this development was the result of a number of factors, including the National Institute of Education's (NIE) focus on research (after it assumed responsibility for the assessment), the foundation-funded evaluation of NAEP by Willard Wirtz and Archie Lapointe, the NIE Request for Proposals in a re-competition to become the NAEP grantee, and the emphasis given to background questions in the Education Testing Service's (ETS) proposal in the competition.

Over the period when the form of this aspect of NAEP was taking shape, different perspectives were brought to bear on the purposes of background questions. At the outset of this period, the early 1980s, the purpose was to explain educational achievement and to answer important policy questions; NAEP was to be an instrument of educational policy research. A view also emerged that background questions should yield data about issues

research studies had already found to be related to student achievement, (which is not at all the same thing as finding out what is related to student achievement through NAEP).

Emanating from the educators and researchers who served on subject matter assessment committees, questions were formulated to reveal the extent to which favored theories were being put into place, such as whether the "writing process" was being followed by teachers, with NAEP reports showing score differences for students where it was being used and not being used.

Over the years, I have heard the following:

- We need to find out if teachers are following currently accepted "best practices."
- We need to include questions that will test the hypothesis that . . .
- We have asked that question several times but it does not differentiate students on the basis of the scores (of course, simple two-way comparisons of the kind used do not necessarily reveal this).
- We have asked that question several times and it consistently shows a relationship to achievement; we do not need to ask it again.
- We need to poll policy makers to find out what questions they want answered.
- We need to find out what is useful information/knowledge to practitioners—principals, for example.
- Well, we need that question because we use it for "conditioning."[1]
- I know that the question gets at something related to achievement, but it involves a sensitive matter.

[1] Conditioning is a statistical process in which answers to background questions are used to improve estimates of achievement scale scores.

My own perspectives have likely evolved as I have had different experiences and involvement with NAEP and the field of indicators generally. These perspectives range from being involved, in a very small way, in the initial discussions about NAEP when I was at the Labor Department, to using NAEP to get my bearings when I assisted Willard Wirtz as Chair of the Blue Ribbon Committee on the SAT Score Decline, to being a participant in the Wirtz/Lapointe evaluation of NAEP, to being liaison between ETS and the Assessment Policy Committee, and to spending a decade using NAEP as a database for reports issued by the ETS Policy Information Center.

Generally, I view the use of background questions as an important addition to NAEP, and one that has broadened its reach—or its potential reach—in understanding educational progress. I also believe that the use of background questions is an under-attended aspect of NAEP, reflecting the grafting of questions onto an already well-developed system of ascertaining what students know and can do—a mission that has required the principal attention and one that is voracious in its need for available resources. NAEP can only benefit by a discussion of the role and purposes of these background questions.

The following sections address roles and purposes.

1. DISAGGREGATION OF ACHIEVEMENT BY POPULATION SUBGROUPS ("REPORTING GROUPS")

The first NAEP assessments used background questions almost exclusively for the disaggregation of test scores, and that purpose has renewed impetus from the recent education legislation. I think there are two areas of NAEP where considerable improvement and refinement could be made. One is in the further disaggregation of

racial and ethnic classifications. The other is in getting a better handle on socioeconomic levels so we know how children from different economic backgrounds are faring.

Racial and Ethnic Subgroups. While the existing classifications are a considerable advance over the early first assessment where the distinction was made only between White and Nonwhite students, it is perhaps time for further evolution. I assume, with the new sampling arrangement where one sample will provide national data and data for all the individual states, the total will be larger than national NAEP has been.

Even in the high achieving Asian American community, considerable concern has been expressed about all Asian Americans being lumped together. In the mid-1990s, the ETS Policy Information Center showed that there was diversity (*Diversity Among Asian Americans*), and showed differences among subgroups within that population. Likewise, diversity exists among the subgroups that make up the Hispanic community. Effort should be made to report results for more racial and ethnic subgroups.

Socioeconomic Status. NAEP has been weak on the ability to differentiate levels and trends in achievement for students from different socioeconomic background. We know that this difference is so pervasive that the standard practice of researchers in establishing whether a particular variable is related to achievement is to control for socioeconomic status (SES). The accepted measure has been a composite of parents' education, income and occupation. In NAEP, parents' education has been used for this purpose, but this does not itself represent SES (and in the lower grades, it is not reported accurately by students). It helps that we know eligibility for the school lunch program, but that only divides the population into two groups. The possibilities are two.

- Create and test out a composite index as a proxy for SES, using such things as school lunch eligibility, reading materials in the home, education of the parent (at least for twelfth graders, and perhaps for eighth graders).[2] The National Education Longitudinal Survey (NELS) gets SES, and may also have the components of a NAEP proxy. The two could then be compared with actual data, as a check on the validity of an index for this use. Or the exercise might identify other factors NAEP could collect that could be added to a composite index. (Also, we might learn more about the socioeconomic level of the community served by the school through use of zip codes and census data.)
- From time to time, adding a parent survey to NAEP, as has long been done in the High Schools and Beyond (HS&B)/NELS longitudinal surveys, has been discussed. This would, of course, add a considerable burden to NAEP administration.

Such an index would enable NAEP to report, for example, on the basis of SES quartiles, the way NELS does. It would also greatly increase the ability of the research community to analyze NAEP data. Now, only parents' education, or other variables separately, can be used as a control. The index might be incorporated into the WEB NAEP tool so users can go beyond two-way comparisons.

[2] Having a computer in the home also might be used. The survey by the Annie E. Casey Foundation (2001) found that 95 percent of families with incomes of $75,000 a year or more have computers, compared to 33 percent of families in households earning $15,000 or less.

2. TRACKING FACTORS RELATED TO ACHIEVEMENT

The history of the expansion of the use of questions about background and characteristics of instruction is one largely of seeking to use NAEP to *explain* levels of achievement (in addition, of course, to the reporting groups discussed above). A brief history of this is provided in Section 4. The proposition here is as follows: that NAEP be used to track trends in factors and conditions that have already been reasonably well established by the research community to positively or negatively affect educational achievement (or to express it alternatively, to track trends found to be strongly associated with achievement).[3]

I say "reasonably well established" because it is often the case that the research community does not (or has not yet) come to closure on the matter. A requirement of complete acceptance by all might narrow to an alarming degree what is thought to be known. And perhaps some statements are so completely commonsensical that they can be generally accepted. For example, commenting on the debate about whether resources matter, Jonathan Kozal observed that students in schools with French teachers will learn more French than students in schools without French teachers.

In this category of purposes of background questions, NAEP is not expected to create a conceptual model of learning and a set of hypotheses about the variables that contribute to learning outcomes. Instead, NAEP uses the results of in-depth research, likely conducted on a much smaller scale than NAEP, to select indicators that are worth

[3] This is no place for an extended discussion of determining causality. Clearly, a statistical agency reporting associations does not claim to establish causality. However, a research scientist advancing a hypothesis based on theory and accumulated research findings, and subjecting that hypothesis to empirical testing, may well lay claim to having explained a phenomenon. The finding may become accepted after being replicated by others. An extended discussion would eventually reach the question of the validity of applying the physical science model to social phenomena.

tracking because they illuminate the things working in the direction of supporting achievement and the things working against it.

NAEP would not itself need to distill such factors and conditions from the whole body of educational and social science research, although it might commission some work where research syntheses are not adequate. Instead, it would rely on efforts made to distill the findings and on the meta analyses that have been performed.

Shining examples of such syntheses are those stimulated by the 1991 report of the congressionally-mandated Special Study Panel on Education Indicators for the National Center for Educational Statistics (NCES). The report recommended the identification of indicators of the health of the nation's educational system. Following up on that recommendation have been two such efforts: *Education and the Economy* and *Monitoring School Quality: An Indicators Report* (December 2000).[4] The nature of the effort and the utility for the use here proposed is seen in the latter's table of contents, reproduced below.

[4] The study was carried out by NCES by Mathematica Policy Research, Inc., and authored by Daniel P. Mayer, John E. Mullins and Mary T. Moore. There was also a panel of external advisors. At NCES, John Ralph was the project officer.

C. Indicator 7: Technology
D. Indicator 8: Class Size
F. Summary

IV. Schools
A. Indicator 9: School Leadership
B. Indicator 10: Goals
C. Indicator 11: Professional Community
D. Indicator 12: Discipline
E. Indicator 13: Academic Environment
F. Summary

V. Conclusion
A. Quality of the Data
B. The Status of School Quality

Other variables were examined and rejected because the research had not established a relationship to educational outcomes. And the reader may recognize a few in this list on which there is not universal agreement. Of course, new research may modify the list, and it will likely be added to. Even in the hard sciences, long established understandings are subject to revision, as in the recent startling findings on hormone replacement therapy for women.

An example, and one that engenders little controversy, is the matter of discipline. The report has this to say:

> "Researchers have found that a positive disciplinary climate is directly linked to high achievement (Barton, Coley and Wenglinsky, 1998; Bryk, Lee and Holland, 1993; Chubb and Moe, 1990). An orderly school atmosphere conducive to learning could be an example of a 'necessary, but not sufficient,' characteristic of quality schools."

In the 1990s, discipline in the schools deteriorated on several measures, some of them coming from NAEP. In looking at NAEP's assessment of reading in Kentucky, where NAEP was showing little progress in reading despite huge efforts in school reform, I found considerable deterioration in discipline. School reform measures were facing an uphill struggle.

The NCES report referred to above is about school factors related to achievement. Some syntheses of research on out-of-school factors also can be drawn on. For example, we know that low birth weight affects cognitive development, that reading to children in the home helps them, and that frequent moving and school changes retards achievement. The latter two of these are reflected in NAEP data.

Where such indicators can be incorporated in NAEP through the use of background questions, we can look for their movement alongside movement in scores of fourth, eighth and twelfth graders in different subject areas, and achieve a broader picture of educational progress. Identifying an indicator is, of course, the first step. The background question or questions still have to be created. It is also possible that answers to several questions would need to be combined into an index. Such an index will likely be more useful for analysis than answers to a single question.

In the economic arena, we have direct measures of the performance of the total economy, such as Gross National Product and Personal Income, and many other indicators that help assess the economy. We would have a very long way to go to be able to do this in education (more below). NAEP can contribute to helping policy makers scan the horizon for trends that are detrimental and trends that are supportive and need to be encouraged.

Trend information is, of course, gathered in a number of large-scale statistical undertakings, and it would be necessary to decide what is best incorporated into NAEP and what is best collected elsewhere. The advantage of using NAEP is that cognitive and non-cognitive measures are collected from the same students, and we see the trends in

juxtaposition with the students, classrooms and schools in the assessment. And of course, NAEP is in the nation's classrooms every two years.

One approach to the selection process for background questions in this category would be to:

A. Have staff assemble the several best syntheses of research on school and non-school factors;

B. Form a committee of respected researchers and ask them to assist in the selection of indicators to be treated and the questions best used in the assessment,

C. Circulate the resulting indicators/questions widely in the research community; and

D. After the indicators/questions are chosen, seek advice from policy people on priorities among them, from their perspective. Since NAEP is designed to help policy makers and practitioners, this advice is critical.

This approach to background questions does not require NAEP to theorize on the causes of achievement outcomes or engage in relational analysis after fielding the assessment. The indicators can be as good as the results of past and present education research and, of course, cannot be better. And NAEP will be tracking educational progress directly with achievement measures, as well as conditions that further progress or impede it—conveying the kind of information to the education and policy community useful for managing the enterprise.

As a result of the outside research referred to in Section 4, it also may well turn out that the NAEP body of data will be used to improve researchers' knowledge about the strength of relationships between indicators and achievement.

3. WHO ARE THE CHILDREN LEFT BEHIND?

The new education act puts the spotlight on leaving no child behind. NAEP will be the most important means of tracking the groups of children left behind. The background information that permits reporting by population subgroups will enable tracking in terms of the characteristics of who the children are. We can, for example, compare the progress of Black male students who live in rural areas in the Southeast to their White and Hispanic counterparts, assuming the sample size supports that degree of disaggregation.

But NAEP can usefully go a step beyond. A disciplined approach such as discussed in Section 2 above opens possibilities to the use of background questions to track the kind of indicators that have been reasonably well established by research to be related to student achievement. And in the priorities set for selecting indicators to be tracked with background questions, weight could be given to the indicators that fit best with the policy objectives of the Administration and the Congress.

I will give a few examples of the kind of statements that might be made as a result.

- Eighth graders showing the least improvement since the last assessment were Hispanic students living in large cities and going to schools where there was a high proportion of teachers inadequately prepared in academic subject matter.

- Students in high schools with a deterioration in discipline have, on the average, shown no improvement in mathematics.
- In state X, the Chief State School Officer has noted a lack of progress for eighth grade students identified as having changed school three times or more, and has asked that transfer students entering schools be given special catch-up instruction for the first month after transfer.

These indicators can also be used in describing the characteristics of children who are "left out" and therefore are at risk of being "left behind." Achieving greater equality in educational achievement is dependent on equality in access to conditions that are supportive of educational achievement, as discussed in Section 2 above.

Illustrative hypothetical statements might be as follows:

- While White fourth grade students continue to have greater access to instruction by teachers with the most experience, the gap between White and Black students has been narrowed in this respect over the last four years.
- In 13 states, the gap in access to mathematics classes with more rigorous content has been reduced significantly between high and low SES students. The gap has remained unchanged in 35 states and has widened in two.
- In the fourth grade, constructive parent involvement with teachers and schools remains highest for suburban families and lowest for families in inner cities. Little change has been seen in the gap since the last assessment.

A more systematic approach to using proven education indicators (tracked with background questions) could lead to a broadening of the reporting of "educational

progress," better informing those who must guide the education enterprise toward higher achievement and greater equality.

4. "EXPLAINING" EDUCATIONAL PROGRESS

The use of NAEP background questions is relatively straightforward in deciding on "reporting groups" to track level and change in achievement, and in going beyond this to portray in more detail the subgroups being "left behind." We can remain on reasonably solid ground if we use NAEP background questions to track trends in factors that in-depth research has found to be associated with achievement, although we are here dependent on the quantity and quality of the research, and we must face the fact that such research often does not come neatly to closure. However, using NAEP for finding explanatory variables and pinpointing the reasons for (1) a change in educational achievement, or (2) why students reach different levels of achievement, is anything but straightforward.

In considering the desirability of structuring NAEP to explain such things, in scientifically acceptable ways, we face some high hurdles.

- The general problem of establishing causality in social science research, particularly when the controlled experiment in educational settings is seldom possible. A controlled experiment was used in the case of class size in the STAR experiment, but even there, changes had to be made in midstream because of an outcry from parents.
- Absent a parent questionnaire, as in HS&B and NELS, it is difficult to construct a traditional measure of socioeconomic status, but we can do better than has been done.

- The cross-sectional nature of the data limits the possibilities of inference as compared with longitudinal surveys such as NELS-88.
- The complexity of the NAEP design with matrix sampling, conditioning and plausible values limits secondary research to researchers who have mastered these complexities and have advanced methodological skills.

All this said, NAEP has some advantages over all other larger scale assessments and surveys. It has broad coverage of subject matter with its matrix sampling, permitting a test equivalent to one of several hours duration. Also, its test questions are driven by a "framework" where effort is made to keep the assessment closely related to consensus on content standards, such as those of the National Council of Teachers of Mathematics. NELS, on the other hand, has tests that all the students in the sample take, and in more limited subject matter areas. This makes it much easier to use, and NELS is very much more made for analysis than is NAEP.

Many discussions have taken place, from the very beginning, about the potential of NAEP, through background questions, to sort out the factors that determine educational achievement. While this is not the place for a complete history, knowledge of a few milestones will help those who must think through what makes sense today.

- The use of NAEP to explain as well as report achievement was thoughtfully debated at the outset at the National Testing Project Conference in 1963.[5]

[5] December 18-29, 1963, Carnegie files. For a nine-page excerpt of the dialogue, see William Greenbaum et al, *Measuring Educational Progress,* McGraw Hill, 1977, pp. 109-115.

Arguments were made for such explanatory uses, and were rebuffed by Ralph Tyler and John Tukey on basically two grounds. One was that it would make the NAEP project so large that it would be impossible to get it approved, funded and off the ground. The other was that, given the state of social science research, this was very difficult to do.[6]

- NAEP started with a relatively few reporting categories for population subgroups, expanding these somewhat after the first assessment. Later, NIE prevailed upon the Education Commission of the States, where NAEP was housed, to prepare public use data tapes to enable secondary researchers to do relational analysis, and launched a small grants program for such analysis. Responsibility for NAEP had been shifted from the National Center for Education Statistics to the National Institutes of Education in 1978; NIE was a research agency and wanted further analysis.
- The early 1980s saw a series of connected happenings that resulted in a considerable expansion of NAEP background questions and a new impetus to promote analysis.
    - NIE was critical of ECS's slow movement toward research. As originally established, this was not considered to be NAEP's charter.
    - The foundation-funded evaluation of NAEP, *Measuring the Quality of Education*,[7] recommended "that assessment data be

[6] At this time, the debate was dominated by the attribution to family background factors versus school factors, influenced by the debate over the Coleman report.
[7] Willard Wirtz and Archie Lapointe, 1982.

developed in forms facilitating their use for research purposes, including particularly the analyses of factors that may relate causally to student achievement" (p. 43).

- A competition for the NAEP grant resulted in an award to the Educational Testing Service in 1983. The conceptual framework for the ETS plan and proposal is summarized in *National Assessment of Educational Progress Reconsidered: A New Design for a New Era.*[8] A section on "Policy Issues NAEP Should Be Able to Address" (pp. 11-14), concluded: "To address them, we must attack issues of statistical inference, sampling efficiency, age and grade sampling, timely data collection, covariance estimation, construct validity, dimensional analyses and scaling, trend analysis, correlation and background and program variables, and 'causal' analysis." A database of huge dimensions was created. However, the NAEP project itself was fully devoted to creating it and reporting the assessment results, with no time or money for performing such complex analyses. In fact, few such analyses were performed anywhere. In the words of a Charles Shultz character, it was "an insurmountable opportunity."

[8] Samuel Messick, Albert Beaton and Frederic Lord, National Assessment of Educational Progress, Educational Testing Service, March 1983.

- In the 1980s, for each assessment, committees comprised of people from the research and policy communities recommended background questions that cut across all subject matter assessments. Also, the committees advising on subject matter assessment made recommendations on appropriate non-cognitive assessments relating to instruction. The Assessment Policy Committee issued a policy statement to guide the use of background questions.[9]
- With some adjustments from time to time in specific questions, the general structures developed in the 1980s remained in place, at least until the late 1990s.

The policy of NCES generally has been to not perform research seeking "causal" relationships, although it has issued reports involving multivariate analyses, and it has given encouragement to "secondary" research.[10] Two separate but related matters arise regarding NAEP and research into causes of achievement and non-achievement; one regarding research inside NAEP and one regarding research outside it. The first matter is whether this database, created with both cognitive and background questions, should be used to pursue explanations, to label various factors as affecting achievement, and to assign weights among them.

[9] Under the terms of the law, once appointed, all authority for NAEP policy lodged in the Assessment Policy Committee.

[10] I think this should actually be considered "primary," given that researchers tackle a huge and complex database that has not yet been "analyzed" by the collectors of the data.

My view is that it is a matter best left up to the researchers outside government or to those who have responsibilities in government for education research outside of NAEP.[11] A body of data exists here that quantifies many aspects of the context in which achievement takes place, and it measures achievement very well, so it is a database that others can use for relational analyses. A research person or organization using NAEP takes the full responsibility for the proper use of data and the quality of the research. If looking at the national level, few researchers or research groups are in a position to design and carry out a national assessment. It is necessary for researchers to examine the large scale data collections that exist and to select on the basis of what best fits the objectives of the research, recognizing (a) the strengths and weaknesses of each, (b) the problem of ascribing causation and (c) the utility of identifying "associations" where causation is not claimed.

The findings of such research will be judged in the critical reviews that take place. And the research results will find, or not find, their niche in the mosaic created by the world of thesis and antithesis. No single project based on NAEP data will likely emerge as the definitive answer to an important question. It takes many efforts and data sources to establish a consensus on a finding, and NAEP data may be one useful resource to researchers. I note that a book has just been published on the still competing views of the role of class size in achievement, after decades of both small and large scale studies, as well as one sizeable effort with an experimental design.

The second matter is whether NAEP should explicitly structure this large set of background questions to further explanatory analysis. My view is that NAEP cannot

[11] The National Science Foundation, for example.

design the assessment around its own explanatory model of the causes of achievement, making its own set of hypotheses. It is hard to imagine one such model, and an all-encompassing set of hypotheses, particularly one that represents perspectives of the different disciplines.[12]

But NAEP is a government-created measurement and statistical system. As such, it is responsive to the views and needs of the users of the data. Independent researchers and research organizations are users of the data, and like other users, can make their views known about the context information they would like to see collected. And NAEP can explore various means of listening and deciding. The possible hypotheses emerging and the data needed to test them are considerable and will not all be anticipated. And the number of questions that can be asked are finite. But we can get clear on the role NAEP is to play.

I was personally exposed to how wide such a range of hypotheses can be about 25 years ago when I assisted Willard Wirtz in his role as Chair of the Blue Ribbon panel on the SAT Score Decline. We wallowed in the hypotheses offered by scientists and educators, and many of the hypotheses were carefully explored. They included birth order, a watering down of school texts, television watching, scale drift, and fluoride in the

[12] I think it would be good to take another look at the practice of attaching achievement scores to the various answers to background questions, with statements such as: "Scores are higher for students whose teachers have subject matter certifications compared to those without such certifications." Intuitively, that sounds logical, but we need to know whether affluent schools have a higher proportion of certified teachers where the student scores are generally higher than in schools with poorer students. Another example is counter-intuitive. NAEP data show that students who work from 6 to 15 hours per week have higher achievement scores than students who work from 0 to 5 hours per week. A closer look reveals that (1) the higher the parents' education (as a proxy for SES), the more hours students work, and (2) student labor force participation rates are relatively low for Black students. Lower income students in central cities have fewer job opportunities, and they also have lower achievement.

NAEP might just report the results of the background questions. At a minimum, if it ties scores and background questions in its regular reports, it should be at least a three-way comparison; for example, scores, hours worked, and the constructed proxy for socioeconomic status.

drinking water. The hypotheses could not have been all anticipated in advance in one grand research design.

NAEP can focus on the more straightforward approaches described in Sections 2 and 3 above. If a set of background questions is derived from the identification of reporting categories to describe in useful ways the "children left behind" and to track trends in context indicators that research has identified as associated with educational achievement, a body of data will emerge that researchers can mine, particularly if the data have been enriched by additional questions suggested by the research community.

In the use of NAEP for such research, I believe it well to keep in mind the conclusions reached in TIMSS.[13]

> Ideally, one would like to link the attributes of teachers and teachers' instructional practices to the demonstrated achievement of the students they teach and, in this way, identify effective teachers and effective teaching practice. Such a linking is possible within the TIMSS data, but cross-sectional designs of the kind that characterize TIMSS (and IEA studies in general) are not well suited to this purpose. Students enter eighth with knowledge, beliefs, and orientations accumulated over 7 years of schooling and some 13 to 14 years of family life. What teachers do within the space of a school year is unlikely to radically alter the achievement level of the class as a whole and so create a sizable correlation between teacher instructional practices and student achievement at the classroom level. The best hope to demonstrate the relationship between teachers' instructional practices and student achievement is to look at the relationship to growth in achievement over the year, rather than absolute levels of achievement. Recognizing this, the original design of TIMSS was one that required a pre-and posttest to measure this growth. Unfortunately, most of the participating nations were unable to support both a pre- and posttest, so the study reverted to a simple cross-sectional single testing design.
>
> As a result, the present analyses and those which took at influence besides instruction (curriculum, for example) can offer no more than circumstantial evidence of the context for learning mathematics and science and, hence, of what might move U.S. students toward the realization of the goal of being first in the world."

[13] *Mathematics and Science in the Eighth Grade: Findings From the Third International Mathematics and Science Assessment*, National Center for Education Statistics, July 2000, p. 91.

5. PURELY INFORMATION

Another category exists beyond proven indicators that track who is "left behind" and explain educational achievement—the categories described above. The other category is the use of background questions purely to collect needed information. The two criteria are:

- A judgment that the information is needed to better inform the education enterprise; and
- A judgment that NAEP is the best place to collect the information, relative to other large scale surveys.

I am not here talking about relating this information to scores on the NAEP assessment, or to proven relationship to achievement.

NAEP has some real advantages for collecting some kinds of information. NAEP is regular, allowing trends to be tracked for every couple of years, and it reaches into the school, classroom, home, and to the individual student.

The most obvious examples are in periods where there is a widespread reform movement; we need information on how it is progressing and how it is being viewed at the point where it is applied. It might have been the science and math reforms after Sputnik, the minimal competency movement of the 1970s, or the reform agenda of the 1980s, such as raising requirements for core courses, demanding more homework, and lengthening the school day. And it might be the current standards-based reform movement. An example of the latter is described on page 30 regarding questions revealing conceptions of principals and teachers as to progress of such reform in the schools and classrooms. Another example might be in asking teachers what they learn

from the standardized tests that helps them improve instruction, and what more they need to know from tests.

6. SOME CUSTOMIZING OF STATE ASSESSMENTS?

One significant change brought about by the recent legislation is the mandatory participation of the states. The fielding of state assessments requires cooperation to be successful, whether the assessment is mandatory or not. And given that the states have to participate, and that they are being pressed hard in this legislation to raise achievement, it would be desirable if the results in states were actually helpful in charting the course of improvement. Making the results of incorporating background questions more useful is, of course, the purpose of the whole exercise being undertaken by NAGB. Beyond the approaches discussed above that apply to the assessment system as a whole is the possibility of providing for some customization of the background questions in the state assessments, with each state able to add a set of questions, replace some of the standard ones, or choose among blocks of questions.

At any one time, each state has its own set of concerns and policy priorities. An individual state has its own ideas about its educational deficiencies and the measures needed to help grow achievement. Or it has its information needs for shaping its approaches. If a state has a "a piece" of the background questionnaire, it may see more utility in the state assessment, and it may make a larger investment in analysis.

One thing we know is that beyond test scores, states learn almost nothing from their own standardized testing systems; in NAEP, they can learn about the context in which instruction takes place, and the progress they are making in getting changes installed in schools, classrooms, and instruction.

It has been shown that NAEP questionnaires, with customizing, can be used as an information system to help lead an education reform movement. I have been very impressed with what the Southern Regional Education Board (SREB) is doing in this regard in the High Schools That Work (HSTW) consortium. SREB, from the beginning of the program in 1986, has used a NAEP-based assessment, using assessments constructed with blocks of released NAEP test items. It has also used the NAEP background questionnaires, but has modified them to meet the specific needs for giving leadership to the program.

An example of an SREB modification is in adding information on course contents. The objective has been to get participating schools to get rid of the watered down general courses and switch to courses that are on a par with courses in the college preparatory track. Gene Bottoms' work shows how other schools that have done this have raised their scores on the assessment.

Of course, there is the question of how to handle this in assessment administration and scoring, as well as cost considerations. But this may be a useful step in getting greater use of NAEP results within a state. States need to see more value from this expensive operation than only an external check in the overall accountability system. My impression is that states have used the information generated by the background questions very little. It might be useful to have visitations to states to help them with what the data mean, as is done in the HSTW program.

7. CONDITIONING AND VALIDITY

Conditioning. Discussions of background questions are usually held on the particular merits of particular questions from the standpoint of their utility to contribute

to understanding some aspect of the educational and learning process. Recognition is also needed that a large number of these questions are used operationally in the process of "conditioning." The background questions play a role in the estimation of achievement scale scores.

When the large expansion of background questions was undertaken in designing and fielding the 1984 assessment, this was not contemplated. It became necessary when the application of Item Response Theory was found to need supplementation through the use of Rubin's approach to missing data, using relationships between background factors and achievement to strengthen the estimates of scale scores.

This use of background questions should be well satisfied in the approach described in Section 2 above. The more background factors are used that have proven relationships to achievement, the more useful the background questions are for the conditioning process.

Validity. I think issues of validity of the background questions asked in NAEP have not been given the attention needed in this important government statistical system. NAEP background questions, beyond those used to report scores of population subgroups, were grafted on to NAEP about 15 years after NAEP was created; cognitive measures continued to occupy center stage because they were the central mission of NAEP. Where resources were available for it, some very good people have worked on framing the background questions; however, these resources have always been very limited.

An ongoing statistical series needs a budget for the specific purpose of creating and maintaining the quality of the instruments used. For example, ongoing research

programs are in place in the Bureau of Labor Statistics and the Bureau of the Census for the Monthly Report on the Labor Force. Different ways of asking about labor force experiences are tried out, in order to get the questions worded in a way that yields the information being sought. The series measures unemployment, but people are not asked if they were unemployed last week. Instead, they are asked what they did last week, and follow-up questions put them in the categories of being in the labor force or not, employed or not, or unemployed or not. When labor researchers wondered whether "unemployed workers" included people who had abandoned a job search because they had concluded that no jobs were available, the researchers made attempts to identify "discouraged workers" to distinguish them from people classified as "not in the labor force." Different measures of who was a discouraged worker were run at the same time in the research effort to find the best way to frame the questions. Such effort is a likely characteristic of many data collection systems.

I believe it would be instructive to look into several such important surveys to determine the scope of the research experiment and validity effort that takes place in them, and compare that to NAEP. The use of small focus groups for trying out questions may be helpful.

Also, other survey research efforts could be examined to see what they have learned in areas that are similar to NAEP. For example, NELS has both student questionnaires and parent questionnaires; this could perhaps be used to compare parent and student answers to some questions. What do parents say about how much they help

with homework, or whether they monitor TV watching, as compared to what students report?[14]

8. QUESTIONNAIRE DEVELOPMENT AND ANALYSIS

If school, teacher and student questionnaires administered to a large national sample were a freestanding survey research operation, a considerable budget would be allocated for their development and maintenance. A considerable budget would also be allocated for data manipulation, analysis and reporting.

There is, of course, a very sizeable budget for the development and analysis of the cognitive side, as there needs to be. But as pointed out above, these background questionnaires were grafted onto an ongoing and very sophisticated system of cognitive assessment. The background question side has never been adequately funded as the large scale research effort that it has every appearance of being. There was wisdom in adding this dimension to NAEP, for understanding educational progress, and the lack of it, is more than the measure of a final result in terms of test scores, even when the test is a superb one.

In the arena of the economy, we have measured the Gross National Product, a final result. But we have also developed many measures that tell us how the economic system is functioning, so we can understand better how we reach that final result. We measure productivity, price change, inventories, wage change, sector employment, sales, and many other economic activities. But we are groping our way to having such indicators in the education system. The indicators report previously referred to

[14] It may be profitable to look at the minority language study appended to NAEP in the 1986 assessment, where a parent survey was used. I believe it found a very large discrepancy between parent and fourth grade student reports of parent education level, and to some extent, at the eighth grade level.

(*Monitoring the Quality of Education*) is a milestone in the work of government toward becoming systematic about synthesizing research for such purposes. The useful indicators will not come from one source. NAEP, being in classrooms every couple of years, is an important vehicle to contributing to a broader understanding. The opportunity is considerable.

But realizing the opportunity requires more than a two-day meeting of a committee to recommend questions. It requires more than one staff person (and it has not always been a fulltime one) who works on this under the NAEP contact, as was true until the late 1990s. It requires staying on top of educational research. It often requires developing a set of questions that can be collapsed into an index. It requires the expertise of the research community that specializes in framing questions and testing different wordings of them to yield the desired knowledge. It requires a paper trail on the process, just as there is on the cognitive and psychometric side. (One NAEP technical report I looked at has only a few paragraphs on the background questions, but over a hundred pages where those questions were used for conditioning in making the scale score estimates.)

A bit of history is relevant here. When the three large questionnaires were added to NAEP for the 1984 assessment, the budget was $3,880,000, the same as the prior assessment, and the same as for the subsequent assessment. I do not know all the history of the financing of NAEP since that time, but I am not aware that there has ever been a separate research, development, analysis and reporting budget for this huge non-cognitive survey research enterprise. However, I am aware that beginning around 1998, the

American Institutes of Research (AIR) has been under contract to assist in the development of background questions.

The expectations we develop and the plans we make should be funded at a realistic level that will permit their realization.

9. THE END GAME: CONVERTING DATA TO KNOWLEDGE

A strategy for informing the nation and state (and perhaps the district) about the condition of learning begins with inclusion of the background questions in the assessment. Getting clear on the purposes for having the background questions informs the selection of questions. A total design and operation of the assessment, however, contemplates the end game as well—planning for the use of the data and the transformation of raw data into knowledge.

This is not to say that the NAEP enterprise itself assumes a responsibility for all the analysis. The points made above refer to the use of NAEP data in research and analysis operations in academe, think tanks, and government and educational entities. At least, that is the hope—and has been since these background questions were greatly expanded 20 years ago. We also know from that experience that such use so far tends to be quite limited, even as the achievement data get broader use. I would hope that clearer purpose would be of some help in achieving broader use. But it will be necessary to call upon a new word added to the language these last 20 years; it will take becoming "proactive."

A number of possibilities are offered below.

Facilitating Research with Composites. One such composite—and a big one—is the proxy for socioeconomic status discussed above. If four or five separate questions are

asked to get a NAEP counterpart to SES, there is a need to combine these in such a way as to assign an index number to each student, so that the number can be used in analyses in the research community. There is no exaggerating the importance that the creation of the Duncan composite of income, education, and occupation has had in facilitating social science research.

Knowing the very large differentials in achievement among the quartiles of students enables researchers to disentangle associations with a "treatment" variable from the associations with socioeconomic background. And being able to quantify SES enables us to track whether we are reducing the large gap in achievement that somehow has become part of the structure of American society and opportunity in it.

Other composites also could make the information more useable for analysis. If we have many questions about the subject matter content of instruction, we may be able to create some idea of how "rigorous" instruction is. If we probe with many questions about the continuum from strictly rote memorization to advanced problem-solving approaches, we may be able to get some composite of where instruction is along that continuum. We are not here looking for a "data bite" parallel to the over-simplification represented by the TV sound bite, but ways to move from a daunting set of separate questions to the essence of what such a set of questions is getting at.

A Focused Report for Each Assessment? While the NAEP project itself will not be the locus for systematic analysis and report writing, it could well be expected to lead the way with at least one special report/analysis for each assessment, basing it on the background questions.

This would require a decision as to what that special emphasis/report would be, so as to have continuity from the formulation of the questions to the data analysis to the report. A couple of examples are provided only to illustrate the possibilities.

- A probing of depth in mathematics instruction (as compared to breadth), on a state by state basis to see if there is a pattern in terms of student achievement. The effort would be a NAEP parallel to the extensive work done in TIMSS on an international basis, where U.S. instruction has been characterized as a mile wide and an inch deep.
- Conceptions of teachers and principals as to how much state content standards have resulted in:
    - Modifications in content of instruction;
    - Provision of teaching materials (texts, workbooks) that reflect the new content standards; and
    - Alignment between the tests used and the instructional content.

I understand that a move has been made in this direction through a contact with AIR to design a focused set of questions for each assessment and to follow through with the preparation of a report.

Service Center for Users. While it is likely possible for individual researcher and data users to get some questions answered now, the service function could be expressly organized and provided for as NAEP becomes more important to the reform and accountability process. States will likely be asking for more help in understanding and "getting behind" NAEP results, at least in terms of achievement reporting. There may be opportunity for a closer relationship with the state analysis/research staff and an

opportunity to create more interest in analysis at the state level. And as NAEP takes on more importance in accountability, there will be more interest in NAEP in the research community. Where NAEP and state test results are not moving in tandem, there will be a desire to investigate why, both in state offices and in research/policy organizations.

Grants Competitions for Increasing Analysis and Attracting Young Scholars. From time to time, NCES has sponsored grant competitions to pose projects for the analysis of state data. I am reminded of the admonition to fill the birdfeeder regularly or the birds will abandon it and go elsewhere. If such grants were a regular and expected funding opportunity, there would likely be more specialization in using NAEP for analysis. An investment is involved for even an experienced researcher to achieve competence in the use of the NAEP database. Researchers will not likely make that investment unless there is a perceived return. And how about a NAEP primer for researchers?

There is also the use of very small grants to support Ph. D theses. The experience of the Labor Department, beginning in the mid-1960s, is instructive here. In the early 1960s, human capital economics sprang forth and was encouraged by the Manpower Development and Training Act of 1962, which provided for a substantial research program. To attract young economists to the field, Labor established a small grants program to help fund Ph. D thesis. This resulted in attracting a goodly number of good economists. In my own opinion, we also got some good research, with more creativity and insight than sometimes emanated from think tanks and large-scale projects. This could be a way of creating a cadre of researchers schooled in the use of NAEP for analysis.

Promoting State Level Partnerships. I am here referring to partnerships between state education officers and a policy research center in the state. These centers are usually located in universities, although they may be stand-alone nonprofits. The purpose would be to mine state NAEP for better understanding of achievement and its sources within the state. Perhaps small grants could be made to encourage the partnerships, covering at least the expenses of a liaison between the two, who would also serve to identify joint projects. NAEP could recognize and encourage these partnerships through regional meetings, for example.

Information Bulletins on NAEP Use. The labor market longitudinal survey (NLS) originated by the Department of Labor in the 1960s, and still going, has long had a newsletter informing about research projects using NLS data, the kind of data tapes available, and the availability of reports. Its primary audience is research users.

* * * * *

One last word. It is continually the case that some groups have problems with specific questions or specific types of questions asked of students. It is always necessary to balance the need to well inform the education system and the reform movement with the concerns of interested organizations and citizens. I think that clarity in the purpose for the questions will help in decisions about dropping or adding questions, and in striking a desirable balance among all interests. And I hope that there is a strong enough conviction about the importance of the purposes to defend the use of questions necessary to realize those purposes.