DISCUSSION

Robert F. Boruch, Northwestern University[1]

1. Vigorous Competition and its Risk

The concerns, during 1984, about contract competition for the National Assessment of Educational Progress were serious for anyone interested in high quality measurement of student achievement. In particular, the Education Commission of the States had been conducting NAEP for over 12 years, at over $3 million each year, and had done a decent, if sometimes pedestrian, job. Moreover, ECS developed good relations with state departments of education and school districts, relations essential to producing good data. Further, ECS had considerable experience in dealing with the technical, professional, political, and management problems engendered by NAEP. It is positioned well to give NAEP visibility, credibility, and utility.

The consequences of choosing a poor contractor and consequences of bad decisions by well-chosen contractors would be serious: Disruption of the single vehicle for making statements about the state of education in the U.S. based on a good national probability sample. The problem is distinctive to NAEP but not unique, of course. As one of our astronauts told interviewers at the start of a moon shot: Every part in this vehicle was got from the lowest bidder.

2. Initial Performance and the New NAEP

The choice of a new vehicle was, to judge from its performance in its first major test, very good. It is awesome that ETS and Westat have, in the first year, managed to obtain nearly 90% cooperation rates from school districts, well over 80% from children within schools, and to initiate new parallel surveys of samples of teachers and principals (Hansen et al., 1984).

It is no less awesome that they have been able to introduce new technical ideas and new research policy ideas to the process. The product is lovely to those who like to see creative adoption of fundamentally good ideas in novel settings and who aspire themselves to be among the best of science-engineers in applied social research (Messick, 1984; Beaton, 1984).

It is remarkable that this has been accomplished without the political visibility and influence of the Education Commission of the States, without (in the case of Westat) the professional visibility and long-term relations with the school community, and without the benefit of a long relation between the contractors themselves.

The quality of people reflected in all this is, to put it mildly, unusual. By "people" here I mean not only ETS and Westat management but staff. I mean not only the abiding operations staff, but the creative scientists and technicians. I mean not only the contractors, but also the administration and staff at the National Institute of Education: It did take bureaucratic craftmanship and courage to get other ponies into the race at all, and to assure fairness at the starting gate, during the race and at the finish.[2] Finally, the accomplishment also reflects professionalism among ECS-NAEP staff members: without their assistance in the first year, none of this would have come about.

3. Specific Merits

Part of the merit lies in what can be called assessment policy, policy that was determined not by only a political committee but by a scientific/technical group that is sensitive to policy interests. The elements of this new policy that can enhance utility of NAEP include choice of time of testing (spring), regular biennial testing of four subjects, accommodating the cohort effect problem, and grade level (and age level, temporarily) testing rather than age level testing (Messick, 1984). There is not much innovative technical contribution here. But these decisions on research policy are hard and they are very likely to make the data much more useful. For instance, the potential user now knows when and what will be produced, and how it relates to an organizational dimension such as grade level that is important for interpreting results, and for educing the implications of results.

The merit lies also in technical creativity - the production of ideas that enhance utility by decreasing the cost of the enterprise or enhancing the data's usefulness. In this respect, Youden, of course deserves resurrection for a variety of reasons. Here, his protean efforts are reflected in a fine enlargement in application of balanced incomplete block designs (Beaton, 1984). It is perhaps no surprise to see how useful BIB's have been in weighing designs for standards of weights and measures, for assuring privacy in social surveys on sensitive topics, and now, for testing achievement in a massive program. It is a pleasing coincidence that Youden's design is being used to assess secondary students some of whom are likely to have read Youden's (1969) nice monograph, written for such students, on weighing designs in the context of measurement.

The benefit of this kind of application lies, more importantly, in the production of data that permit us to better understand the structure of the data, notably the relations between performance on achievement tests and the actual ability of students. At its best, the application of BIB technology will help us to understand mathematically and substantively the relation between student performance and characteristics of schools and students. And, in the interest of future work, the data may be a far better vehicle for allocating scarce resources and estimating the effect of allocation than we have had. For example, the scale invariance engendered by models that may be fitted to such data will facilitate exploitation of regression-discontinuity designs in allocation and estimation (Trochim, 1984). The ceiling/floor problems of other ostensibly simpler ways of scaling have impeded progress in this arena.

There remains, as Beaton suggests, some technical problems that will require creative solution, nonpositive definite matrices being one. The solutions, we expect, will advance our understanding of how to exploit approaches to missing data. They may also encourage search for still other approaches to obtaining relatively complex data. For example, it may be sensible to compare a fractional factorial approach, adopted and

extended nicely by Rossi and Nock (1982) to complex social surveys, to the benefits and costs of a BIB approach.

4. New Issues

This performance will suggest to many observers that Westat seems to sit at the right hand of god with respect to sampling, and ETS serves as His (or Her) left hand with respect to measuring ability. Despite the seating arrangements, some important chores remain. Chief among these, I think, are the problem of assuring the data's usefulness and assuring the evaluability of the operation apart from technical excellence.

Piggyback Policy

Technical creativity does not guarantee that NAEP data will be used in the ways it can and should be used. Technical creativity aids, but cannot substitute for creativity in expanding, to the extent possible the nature, frequency, and quality of the data's application. And it remains to be seen how new NAEP will fair in this respect.

Generalized "Piggybacking Policy" seems especially promising here given some states' interest in better indicators of educational excellence and to substitute for poor approaches to indexing such as the input-out charts recently released by the Secretary of Education. States can augment samples to better exploit NAEP at the state level (Sebring & Boruch, 1983). Local districts can augment to permit better assessment at this level. Moreover, it appears to be feasible at low cost relative to state assessments. The latter cannot be replaced by the current NAEP, for state assessments can and do recognize curriculum content better than NAEP. Bridges between the two are bound to enhance understanding of each and, in this limited sense, at least, to enhance the new NAEP's utility.

How to enlarge and implement broad Piggyback Policy, how to build those bridges is not yet clear. These jobs do, however, have strong implications for making an excellent technical product a remarkably useful one at national, state, and perhaps local levels of government.

Evaluation

The second area that warrants exploration in how to better evaluate performance of NAEP. Technical standards are relevant to be sure and these have to be used. Still other kinds of standards can also be regarded as relevant. We ought to understand them better in the interest of fair and catholic assessment of assessment. NAEP needs to be marketed: How do we evaluate the marketing plan and it's execution, taking into account the public's often sturdy indifference to data? NAEP needs to be supported financially by groups other than the National Institute of Education. How do we evaluate the support development effort? By the nature and frequency of interesting and creative joint ventures? NAEP should be visible, to be useful in more than academic arenas. But how do we evaluate efforts to secure good competent press coverage? NAEP is, like a linear accelerator, an instrument of science technology. How do we evaluate its productivity in this respect, especially in getting beyond citation counts (Should we get beyond counts?)

The point here of course is that the state-of-the-art in evaluating such an assessment is underdeveloped, or at least fragmented. Developing a more coherent approach seems sensible, partly because of the increase in state assessments, partly because new NAEP will in five years again be subjected to serious scrutiny in recompetition. The theory and data that will have to be used to evaluate at that point ought to receive attention now.

Footnotes

1. This invited discussion was presented at the American Statistical Association meetings, Philadelphia, August 13-17, 1984, in response to papers on NAEP by Hansen, Beaton, Messick and their colleagues. Support was provided by the Center for Statistics and Probability at Northwestern University (Reference: A-244) Larry Rudner of the National Institute of Education was Chairman of the Symposium on NAEP.

2. An earlier attempt to elicit proposals for conducting NAEP, was imperfect in a remarkable respect. Only one organization submitted a bid: ECS. NIE and its academic advisors learned how to do the RFP issuance better. This is no mean thing.

References

Beaton, A. E. Statistical issues in data analysis for the National Assessment of Educational Progress. Presented at the Annual Meeting of the American Statistical Association, Philadelphia, August 13-17, 1984.

Hansen, M.H., Tepping, B.J., Lago, J.A. & Burke, J. NAEP-the sample and data collection scheme for year 15. Presented at the Annual Meeting of the American Statistical Association, Philadelphia, August 13-17, 1984.

Messick, J. The new design for the National Assessment of Educational Progress. Presented at the Annual Meeting of the American Statistical Association, Philadelphia, August 13-17, 1984.

Rossi, P. and Nock, S. L. (Eds.) Measuring social judgments: A fractional factorial approach to survey design. Beverly Hills, CA: Sage, 1982.

Sebring, P.A. and Boruch, R. F. How is National Assessment of Educational Progress used? Results of an exploratory study. Educational Measurement: Issues and Practice, 1983, Spring, 16-20.

Youden, W. J. Experimentation and measurement. New York: Scholastic Books, 1962.