4 Trends in Statistical and Analytic Methodology: Implications for National Surveys

THIS PAGE INTENTIONALLY LEFT BLANK

"So What?" The Implications of New Analytic Methods for Designing NCES Surveys

Robert F. Boruch
George Terhanian

SUMMARY

This report was commissioned to address the question "How can advances in statistical analysis be used to improve the design of surveys?" The surveys of paramount interest are those sponsored by the National Center for Education Statistics (NCES). The "advances," as initially conceived, include new approaches to analysis that have been invented by statisticians, mathematicians, and methodologists.

Advance: Mathematical statisticians and methodologists, at times, remarkably improve the way we analyze statistical data. But they rarely describe how their advances can improve the *design* of surveys. Scholars who apply the new (or old) methods to NCES data, at times, speculate on how NCES surveys might be improved and report their suggestions in journal articles.

Implications: First, NCES can encourage scholars who invent new analytic approaches to educe the implications of their advances for improving survey design. NCES should not expect to find explicit implications absent such encouragement. Second, NCES can encourage scholars who apply new (or old) analytic approaches to NCES data to educe the implications of their results for better survey design and to contribute more effectively to a common pool of implications. Third, NCES can exploit mechanisms that NCES and other federal agencies already depend on to build this knowledge pool, e.g., external committees and internal staff. Fourth, NCES may exploit new technology to do so, notably on the World Wide Web (see section on New Technology).

Cross-Design Synthesis

Advance: Recent work on cross-design synthesis suggests that, at times, survey-based studies of the effect of national programs, based on probability samples or administrative records, can be combined with local controlled experiments on the programs' effects so as to produce better national estimates of the impact of the programs. In the long run, combining such information is arguably important to advancing knowledge and to the efficient exploitation of resources in both the survey sector and the experimentation/evaluation sector.

Implications: First, to foster good cross-design synthesis, NCES surveys can be designed so as to permit linkage of the surveys to controlled experiments. Experiments at the local level, over which

NCES has no direct control, can be designed so as to permit linkage with NCES surveys. That is, both the surveys and the independently conducted experiments can be designed cooperatively so that response variables, treatment variables, target populations, and propensity variables are measured in the same way. Second, NCES can ask, or learn how to better ask, about propensity so as to enhance analyses and synthesis. NCES can do so in ways that others have not, through cognitive research and other approaches.

Hierarchical Models, Models in General, Theory, and the Design of NCES Surveys

Advance: Hierarchical models and associated models and analysis help to frame the way we look at data that are generated at the national level, state level within nation, school district level within state, classroom within school or district, and children within schools, and the way we examine data on each child or classroom, and so on across a time frame. A notable advance lies in contemporary software.

Implications: The claims made by the developers of hierarchical models are sufficiently broad as to allow vague statements about how new NCES or any other surveys should be designed. A first such implication is that NCES should collect multi-level data, as it has in the past. A second equally vague implication is that the NCES effort ought to be expanded, invigorated, and made more disciplined in the context of HM technology, e.g., figuring how whether and how to enlarge sample size at certain levels. Although proponents of HM may merely identify general implications, at least some who employ the approach are more specific. A third set of implications is that NAEP 1) should measure socioeconomic status more directly or less indirectly than it now does; 2) should get at teacher instruction variables better; 3) should elicit information from more teachers if indeed we want to know about their influence; and 4) may have to sample more students within schools. A fourth design implication for NCES is that investments have to be made in understanding how to estimate sample size within each level in a hierarchical scenario. A fifth implication is that NCES has to decide where HM-driven implications ought to be exploited, e.g., in designing NAEP versus NELS:88. Other simpler models and analytic approaches may be better and, in any case, theory ought to drive some of this. NCES has to take theory into account somehow.

Advance: Meta-analysis, which can be construed in terms of hierarchical models, involves the combination of multiple studies.

Implications: NCES surveys can be designed so as to exploit the results of meta-analyses to design a survey. This requires, in the design or modification of each survey, the invention of a mechanism for linking the survey at hand to other related surveys or experiments (Sections 2 and 6).

Counting the Hard to Count, Measuring the Hard to Measure

Advance: New developments in analyzing count data suggest that a social network-based estimator of the incidence or rate of a sensitive behavior can, at times, be informative. Such estimators avoid certain privacy problems in educational and social surveys, and avoid the appearance of problems. That is, they are based on questions asked about unidentified people, not on questions about the respondent's own potentially embarrassing behavior.

Implications: When privacy is an issue but understanding the incidence of a sensitive behavior is important, NCES can consider the design surveys that exploit network-based estimators. A second implication is that some basic research, pilot work, and verification research are, as usual, necessary.

Advance: Cognitive approaches appear not to have been employed often in test development despite their use in other survey areas. Full information matrix factor analysis is alleged to be a relatively new way to get at the structure underlying test results. Neither analytic approach itself has obvious implications for NCES survey design. A Stanford group used these, together with other methods, and applied them to mathematics and science data and other information from NELS:88. They produced implications for design of NELS:88 and perhaps other surveys.

Implications: Mathematics reasoning and knowledge are two distinct latent factors underlying test scores generated in NELS:88. They ought to be treated as such inasmuch as total scores are arguably misleading. Science scores are characterized by many different factors. Moreover, each type of factor is influenced in theory and predicted empirically by different variables whose measurement in NELS:88 can be improved. Some variables that may relate differently to each factor are not measured at all, e.g., instructional practices such as discovery learning or reciprocal teaching. A main implication is that NCES can exploit theory of how knowledge and reasoning are affected by various factors. The theory and analyses can be used to drive NCES decisions about what to measure, how deeply to measure, and why.

Small Area Estimators, and So On

Advance: Recent work on indirect estimators suggests that it is possible, at times, to develop good small area estimators based on 1) data from a national probability sample, 2) information obtained independent of the national sample, and 3) a model that links the two. "Good" estimators here means that they are more plausible than any alternatives.

Implications: NCES surveys can be designed so as to exploit new work in domain indirect, time indirect, or time and domain indirect estimators. Time indirect estimators might be tested to understand whether they suffice to permit reducing NCES annual data collection efforts to biennial efforts, or to lengthening the time between points of measurement in NAEP and other periodic surveys. Domain indirect or time indirect estimators might *now* be tested to determine if satisfactory local area estimators can be produced or if certain area surveys now producing direct estimators can be reduced. Validation tests are possible because NCES now relies heavily on direct estimates.

Satellite Policy

Advance: NCES survey data are at times used to sustain analyses of cause and effect. The problems in doing so are complex, numerous, and have been discussed often and in numbing detail.

Implications: NCES surveys can, at times, be designed to facilitate local controlled experiments, for example, by oversampling the subgroups that are targeted for experimental programs. This requires survey designs that permit linkage between the surveys and experiments.

Linking NCES Surveys and Data Sets from Other Sources

Advance: Multiple independent surveys are undertaken often, and with good reason, by NCES and various other federal agencies. To judge from recent analytic work, the independence of surveys mounted by different agencies or units within the Department means, however, that the results of different surveys often cannot be easily integrated, compared, combined, or otherwise linked. More important, NCES has had substantial recent experience in the problem of integrating certain data collection efforts, e.g., the CCD.

Implications: NCES can take a leadership role in learning how to run independent surveys or studies more generally so that linkage, comparison, integration, or merger is possible despite their independence. The task hinges on enhancing the extent to which major factors are common to different databases, e.g., variables, ways of measuring the variable, target population. It hinges on the invention of ways to specify the lack of commonness, and on the invention of ways to induce artificial commonness.

New Technology

Advance: The development of the Internet, especially the World Wide Web, does not fall into the category of advances that concern us here. Nevertheless, it is too important to ignore.

Implications: There are a variety of tactics that might be exploited in interest of better design of NCES surveys. They include Web-based surveys of data users and analysts to 1) elicit direct information on questions, design characteristics, and so on; 2) build a registry of users, uses, and products; 3) distribute spreadsheet files; 4) track the emergence and development of new analytic methods; 5) create electronic discussion groups among analysts and designers; 6) post frequently proposed questions and their answers; and 7) exploit Adobe functions to better disseminate information.

INTRODUCTION

This report focuses on what new analytic methods imply for the design of better surveys. The surveys of special interest here are those conducted by the National Center for Education Statistics (NCES) (Davis and Sonnenberg 1993; Davis and Sonnenberg 1995).

The report's topic was determined jointly by the author, NCES, and an NCES contractor, MPR Associates. It was chosen to assure that NCES could exploit new opportunities to enhance survey design on education in the United States if indeed such opportunities are engendered by new analysis methods (NCES 1995). The various parts of this report vary in their length, developmental stage, and depth. Some are better thought out than others; some implications are stronger than others.

The first section examines the broad question: "What are the implications of new analytic methods for the design of NCES surveys?" It describes why the answers to the question are hard to produce. It also describes why and how implications can be produced.

The next section concerns recent work on cross-design synthesis. It argues that data generated in surveys of the sort undertaken by NCES can be combined at times with controlled experiments sponsored by other federal agencies or by private foundations. This combination of data is in the interest of better estimating the effects of federally sponsored education programs and policies.

The third section focuses on recent work on hierarchical models and other statistical models. Some implications are obvious, provided there is some agreement that measuring individual growth trajectories, or estimating the effects of schools is important.

Counting the hard to count and measuring the hard to measure is considered in the fourth section. We focus on network-based estimators and on recent analyses of NELS:88. NCES cannot always elicit information directly about the private behaviors of students, teachers, parents, and so on. This is despite the fact that these behaviors, such as criminal or sexual or disruptive activity, may be important on policy grounds. One new method, invented by a quantitative anthropologist and a physicist, is reviewed here and the implications are laid out. A second section covers the product of an interesting effort by Stanford scholars to learn how to improve NELS:88.

The fifth section of the report concerns indirect estimators, including small area estimators. The object is to understand how NCES, whose efforts are routinely based on large scale periodic national samples, can estimate the incidence of problems in small geographic areas or can abstain from one cycle of a national data collection effort. Achieving either object is not trivial, given NCES's mission to produce data based on national probability surveys and the pressure to say something at the subnational (small area) level, and given the pressure to produce information on a regular cycle, and given restricted resources.

Section six is entitled satellite policy. It argues that the NCES surveys and others ought to be an unobtrusive platform for controlled experiments run by other technical agencies or private foundations.

Section seven concerns the idea of linking surveys and data sets. Linkage, combination, comparison, and related ideas are considered briefly. This essay exploits research that was sponsored by the National Science Foundation and is relevant to NCES interests.

The last section of this report considers new technologies and how they might be exploited to enhance the design of NCES surveys. The focus is on the Internet and how the "Net" can be exploited in the interest of designing better NCES surveys.

EDUCING THE IMPLICATIONS OF NEW ANALYTIC METHODS FOR THE DESIGN OF SURVEYS: SOME PECULIAR DIFFICULTIES

The question at hand is "What are the implications of new approaches to statistical analysis for the design of surveys?" Put another way: "How can surveys be improved, based on advances in analytic methods?" A basic reason for posing the question is that it seems important. Or at least interesting.

The presumption is that an agency, such as NCES, can exploit advances made by the inventors of new ways to analyze data. A further presumption is that exploitation can enhance the design of the National Assessment of Education Progress and other surveys. It seems then sensible for the agency to do so. In the abstract at least, one might speculate that advances in analytic methods might for example, lead to designs that enhance the precision, informativeness, or usefulness of surveys or decrease their costs or difficulty. The phrase "in the abstract" is of course important here.

A second reason for asking the question has to do with an early partial flop. A decade ago, a Social Science Research Council Committee on Evaluating Longitudinal Surveys addressed the question. Some good products were developed (Pearson and Boruch 1986; Boruch and Pearson 1988). However, the SSRC conversations on how new statistical models and methods could be exploited to improve longitudinal surveys led nowhere.

One simple way to uncover answers to the question is to examine the writings of statisticians who invent new analysis methods. The presumptions are that these experts are in a good position to understand the implications of their work and, further, will have written about it. In the following section, we pursue this line of thinking and examine what appeared initially to be a promising approach and examine the published literature, proceedings, journals, and books.

Proceedings of the American Statistical Association

To understand what new analytic methods imply for survey design, it seems sensible to peruse the *Proceedings of the American Statistical Association: Survey Methods Section*. The 1993 edition was examined for papers describing new methods. These, in turn, were examined for a section on "Implications" or "Conclusions" that might educate us about answers to the question. We found none. (We did find implications in papers *other* than those dedicated to the mathematical invention.)

One might surmise that ordinary sessions of the ASA are usually not oriented toward the future. Rather, it may be more sensible to examine a source that is less time constrained, such as the *Proceedings of the Sesquicentennial Meeting of the American Statistical Association* (Gail and Johnson 1989). Boruch read each of the *Sesquicentennial* papers and looked for a sentence, paragraph, or section on implications and for conclusions that might have implications for the design of new surveys.

With a few exceptions, no paper in these special *Proceedings* directed attention to the matter. One of the exceptional papers described interviews with two able statisticians, Ron Gallant and John Pratt. The interviewer elicited their expert opinions about the implications of statistical theory for the design of a better census. Roughly speaking, both answered "I don't know."

Sending an e-mail inquiry to colleagues who are inventive about analytic methods seemed a sensible thing to do. So, a few of them were asked if they had *written* about the implications of their work for designing better surveys. Each individual had made remarkable contributions to analysis. Only one response is given here because it is instructive. It is from a colleague whom I admire on account of his inventiveness and industry.

> Thanks for your note. There's no doubt that better methods of analysis can lead to better designs. That idea permeates so much of what I do, I don't know exactly what to send. So I've decided to send you a CV and you can pick by interesting title. Also, I'll think harder to find particular appropriate articles.

We have depended on this scholar's work elsewhere in this report. His response reiterates the notion that implications of invention are important. But for able inventors, they cannot be drawn plainly, or will not be drawn plainly for many reasons, including the fact that "the idea permeates."

New Approaches to Analyzing Cohort Data: A Volume

Mason and Fienberg's edited volume (1985) handled advances in analyzing cohort data. The approaches to analyses are relevant to NCES surveys inasmuch as NCES sponsors surveys that attend to different cohorts of students in different time periods. Understanding the differences among cohorts and determining what may account for similarities or differences seems important. None of the papers in the Mason-Fienberg volume are explicit about how new analytic methods can be employed to improve any surveys, much less NCES efforts.

Failing to identify an explicit discussion of implications in Mason and Fienberg should not deter us, of course. Some implications may not be labeled as such. David Freedman's essay (1985) in the volume begins with the announcement that "[r]egression models have not been so useful in the social sciences" (p. 343). These models, for Freedman, include logics, time-series, and LISREL. His definition of social science includes education and psychology. His paper preceded recent developments in hierarchical linear models (HLM), but it seems reasonable to include HLM in his ambit.

Freedman argued that conventional statistical approaches to data analysis, as they are conventionally applied, have not had much yield. More important here, Freedman suggested that *any* new advances in statistical methods of analysis are likely to be uninteresting without major changes in the way that we think about data and about educational research and the behavioral and social sciences.

That is, the question posed earlier in this report, "Do new models and analytic methods have implications for better survey research design?" would have little merit for Freedman. It is the scientific thinking that underlies the models and methods that is important for him. Indeed, he argues that many of the models and methods are not sustained by good thinking about the processes that generate the observations in the first instance, i.e., a social theory.

It may not be difficult to agree with Freedman. Agreement, however, implies that the topic of this paper is misguided. Let us keep this implication in mind and resurrect it later.

The *Journal of Educational and Behavioral Statistics*: A Special Issue on Hierarchical Models

A recent issue of the *Journal of Educational and Behavioral Statistics* focused on hierarchical models (Kreft 1995). The issue's contents were reviewed to understand whether its

authors suggested how surveys could be improved, based on advances in the subtechnology of hierarchical models. Only one author of an article in the journal stated that there are implications for design survey. His statements were opaque.

The Society of Industrial and Applied Mathematics

Curiosity and opportunity led us to ask about the topic of this report of a founding member of the Society for Industrial and Applied Mathematics (SIAM). SIAM's members, one might expect, would at times educe the implications of new analytic approaches in mathematics, including statistics for the better design of empirical research.

The interview with this scholar suggested that our mathematical colleagues are not inclined to speculate about how their work can be used to enhance future research. That is, mathematicians do not often educe the implications of their innovations for further work, at least not in print. The disinclination may, of course, be influenced by proprietary interests. Some members of SIAM are employed by profit-making corporations. University-based mathematicians who are also members of SIAM presumably have a taste for applied work. They invent new solutions to problems. But they also appear to infrequently educe the implications of their work and to make the implications plain in their published work.

The *Journal of Educational Statistics*: A Special Issue on Models

A special issue of the *Journal of Educational Statistics* (Shaffer 1992) reviewed the "Role of Models in Nonexperimental Social Science." David Freedman and Howard Wainer wrote their papers on structural models and on analyzing survey data, respectively. The commentaries and the authors' responses to criticism are important additions.

The authors did draw implications that bear at least indirectly on the design of some surveys, including perhaps NCES surveys. Freedman argued that "investigators need to think about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models" (p. 27).

In effect, this again suggests that we may have gotten off on the wrong foot in this report by focusing on the implications of new analytic methods. That is, for those of us who are interested in science, the theory ought to drive the way a model is built. The model, in turn, drives analysis: parameters that ought to be estimated, hypotheses that should be tested, and so on. This in turn can perhaps improve design of surveys, e.g., identifying assumptions whose tenability might be informed by certain designs and this leads to new models and analyses.

Wainer's conclusion was to "think hard" about nonresponse. In effect, this means inventing small theory whose elements might be informed by new data; he suggested that the new data are essential in understanding the nonresponse. Critics of the Freedman and Wainer papers argued along similar lines. Hope, for example, concluded "[t]here is no methodology that will write our theories for us" (p. 46).

To put the implications of these analysts bluntly: better theory (thinking) is warranted. This may not seem much like guidance for improving NCES surveys. But it does introduce some interesting choices for NCES that are discussed elsewhere in the report.

The Meaning of the Question at Hand

What was meant by "implications" at the outset of this essay was not made clear. Finding even a few implications reminds us to be more specific about what we seek. The word here means that, as a consequence of a new analysis approach, we might better understand any of the following (Exhibit 1):

1) What variables to measure or not to measure;

2) How to measure;

3) Whom to measure;

4) How many to sample;

5) When and with what frequency of measurement;

6) With what periodicity;

7) With what sample design characteristics (strata and so on);

8) In connection with what other data collection;

9) Why; and

10) How to report.

This list accords with at least some efforts to understand how to improve surveys generally. Items concerning what variable to measure, when, and on whom, are embodied, for example, in the products of a recent NAS-IOM workshop on integrating federal statistics on children (Board on Children and Families and Committee on National Statistics 1995). The list also accords with how users of new analytic approaches and data sets suggest improving survey design on the occasions that they do so, for example (Boe and Gilford 1992).

The list seems promising enough to use as a template for further work. Internet-based facilities that are discussed in the light of this report are suggested as a device for orderly acquisition of information on such items. Such a facility, a list server, for example, then provides a continuously updated archive of possible improvements based on the experience of users of NCES and other survey data.

The phrase "new analytic methods" as used in the title question may seem clear to some, but it is deceptive. Implicit in the phrase is the presumption that buried in any new method is a new model. A further presumption is that it is better to have explicit models to drive an analysis of data than to have analysis driven by implicit models. Both approaches are functional, however, to judge from the history of science including statistics. The former is regarded here as more functional.

Further, a new model might or might not have to depend on substantive scientific (educational) theory.

What can be regarded as "new," of course, is not obvious. Hierarchical models, though new to many users, are based on mathematical efforts that extend at least to Kempthorne and Cochran, and Cox in the 1940s and 1950s. So called network-based estimators are based in a fundamental way on elementary ideas about the probability of independent events. Spiraling methods now used in NAEP have their origins in balanced incomplete block designs developed over 30 years ago, and so on. The point is that here, when we denominate a method, model, or approach as new, the denomination is merely a convenient label.

And, of course, what a model or method *is* can be similarly complicated. Here, the focus is on a model that contains a stochastic error term and is suppose to represent reality—reality itself being partly represented by survey data. The models and methods examined here include hierarchical models, indirect estimators, design synthesis, and projection models, among others.

Published Analyses of Specific Data Sets

A search of education journals for 1991–95 uncovered 31 reports of analyses of data from NELS:88. Most of the authors employed conventional analytic methods such as OLS linear regression; perhaps three employed newer methods. Disregarding the analysis method, 15 out of 25 papers that we were able to review contained some form of implication. Nine articles contained no explicit statement of implications for better designing NELS:88.

Two papers were direct in providing very broad implications and indeed were developed to do so. These concerned the construction of math and science achievement tests so as to better recognize the multidimensional character of such ability. Of the 13 remaining papers, most called for new variables to be measured. Authors said that NCES should measure "global self-esteem" (instead of academic self-esteem), ask about criteria for placement of students into ability groups (instead of just asking whether students are grouped), ask how long students have lived in a single parent family (rather than just whether they do), elicit information on parental education and indicators of middle school philosophy (rather than just the existence of middle school). This list is idiosyncratic. That is, the implication drawn by the data analyst depends heavily on the analyst's particular theoretical framework and objective. This varies dramatically across analyses.

Only a couple of papers suggested that samples of certain groups be "beefed up," e.g., Hispanic students. And of course, some papers reiterated the need to collect similar data in the next wave of measurement, a tactic that NCES examines routinely.

This evidence suggests to us that some orderly way of identifying implications is warranted. The Terhanian Home Page model discussed later in this report is one option, a way of summarizing articles, implications, and analytic methods. It also implies that some method for routinely screening the published analyses is warranted; existing NCES advisory groups, for NELS:88 for instance, are an option.

Conferences, Working Groups, and Other Integrative Instruments

NCES and other federal agencies rely, from time to time, on specially convened groups to say something sensible about its activity. The group may be appointed by a department, as in the case of the NCES Advisory Council on Education Statistics, or the group may be appointed independently, as in the case of a National Research Council Committee. These and other groups might be expected to develop the implications of contemporary research for the future of the agency, including perhaps the design of specific studies. Some groups do so.

For example, researchers at the Educational Testing Service have occasionally tried to learn whether and how disparate databases that concern science could be used in combination. The Hilton (1992) effort, sponsored by the National Science Foundation, was unsuccessful in a few respects; it was successful in others. It employed rather than invented new methods or models. Surprisingly, Hilton's work (1992) dedicated little attention to how their lack of success could be rectified. That is, not much was said about how the design of independent surveys could be improved to foster their combination (see section on Linking NCES Surveys and Data From Other Sources).

Two other groups, which neither directly employ nor invent new statistical analyses, were also examined. Both dealt with the problem of "linking" data sets, the first being on teacher supply and demand (Boe and Gilford 1992) and the second concerning statistics on children (Board on Children and Families/Committee on National Statistics 1995). Both contain what amount to implications of prior empirical analyses and thinking, based on new methods and otherwise.

Teacher Supply, Demand, and Quality

Boe and Gilford's volume (1992) covers the NRC conference on this topic. Supported by NCES, the group was convened in the interest of enhancing the teaching force in the United States by focusing on major issues in the area and the information needed to understand them. This effort entailed reviews of the data that are produced, the data that might be produced, and the models that are used in forecasting supply or demand. The reviews perforce cover earlier analyses of the data, analyses that employ new methods or old.

This paper deals with "implications." In a sense, the NRC Conference on TSDQ also did so. It was "designed to reach a consensus . . . to stimulate suggestions concerning 1) information . . . and 2) further development of projection models and databases" (p. 3). The conference summary then provides NCES with another choice about how to characterize "implications." It and the main report are also interesting because they categorize the ideas/implications into two broad and arguably instructive categories: "information needs" and "suggestions." The needs usually refer to what variables ought to be measured. The suggestions focus on more specific implications. (Note that *none* of these are "recommendations"; the conference was not empowered to make them. This is a virtue in many respects.)

Exhibit 2 outlines the TSDQ Conference's summary of information needs. It is a short list of what variables ought to be measured by NCES and other communities of scholars even if we do not yet know how to measure them. We are told that we need, for instance, to measure teacher quality

(Information Need 1). The rationale is to better inform decisions about quality-quantity trade-offs and model-based forecasts of whether and how we might improve.

Implicit in some items is theory. Information need #3, for example, suggests that the demographic mix of teachers ought to be examined with respect to the demographic mix of students. Are old white people teaching young Hispanic people? What kinds of people are teaching whom? And, does it matter? Each question has an implicit, and rudimentary, theoretical basis. It seems important to recognize this basis and NCES can do so.

Some of the TSDQ Conference suggestions are outlined in Exhibit 3. Several points are worth noting. First, all the suggestions can be categorized using the generic list of implications in Exhibit 1, which reinforces the notion that this list may be a reasonable way to summarize such things. For example, Suggestions 1 and 2 bear on research to inform the use and measurement of the variable called teacher quality (Items 1 and 2 in Exhibit 1). Suggestions 18 and 19 bear on connections to other data sets, e.g., linking SASS to state databases bears on Item 8 in the list.

A second point worth noting is that the TSDQ Conference suggestions are a matter of collective judgement based partly on the expertise of participants and the papers commissioned for the conference. Backtracking to the volume's papers, we find most are based on rather simple but informative analyses. Murnane (1992), for instance, argued that state licensing records on teachers is a valuable resource and ought then to be linked somehow to the NCES effort, based on analyses showing downward trends in licensing and in licenses given to black college graduates and in their probability of returning to teaching having left the profession some time earlier, all from North Carolina records. Murnane also argued tersely for redesign of state record systems on account of the great difficulty he and his colleagues had in exploiting them. He did not recognize the NCES expertise in this area. But the crude implication we draw from this is that NCES' expertise on design of data systems and linkage is a major resource that might well be exploited in any effort to better capitalize on state data.

Only one paper in the TSDQ Conference *Proceedings* focused on models, and using them in the context of NCES surveys and state data. Barro's concerns (1992) lay solely with projection models of different kinds and the data used to sustain their use. His paper is nonetheless instructive because of the implications that were drawn from it by Boe and Gilford (1992), such as Suggestions 18 and 19, and on account of Barro's own thinking. Indeed, Barro's entire paper can be regarded as an exercise in drawing implications. For instance, he argued that the mechanical (demographic) demand models in contemporary use are far less informative than new behavioral models that help one address "what if" questions. His implication is that NCES ought to use the "what if" theme to drive design; NCES' current projection models are of this variety (Gerald and Hussar 1992). Improvements, according to Barro, lie partly in adding variables such as pupil-population ratio and teacher salaries. It lies partly in treating a measured variable, notably state aid to schools, *not* as exogenous but as a variable that itself ought to be forecast from other (unspecified variables). Other suggestions lie in frequency of measurement (Item 5 in the generic list); more being better, in using state-level data to build more detailed and policy-relevant models (Item 8).

The idea he produced for better designs based on the supply side are sustained by simple rather than elaborate models and findings from their application. His implications are numerous.

Among other things, he reiterates the need to exploit SASS to get better forecasts of teacher attrition rate, especially 1-year followups of subsamples to get at turnover.

Many of the implications that Barro drew seem important. They are certainly ample. About one implication appeared every page and a half in the discussion on demand. One implication that we draw from the way Barro approached his task and the TSDQ Conference *Proceedings* is, again, that the generic list in Exhibit 1 is helpful in classifying implications. A second, more important, concern is the mechanisms available to NCES to uncover implications on new *or* old models and their application. A conference was organized to do so. Third, when implications are ample, we need to keep track of them and their bases. The generic list in Exhibit 1 helps the orderly acquisition. Sharing such information beyond print would arguably help (see the section on New Technology).

Integrating

The Board on Children and the Families and the Committee on National Statistics (1995) of the NRC/IOM convened a workshop "to examine the adequacy of federal statistics on children and families" (p. 1). Its joint sponsorship, by the Board and the Committee, and the topic itself led us to expect "implications" to be produced and indeed they were. The final report, *Integrating Federal Statistics on Children* (hereafter called *Integrating*), is plentiful in its supply of them.

The summary of *Integrating* outlines cross-cutting "suggestions" (p. 2) based on collective expertise and commissioned papers, as in the Boe and Gilford (1992) effort. But the summary is rather broader in its handling of them. We are told the following, for example:

> Improvements in data are needed to understand the connections between resources and child outcomes, as well as family and community processes that translate resources into outcomes (p. 3).

This is rationalized by recognizing the availability of data on input variables (e.g., PSID) and offering the opinion that "data on child outcomes are substantially more limited" (p. 3). This "implication" does not recognize, much less exploit, the notion that children's education achievement is an outcome, that NCES routinely obtains such information *and* information that bears on some resources. Two sub-implications were drawn: that data ought to be collected for "more than purely descriptive purposes . . . " and that the use of time by parents and children is a major variable that is rarely measured. The first item is relevant to NCES in that the agency is often confronted by the need to incorporate substantive theory into debates about design of surveys. The second is relevant inasmuch as NCES has asked questions about how time is spent in some surveys, e.g., time on teaching certain topics and time in watching TV. Again, this is unrecognized in *Integrating.*

Integrating's summary is about what variables to measure, as in the example above; about family relationships (e.g., biological, adoptive, step, and noncustodial parents); about the need for service-related data at subnational levels; about new strategies (designs) for oversampling certain groups; about "improved longitudinal data . . . to address . . . policy issues [on] changes in family

resources, predictors of successful development . . . precursors of serious problems; . . . " and about cross-agency planning and coordination.

Exhibit 1 catalogs the implications that are drawn in *Integrating's* summary. The crude enumeration suggests that the generic list of implications developed earlier seems reasonable. It is important to note that the implications are not drawn directly from new analytic methods nor are they drawn specifically from *any* method. They are drawn in unspecified ways from various and often unspecified analyses. In other words, the coupling between "implication" and analyses is often loosely specified. Brooks-Gunn, Brown, Duncan, and Moore (1995), for example, recognized that hierarchical models can be employed to analyze NELS:88 on account of this survey's design (p. 63).

Let us backtrack to the papers that were written for *Integrating* to understand more specific implications for NCES surveys. What do we learn? First, Brooks-Gunn et al. (1995) admired NELS:88 for the survey's attention to eliciting information from multiple sources, such as parents, teachers, children, and school administrators to produce data that help us to understand outcomes and inputs and process overtime. The only implications drawn by Brooks-Gunn et al. are that 1) the 1996 wave of measurement of children who would be in the 20—24 age range ought to be done; and 2) NELS:88 ought to be continued until the cohort is at the age of 28 or so (in the year 2003). The rationale for the authors lies in their view that transitions, from late adolescence to adulthood for example, are important. It is not based on identified data analyses or particular analytic models or methods (p. 76).

Hoffreth's paper (1995) in the same volume focuses on transitions to school. She then emphasizes the need for an entirely new survey, the Early Childhood Longitudinal Survey, that has been considered by NCES. The rationale is that we know less about entry to schools than we should. Further (p. 114), the United States has no longitudinal study underway that begins prior to entry to school. Hoffreth was attentive to linkage among data collection efforts but did not mention NCES in this context nor did she get much beyond the notion that data on mothers from the NLSY ought to be coupled with other data.

The Implications of Looking for Implications

This primitive review of scholarly published works that might have contained implications itself has implications, of course. The zero point implication is that the question posed at the outset of this essay was not put quite rightly. That is, getting beyond the initial question is important. We cannot be content with: "What do new analytic methods imply for NCES survey design?" We must ask the further question, "What are the implications of employing new analyses or old ones for design of NCES surveys?" Also, what do we mean by implications? And who articulates them?

First, we should not expect able scholars who *invent* new methods of analysis to educe and state plainly the implications of their work for designing better surveys. Attention to such implications is sparse in the current culture of mathematical statistics.

Second, we should expect fewer than half of the scholars who apply new or old methods to real NCES data to make suggestions (implications) about improving survey design. Further, we

should expect them to suggest: new variables or deeper/more sophisticated measurement of existing variables. The need to oversample certain groups or to measure the same way is, at times, reiterated.

Third, when the implications are stated at all, they are diverse and depend heavily on the analyst's idiosyncratic interests and theoretical perspective. When the stated implications are unclear, as some are, they can be perfectly uninformative and may require further action. The diversity means that NCES might develop methods for orderly acquisition and screening using vehicles that NCES has at its disposal (the Web, advisory groups, and so on). That is, many implications, can be generated and this is another peculiar problem. Some options for handling the problem via the Internet are described in the last section of this report.

Fourth, mechanisms exist to foster statements about implications of new analytic methods of employing new or old analytic methods to NCES data. The mechanisms include institutions such as NAS. They include grants, e.g., the Stanford group. They include professional organizations and journals to which NCES professionals contribute pro bono. NCES can encourage its contractors to educe the implications of their work for improving survey designs and can influence grant agencies to encourage grantees to educe the survey design implications in their research.

EXHIBIT 1

THE POSSIBLE DESIGN IMPLICATIONS OF A PARTICULAR METHOD, MODEL, OR ANALYSIS*

1) *What* new variables should be measured and what variable ought not be measured?

2) *How* or what should we measure?

3) *Whom* to measure?

4) How many?

5) When and with what *frequency*?

6) With what *periodicity*?

7) With what broad *design* (e.g., strata, and so on)?

8) In *connection*/*coordination*/*link* with what other survey, database, or experiment?

9) How to *report*?

10) *Why* for each of the above?

*For example, Mullis, Jenkins, and Johnson's HM analyses of NAEP data (1994) suggest that NAEP should better measure sets (Item 2), more instructional variables ought to be measured (Item 1), and information ought to be elicited from more teachers (Items 3, 4) in the interest of understanding the relative effects of classroom/teacher (Item 10).

EXHIBIT 2

INFORMATION NEEDS: TEACHER SUPPLY, DEMAND, AND QUALITY*

Information Need 1:	Teacher quality indicators
Information Need 2:	Teacher credentials
Information Need 3:	Demographic matching
Information Need 4:	Teacher professionalism
Information Need 5:	Programs to improve practice
Information Need 6:	Assessment of quality of teaching practice

*Excerpted from Boe and Gilford (1992).

EXHIBIT 3

SUGGESTIONS: TEACHER SUPPLY, DEMAND, AND QUALITY*

Suggestion 1:	Teacher quality indications; Sustained research
Suggestion 2:	Tested ability of teachers; Tests of knowledge
Suggestion 8:	Reserve pool; Little is known; Survey applicants in SASS
Suggestion 16:	Teacher demand data; NCES should develop a (better) model for teacher demand projections
Suggestion 18:	Unused databases (e.g., NSY and Supply)
Suggestion 19:	Linking SASS and state DBS
Suggestion 23:	TSDQ Consortium

*Excerpted from Boe and Gilford (1992).

EXHIBIT 4

THE IMPLICATIONS DRAWN BY THE NRC GROUP ON TEACHER SUPPLY, AND SO ON (TSDQ) THE NRC GROUP ON INTEGRATING FEDERAL STATISTICS*

		TSDQ	Summary Integrating
1)	Variables: New/deleted	Yes/no	Yes/no
2)	Measurement	Yes	
3)	Sample units		Yes
4)	Sample size: Increase/decrease	Yes/no	
5)	Time		
6)	Timing		Yes
7)	Survey design	Yes	Yes
8)	Links	Yes	Yes
9)	Reports		
10)	Rationale	Yes	Yes/no

*For example, *Integrating Federal Statistics on Children* (Board 1995) educes implications from other research for the design of new surveys. The implications cover sample size (e.g., oversampling Hispanics) and links (e.g., to state databases). Some implications bear on NCES efforts and they are identified by a "yes" in the column "Summary Integrating."

CROSS-DESIGN SYNTHESIS: IMPLICATIONS FOR THE DESIGN OF EDUCATIONAL SURVEYS AND CONTROLLED FIELD EXPERIMENTS

Background: Cross-Design Synthesis

Cross-design synthesis is a strategy for combining analyses of the data that are generated in controlled experiments with analyses of data generated from surveys or from certain administrative databases. For example, the national data obtained in a NCES probability sample survey on adult literacy in the United States might be used in an analysis that purports to yield estimates of the relative effects of certain literacy programs. The results would then be combined with evidence generated by a dozen experiments on the relative effectiveness of local literacy programs.

The object of this combination of evidence is to produce valid and generalizable estimates of the effect on certain social programs. The rationale for combining the different data sources is that the combination exploits a benefit of controlled tests, notably an unbiased estimate of the treatment effect in local settings, and further exploits a benefit of national probability sample surveys of the kind that NCES executes, the capacity to make generalizations to a larger target population.

In the adult literacy case, controlled experiments in particular sites may yield valid estimates of the effect of literacy programs. But the estimates are local, e.g., of uncertain generalizability. The national database or survey may yield estimates of the effect of programs at the national level. These latter estimates are suspect in that their validity is unclear; the survey or administrative database involves no active control. Rather, analysis usually involves statistical control. A combination of the two sources of evidence might be combined so as to justify inferences that are both valid and generalizable.

The general approach to cross-design synthesis is described in a U.S. General Accounting Office report (USGAO 1992) and in Droitcour, Silberman, and Chelimsky (1993). A more recent report (USGAO 1995) describes the approach's application to the problem of estimating the effect of breast conservation versus mastectomy on the 5-year survival rates of women with breast cancer. This analysis is based on data from randomized clinical trials and a large database. In particular, six studies serve as the evidence in the randomized trial category; they include single-site and multisite experiments undertaken in North America and Europe. The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) system constitutes the administrative database. It provides data on breast cancer patients, their treatment, and prognosis based on reports from practicing physicians in a large geographic region of the United States.

Objective and Assumptions

Recent reports on cross-design synthesis have focused on the analysis of data from two kinds of sources: controlled experiments and databases. Here, the focus is on how the thinking about cross-design synthesis can improve the design of administrative databases and national surveys sponsored by NCES.

To put the objective bluntly, we want to turn "cross-design synthesis" into a vehicle for better design of studies, rather than to encourage its current use as a form of meta-analysis. This objective accords with the theme of this report, i.e., educing the implications of new analytic approaches for better study design. It is also distinctive; the inventors of this analytic approach did not develop this implication (Droitcour and Chelimsky 1995; Boruch 1995).

A first assumption is that it is important to estimate the effects of education programs in the United States. The second assumption is that NCES cannot or should not undertake formal evaluations of the effects of such programs. Other federal agencies, for example, are responsible for running controlled experiments on education programs. Third, we assume that NCES can design surveys that accord with the first two assumptions. Finally, we assume that, in 5 years, we will have to combine results from different sources to reach a conclusion about a program's effectiveness.

The object here is to address the question: How can NCES exploit ideas in the cross-design synthesis approach so as to design better surveys or databases?

Definitions

Survey here means an effort to elicit information from a probability sample of individuals or institutions who are members of (ideally) a well-defined target population. Such a survey involves no active treatment or manipulation of respondents, apart from the act of eliciting information. The survey may be cross-sectional, for example, the NCES 1991 National Adult Literacy Survey. Or, the survey may be longitudinal, as in the case of the National Educational Longitudinal Study undertaken in 1988 (NELS:88).

Administrative database here is defined as a set of administrative records on a well- defined target population. For instance, transcripts on all students in a junior college, containing information about the students' courses and grades, constitute a database. The records on all students in a voluntary service organization's program on literacy also constitute a "database."

A database is a survey of a special kind. It usually includes the entire target population; no sample is taken. It is a "survey" to the extent that any set of administrative records is the product of interviews of a kind that are often done in survey research, albeit under different conditions.

Certain kinds of NCES data collections result in a database of administrative records for eligible institutions in a population. That is, the NCES effort is not based on a sample. The population databases include the Common Core of Data (CCD), the Integrated Postsecondary Education Data System (IPEDS), and the Library Statistics Program. In principle, analyses based on data from any of these sources could be combined with results of controlled experiments in a cross-design synthesis.

A controlled field experiment is a setting in which individuals (or other entities) are assigned to program variations in accord with a plan designed to produce an unbiased estimate of the differences among the program variations and a statistical statement bearing on one's certainty about the results. For instance, one may design a study to compare certain approaches to teaching

English as a second language so as to understand which approach works best, and under what conditions. Individuals or entire organizations might then be randomly assigned to the different program approaches, engaged in the relevant approach, and then measured with respect to their English proficiency.

Because controlled experiments are difficult to mount, only a few are undertaken in a very small number of sites. The results may be relatively unequivocal in the sense that one variation appears to work better than another in one or more of the sites. It will usually not be clear how these results can be generalized. For instance, the experiment sites may include cities in the Northeast; they may exclude the Northwest and Southwest.

An agency such as NCES is mandated to conduct observational surveys. It is not mandated to execute controlled tests of education programs. Other agencies within the U.S. Department of Education, such as the Planning and Evaluation Service, are mandated to conduct controlled experiments to evaluate education programs. Further, private foundations and other government agencies may exploit surveys or experiments or databases to further knowledge about programs or about the educational state of the nation.

Rationale in the NCES Context

The first rationale for focusing on cross-design synthesis is as follows: Users of NCES survey data have often tried to use the data to estimate the relative effectiveness of different sorts of education programs. It seems reasonable to expect these efforts to continue despite the ambiguity in the interpretation of the data that is bound to occur because the survey is a passive instrument rather than an active experiment. Insofar as cross-design synthesis carries a promise to combine such survey data with other data from experiments, so as to produce better information, it is sensible for NCES to exploit opportunities presented by cross-design synthesis.

A second rationale is more ambitious. It is that cross-design synthesis can be a vehicle for the mutual education of survey researchers and experimenters and a productive change in scientific culture. Thoughtful survey researchers cannot always be well informed about controlled field experiments. For example, Clifford Clogg (1989), a sociologist and survey statistician, announced that “experimentation of the classical variety is usually impossible, inconceivable, or difficult to implement.” Economists and educational researchers, such as Henry Levin, and mathematical economists, such as James Heckman, who rely heavily on observational survey of that sort that NCES produces, have made similar claims. They rarely present empirical evidence (see Boruch 1994 and references therein).

Experimenters, on the other hand, depend in only a limited way on survey data of the kind that NCES obtains. Their design of a local controlled experiment on the relative effectiveness of two compensatory literacy programs may, for example, depend on regional or state literacy rates to inform the experiment’s design. As a consequence, experimenters are at times not well informed about surveys run by NCES or other statistical agencies. Few important controlled experiments in the United States rely heavily on surveys run by federal statistical agencies except at the

experiment's design stage, where the experiment and design may recognize survey-based estimates of the incidence of a problem.

A More General Rationale: Government Agencies

A broader reason for inverting the analytic idea of cross-design synthesis so as to focus on design of surveys is that the approach can be a fine bridge between the members of the federal statistics agencies on the one hand and the federal evaluative agencies and private foundations that sponsor controlled experiments on the other. These include, for instance, the Bureau of Justice Statistics responsible for the National Crime Victimization Surveys, and its sister agency, the National Institute of Justice which is responsible for multisite controlled experiments on the police handling of domestic violence, among other topics. It includes the Bureau of Labor Statistics, an agency that continues to run large-scale probability sample surveys on employment and training and the Department of Labor's unit for large-scale experiments on residential Job Corps, the Job Training Partnership, and others. The role of NCES as statistical agency is complemented by the role of the Planning and Evaluation Service at the Office of the Undersecretary at USDE.

The gap between the statistical agencies and the other units that focus on analysis represents a kind of intellectual travesty in this country, given that data from the former *are* often used to estimate program effects, not just to describe them. The insulation of statistical agencies such as NCES has considerable political justification, of course. Statistical data should be and, under current laws, is relatively free of political influences. Analysis units are more vulnerable to the latter although some have a fine reputation for both independence and political sensitivity. The institutions need to keep the two functions separate. But this does not vitiate the idea that as an intellectual matter, the separation is unnecessary and arguably dysfunctional.

The gap between the statistical agencies and those responsible for analytic studies of programs was recognized implicitly and explicitly in a NRC volume on integrating statistics on children. Brooks-Gunn et al. (1995) and Hoffreth (1995), for instance, recognized the distinctive role of the JOBS experiments and the Perry Pre-School Project in the context of NCES and other surveys but did not explore the matter deeply. Pallas (1995, p. 153) recognized the merits of NCES and other statistical systems and the distinctive role of experiments on dropout prevention programs, and more importantly, expressed discomfort with the volume's heavy emphasis on statistical systems. It is a discomfort that we share, discussed briefly in a paper on the future of experiments (Boruch 1994), and explore here.

The First Illustration in the NCES Context

The NCES has undertaken a national probability sample survey of adult literacy in the United States with augmentation for special subpopulations, e.g., prisoners. Reports on adult literacy are available from Andrew Kolstad's Education Assessment Division at NCES (see Davis and Sonnenberg [1995] and other NCES *Programs and Plans*). Suppose that the NCES will run another such survey and that the survey's plan can be influenced.

The U.S. Department of Education's Planning and Evaluation Service, Office of the Undersecretary, has had a responsibility for evaluating the effectiveness of certain adult literacy programs. Suppose that another evaluation at multiple sites will be undertaken by this office.

Regard the NCES survey on adult literacy and any other information obtained by NCES from administrative sources as a database. Regard the USDE/PES evaluation as a source of data generated by controlled experiments.

Consider then the question: How can the cross-design synthesis approach inform the design of new surveys or databases (and experiments) in the adult literacy arena so as to generate better estimates of the effect of literacy programs in 5 years?

The GAO reports on cross-design synthesis approach suggest that in the survey and in controlled experiments we attend to the following:

- Target population and its characteristics;
- Treatments;
- Outcomes; and
- Propensity scores.

Each is considered in the section that follows.

Implication: Target Population and Samples

Cross-design synthesis requires that the individuals who are targeted in controlled field experiments are also represented in the survey sample or database.

A new NCES sample survey on adult literacy in the United States must then include individuals who are targeted for literacy services. Attempts to estimate the effect of the services, undertaken in local controlled experiments, must target similar individuals.

For instance, if programs make major efforts to serve illiterate immigrants from Bosnia, Slovakia, Morocco, or other countries, then NCES must plan to include these in the target population for a new NCES survey. This, in turn, requires that the local literacy agencies be able to specify their main local targets. It implies that the federal agency responsible for support of adult literacy programs, an agency different from NCES, be able to specify target population that is of major interest in any controlled experiments that are undertaken to test the programs.

Implication: Treatments

To combine data in the cross-design synthesis approach, one must know what treatments (programs) are delivered to whom and when. A new sample survey of literacy in the adult

population undertaken by NCES then would have to ask individuals about the literacy programs in which they have participated. Learning *how* to ask such a question so as to secure reliable responses is difficult, to be sure. Figuring out how to exploit local databases of literacy services that maintain such information is also likely to be difficult. Nonetheless, NCES must do so if the object is to produce a cross-design synthesis in 5 years, of who gets what literacy program and to what effect.

For a federal agency or private foundation that sponsors controlled experiments on the effects of certain literacy programs, the implication is that the agency or foundation must record the individual's program participation. More important, the method of recording must correspond with how the NCES national survey asks about program participation. Questions about program participation are framed in a survey and the way they are framed in local experiments must be compatible with one another. The local experiments will usually depend on administrative program records to establish an individual's participation in a certain program. A survey usually involves depending on an individual's self-report about participation in a program; it may also depend on institutional records contained in databases.

To make the two kinds of information compatible for cross-design synthesis, several options might be considered. The local experiments might ask about participation in the same way that the survey asks, permitting one to correlate self-reports with administrative records. Or, both the survey and the experiments might direct attention to local service providers and their clients, eliciting records so as to reduce reliance on self-reports of individuals. In any case, small studies of the matter are needed.

Implication: Outcomes

The impact of adult literacy programs can be registered partly by measuring an outcome variable such as "literacy level" of each individual or of groups of individuals.

To accomplish a cross-design synthesis of the effects of literacy programs, a survey agency such as NCES must cooperate with an evaluation agency such as USDE/PES or a private foundation that sponsors evaluations in developing outcome measures. That is, the organizations must agree on how literacy level is to be measured.

Cooperation of this sort is not easy across local literacy programs, much less across federal agencies or private foundations. For instance, a recurring problem is that local literacy programs, regardless of their sponsorship, have not been able to agree on how to measure literacy. In the absence of agreement, no surveys or experiments undertaken by the federal government are likely to lead to a persuasive cross-design synthesis of whether and which programs work in what sense.

Implication: Propensity Scores

A controlled randomized experiment relies on randomization to produce an unbiased estimate of the difference between two or more groups. In such an experiment, individuals who are eligible to be served by a literacy program and who are willing to avail themselves of the program are randomly assigned to the program or to one of two or more variations of the program. Or, entire

organizations might be allocated randomly to alternative service programs. In ordinary language, the groups are "equivalent" apart from chance because they were randomly composed. A comparison of the groups' performance is then fair. The difference in average literacy level of the two groups following their engagement in the programs, or difference in rates of achievement then provides a good estimate of the relative effectiveness of the program variations.

The NCES does not sponsor controlled randomized tests of literacy policies or programs. NCES does, however, provide an observational survey data platform for estimating effects. Statistical analysts who rely on such a platform have usually developed strategies to approximate the results of a controlled experiment, i.e., compensate for the absence of the randomized test. The strategies vary. During the 1960s, for example, analysts employed OLS estimates of a program effect that was based on a simple, single-stage linear model and observational data (e.g., covariance adjustment).

The focus here is on propensity scores as a device to produce analyses that approximate the results of a controlled test. Such scores were used, apparently to good effect, in the GAO (1995 and Appendix I) report on the differences between two approaches to treatment of breast cancer. The recent work on propensity scores has the benefit of conscientious thinking about how to recognize the fact that people, in ordinary circumstances, do not engage in programs randomly, and how to incorporate this and related selection factors into analysis.

The GAO's application of cross-design synthesis to data on treatment of breast cancer suggested the following were important in developing propensity scores:

1) Year at which the individual is engaged in treatment;

2) Geographic area of residence;

3) Severity of the problem at baseline;

4) Age of the individual;

5) Race or ethnicity; and

6) Marital status.

How and why the variables were chosen is not made plain in the GAO's report (1995).

These same variables *seem* relevant nonetheless to understanding the propensity of individuals to engage in adult literacy programs. The access to such programs was greater in 1990 then it was in 1980, and the efforts to entrain clients has arguably been more vigorous in the past few years. Year of engagement then is arguably important. The geographic area of residence and ethnicity are related and theorists argue that it is important to recognize each. For example, Hmong immigrants have clustered in only a few cities in the west, midwest, and northeast United States. Bosnian immigrants and others from the new independent states of the former USSR make their homes elsewhere.

Marital status may have no obvious influence on one's inclination to become literate. But a conscientious theorist might argue that if one examines the way families develop once marriage occurs, the way adults in the family behave in their children's interest and in their own economic interest, the variable called "marital status" may be a reasonable one to use in constructing a propensity score.

Implication: Propensity Scores, Intentions, and Reasons

Roughly speaking, a propensity score reflects the predilection of individuals to belong to one group rather than another, where the predilection is indicated by some observable characteristics of the individual. More specifically, it is the conditional probability of being in a particular group given a vector of observed covariates (Rosenbaum and Rubin 1983).

For example, high school dropouts and high school stayers constitute two groups. The probability of being in one group or the other can be characterized descriptively as a function of variables such as daily school attendance rates, age, academic grades, and plans for higher education. Similarly, the probability of entry to college or the work force can be characterized as a function of demographic and other variables.

The variables typically used to estimate a propensity score usually include demographic and contextual information. Over 30 such variables were used by Rosenbaum (1986) to estimate a kind of propensity score for school dropouts and stayers. They included those identified in the paragraph above.

The variables used to compute a propensity score are often "indirect" in the sense that they indicate an individual's state, rather than capturing directly: 1) an intention to belong to one group or another, or 2) the observable reasons for belonging to one group or another. Education surveys, with a few important exceptions, do not ask individuals why they dropped out of school or about their intentions to do so.

An implication of the analytic work on propensity scores (and related analytic methods) is that we should consider obtaining information on the individual's intention or on the reasons for membership in a group or both. One rationale for obtaining such information is that it *appears* to be a more direct covariate of membership than less direct ones, such as demographic characteristics. The connection between an individual's declaring that he or she will drop out of school and actually doing so appears more direct, less distant, from actual membership in the dropout group (i.e., becoming a dropout) than, say, the connectedness between "age" in school at one point in time and becoming a dropout in another.

Usually, no formal educational theory underlies the construction of propensity scores. Rather, the justification for their use lies in small and large sample statistical theory (Rosenbaum and Rubin 1983). A second rationale for eliciting information about intentions or reasons then lies in the need to construct better substantive theory in education. To the extent that the propensity approach can be informed by education theory and can help build the theory in a cyclic way, this seems desirable. Better theory, for example, may promote propensity scores that are easier to compute or more

interpretable. They may decrease the need for a large reservoir of cases on which to match when propensity scores are used with matching. This promotion may hinge on eliciting information about intentions or reasons.

Sensible readers can quarrel with the idea that information about reasons or intentions ought to be elicited in surveys. Critics do so with considerable justification. Asking individuals about intentions and reasons is difficult and, in any case, may not be useful. For instance, Rosenbaum's exploration (1986) of a propensity-scorelike approach in a dropout study using NCES' High School and Beyond data uncovered the fact that "the vast majority of students who eventually dropped out said in their sophomore year that they expected to graduate" (p. 208). Was the question asked well? We do not know. We do know that other "intentions" question, about aspirations beyond high school, was indeed useful to Rosenbaum in constructing the propensity score.

At least some scholars would argue, based on good evidence, that the more general problem is of understanding revealed preferences and their usefulness in studies based on observational data. Manski's book (1995) has a chapter dedicated to this and related matters. The NCES' National Longitudinal Study of the High School Class of 1972 (NLS−72) served as a vehicle for his attempts to understand how college enrollment rates would be affected by Pell Grants to needful students. The variables he used as a surrogate for revealed preference included ability, income, and so on as measured in NLS−72.

Recognizing the skepticism that economists have about self-reported preferences, Manski argued persuasively for trying to measure the preferences directly. Part of the argument is tied to theory, notably theory about what variables to use in an analysis. Economists vary, for example, in the variables they have included in studies of returns to schooling (p. 97). Manski's argument is based partly on empirical grounds. He provides citations to research in the arenas of consumer buying intentions, fertility (based on Current Population Survey over the last 50 years), and voting intentions, and to work by social psychologists in the arena to justify his argument that preferences ought to be assessed more directly.

For Manski, one of the implications of agreeing that information on preferences is important in that we must get beyond simple "yes" and "no" answers, e.g., "Do you think you will drop out of school?" He argues, on analytic and empirical grounds, for eliciting a probabilistic assessment of behavior from each individual. To paraphrase his sample question: "Looking ahead, what percent is the chance that you will drop out?" Social psychologists working in the arena would probably go further to argue for eliciting preferences (self-predictions) at points in time that are close to the event in question. Asking in September about students' perceived probability of dropping out is arguably less useful than asking the question in November or December.

To summarize, propensity score approaches suggest that 1) we consider more seriously whether to measure preference (self-declared propensity), and 2) how and when the preferences are measured seems important. But we need to do research on this.

Similarly, one may argue that to do a better job constructing propensity scores, one ought to observe or elicit information on why or how people find their way into groups, e.g., into a literacy program or not. To return to the main illustrative context, NCES might then ask a question of the

following sort: “Which of the following factors influenced your decision to enroll (or not enroll) in the literacy program?”

The responses to the question might then be incorporated into a propensity score that is better than (say) one that relies solely on demographic information. Further, the responses may help to develop a small part of a substantive education theory that helps to understand processes by which people enter programs or, more generally, a substantive theory that complements or augments statistical theory for analysis of observational data.

A question of the sort proposed above appears not to have been asked in any large-scale observational surveys, nor can we find concrete illustrations in the published reports on selection modeling or propensity scores (e.g., Rosenbaum and Rubin 1983; Rosenbaum and Rubin 1984; Rosenbaum 1989). Ways to frame such a question can be developed, based perhaps on NCES expertise and cognitive research in a laboratory or field setting.

Implication: Measurement Issues

In national probability sample surveys, we can often measure a variable using only one or two questions or using an inventory with very few items. Learning about children’s relations with other children in a survey might, for example, involve only a few questions about (say) how many friends that the child says he or she has. A set of local experiments designed to test ways to improve the ability of withdrawn or hostile children to relate to other children usually involves a more elaborate inventory. It is not clear how to link the data from sparse measures made in a large sample survey to the deeper measures made in the small sample experiments.

Similarly, learning about literacy level of individuals in a large sample survey must contend with respondent burden. Local experiments can often depend on inventories that demand more time of the individuals who participate, and do.

The problem here has a delicious analogue in atmospheric weather research. Satellite imaging might be based on measures on grids that are 1,000 kilometers in width. Surface measures may be obtained in far smaller grids, 100 kilometers across for example, yielding more precise local measurement. The challenge lies partly in how to integrate these data across levels of resolution (Draper et al. 1992).

Learning how to measure simply in large-sample surveys and how to measure roughly the same construct with more precision in local experiments are important. Cross-design synthesis and the problem of combining different sources of information generally, invites us to learn how to link the two sources.

Summary

NCES has taken a leadership role in arenas related to cross-design synthesis. This strength suggests that it can succeed in work based on design orientation to cross-design synthesis. For example, the Common Core of Data (CCD) is a substantial product of NCES’ efforts at the national

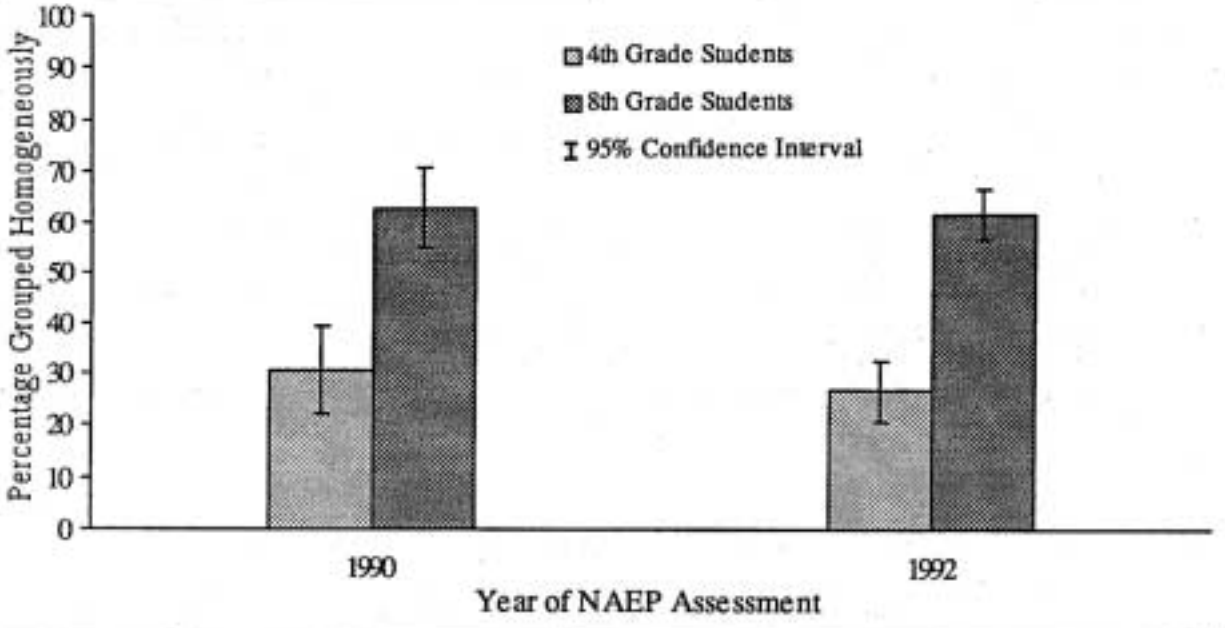
level to develop agreement among the states about what ought to be measured and how what is measured ought to be defined so that resultant data are interpretable.

This achievement is no small one. The experience in negotiating with different jurisdictions and the product are valuable. Both can be capitalized in exploiting a cross-design synthesis approach.

A Second Illustration: Ability Grouping

Schools often sort students by ability (homogeneously) in math classes, particularly in higher grades. As Figure 1 shows, nationally representative data indicate that schools[1] grouped a significantly[2] higher percentage of 8th grade students than 4th grade students by ability in 1990 and 1992.

Figure 1—Percentage of homogeneously-grouped 4th and 8th grade public school math students in 1990 and 1992



SOURCE: U.S. Department of Education, National Center for Education Statistics, *NAEP Data on Disk: 1992 Almanac Viewer*.

Numerous propositions that attempt to explain why ability grouping increases in higher grades seem plausible. Students of mixed ability, for example, may receive academic instruction in several subjects from one teacher in elementary school (i.e., K-5), reducing the possibility of homogeneous grouping. As students reach middle school (i.e., 6-8), however, they may receive instruction in several subjects from several teachers. It may then become more convenient to group by ability; that is, to reorganize heterogeneous groups of students into homogeneous ones. Or differences in achievement may accumulate as students age, becoming more pronounced in later grades, thereby

creating the perceived need for homogeneous grouping. Or schools may intentionally or otherwise sort students by socioeconomic status, gender, and race, as some critics of ability grouping have charged. Or decision makers may believe (perhaps on the basis of research evidence) that comparable students, particularly older ones, learn better in homogeneous classes. Hereafter, this paper will attend primarily to the latter two propositions.

The Concerns of Equal Opportunity Advocates

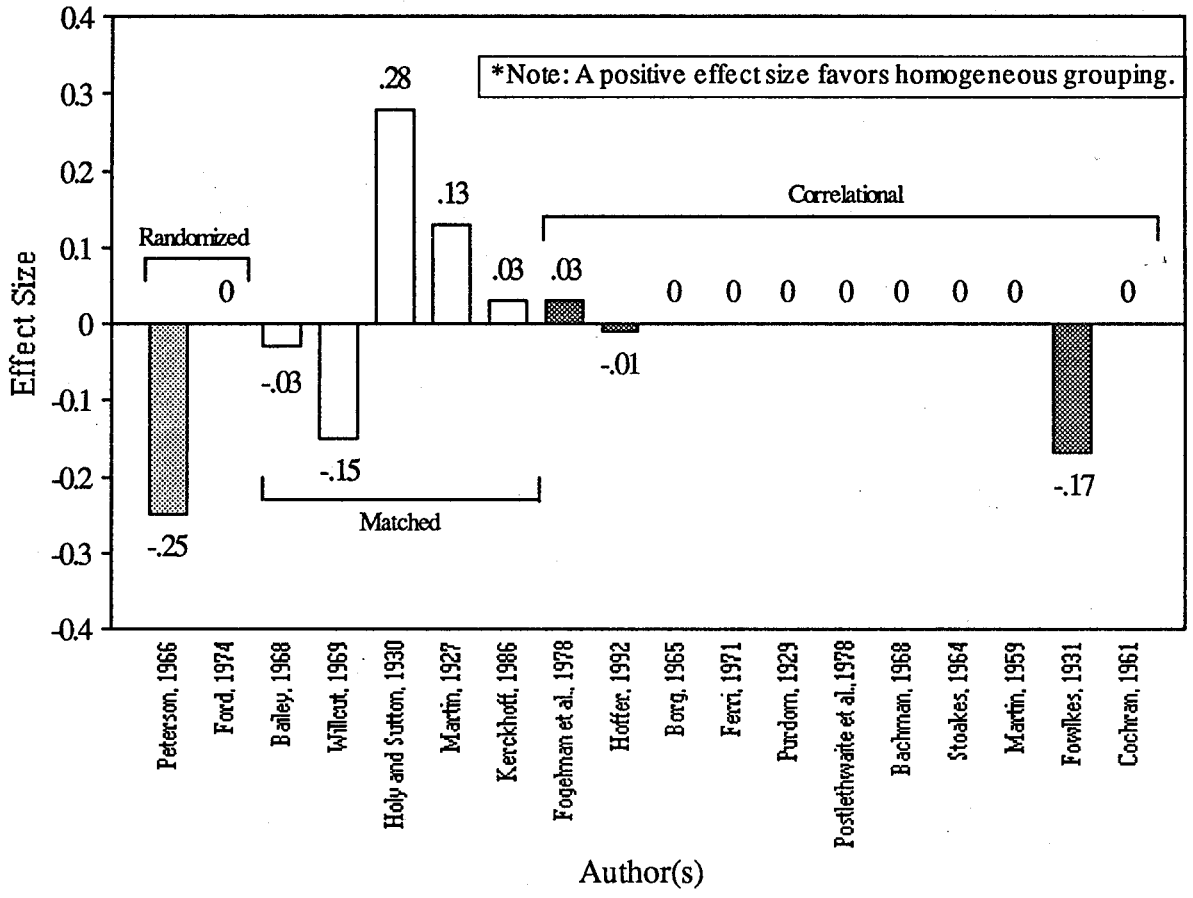
Advocates for equal opportunity often assert that two tracks—one leading to prosperity and the other to poverty—exist in America's schools. That these tracks appear to reflect gender, racial, and socioeconomic differences is cause for alarm. "As a result of the two track system," Beatrix Hamburg, president of the William T. Grant Foundation, writes: "[T]here is educational neglect and underachievement that disproportionately afflicts girls, minorities, and the poor" (1993, p. 9). And "what purpose has desegregation served," Jay Heubert, an attorney and education professor at Harvard University, adds "if resegregation takes place within desegregated schools?" (personal communication, November 1992). Ability grouping, from their collective perspective, may be viewed as one vehicle through which differences along gender, racial, and socioeconomic lines are bred and perpetuated. And indeed, some evidence supports this view. Oakes (1990), for instance, has observed that

- Schools tend to disproportionately place black, non-Hispanic and Hispanic students in lower ability groups;
- Ability groups tend to reflect socioeconomic status;
- Teachers of low ability groups tend to expose students to fewer, less demanding, topics than do teachers of high ability groups; and
- Schools tend to place their least qualified teachers in low ability classes and their most qualified teachers in high ability classes.

Researchers Concerned with Student Achievement

Those concerned with student achievement, meanwhile, often assert that ability grouping either impedes or adds no value to overall math achievement. Understanding whether this is so suggests the use of experimentation. In question form: If a sample of students were randomly assigned to homogeneous and heterogeneous instructional groups, which group would achieve at a higher level? Asking the question is the easy part. Mounting randomized experiments has turned out to be more difficult—there have been none since 1974—and there are only two on record. But there have been several non-randomized (i.e., "matched" and "correlational," in Slavin's terms [1993]) efforts to estimate the impact of homogeneous grouping on math achievement. Slavin (1993) included 16 such studies, plus the two randomized experiments, in his "best evidence synthesis." Slavin found the mean effects of homogeneous grouping to be near zero for the 18 studies. Figure 2 displays each study's effect size estimate.

Figure 2—Effect Size Estimates of Middle School Math Studies That Compared Homogeneous and Heterogeneous Grouping



SOURCE: Slavin, R. 1993. "Ability Grouping in the Middle Grades: Achievement Effects and Alternatives" *Elementary School Journal* 93 (5), pp. 535–552.

But Do the Findings Generalize?

Some researchers (e.g., Elmore 1993) have questioned the potential of evidence from a "best evidence synthesis" to inform practice. Implicit in this question is the notion that a "best-evidence synthesis" (or meta-analysis) contains insufficient evidence to generalize to other settings. This notion is not entirely accurate. Slavin, for instance, includes one analysis (Hoffer 1992) that made use of data from the Longitudinal Study of American Youth (LSAY), a 4-year, large-scale study. He also says that other such studies "provide important additional information not obtainable from the typically smaller and shorter experimental studies" (Slavin 1993, p. 539). However, Slavin does not discuss the premise underlying cross-design synthesis, namely, that evidence from experimental studies and observational studies might be combined to generate more national estimates of effect. LSAY, however, may not be the ideal study for this purpose insofar as ability grouping is concerned. Adequate data, for example, were available from only 1,800 8th grade math students. NAEP's Trial

State Assessment, in comparison, collected data from about 2,500 8th grade students from *each* state.

What Does NAEP Reveal About Ability Grouping?

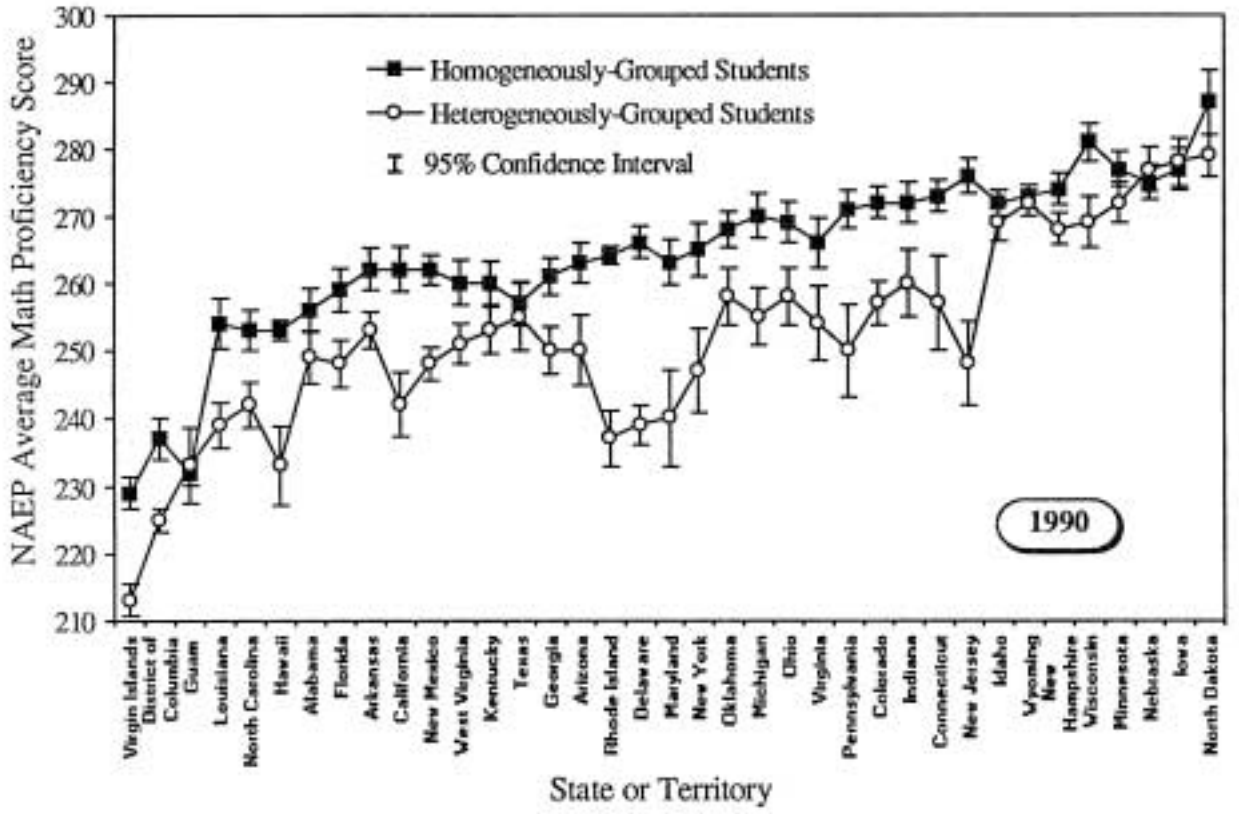
NAEP provides information on student achievement to local, state, and federal policymakers on a biennial basis. It also provides background information on students, teachers, and school administrators. Some NAEP information is demographic, while other information concerns educational practices and policies. NAEP allows policymakers to know, for example, whether student achievement is related to ability grouping. NAEP is an observational study, however. Making statements of impact or effectiveness on the basis of NAEP data is therefore inappropriate without some adjustment. It is imprudent to assume, for example, that students who are grouped by ability are comparable in all ways to those who are not. Schools, for example, may tend to group higher achieving students by homogeneous ability rather than heterogeneous ability, thereby causing an imbalance between groups that may bias achievement-based comparisons. *Unadjusted* NAEP data indicate, for example, that homogeneously grouped 8th grade (public school) math students outperformed their heterogeneously grouped counterparts in 34 of 37 jurisdictions (significantly in 27 of 37) in 1990, in 43 of 44 jurisdictions (significantly in 34 of 44) in 1992, and nationally during both testing years, as Figures 3, 4, and 5 show.

The Need to Adjust NAEP

If one is to use NAEP data to estimate the effects of ability grouping, then one must first employ a substitute for the randomization of controlled experiments, i.e., to assure that the groups do not differ systematically. The focus here is on a "propensity score" adjustment—a technique to produce analyses that approximate the results of a controlled experiment. As applied to the example of ability grouping, the analyst's first task would be to develop a statistical model—on the basis of theory, following the lead of others (e.g., see Hoffer 1992), possibly through stepwise logistic regression, or through some combination of the three—to compute each student's probability of being grouped by ability (homogeneously); that is, to compute each student's propensity score. This approach may benefit from recent advances in the statistical theory for estimating multilevel models. Version 4 of Bryk and Raudenbush's hierarchical linear modeling software, for example, will enable analysts to model categorical dependent variables while taking into account the multilevel nature of NCES data.

After deriving propensity scores, the analyst's next task would be to divide the entire sample into quintiles on the basis of these scores; that is, to subclassify students on the basis of their propensity scores.[3] The analyst could then compare the achievement levels of subclassed ability-grouped (homogeneous) and non-ability-grouped (heterogeneous) students. In a sense, this procedure would generate five estimates of the effect of homogeneous grouping. An example of one possible interpretation is as follows: *With respect to students who were most likely to be grouped by ability, no difference in achievement exists between those who were actually grouped by ability (homogeneously) and those who were not*. The analyst could then combine the estimates by taking

Figure 3—Estimates of math proficiency for grade 8 students by type of instructional grouping (i.e., homogeneous or heterogeneous) for each state or territory in 1990



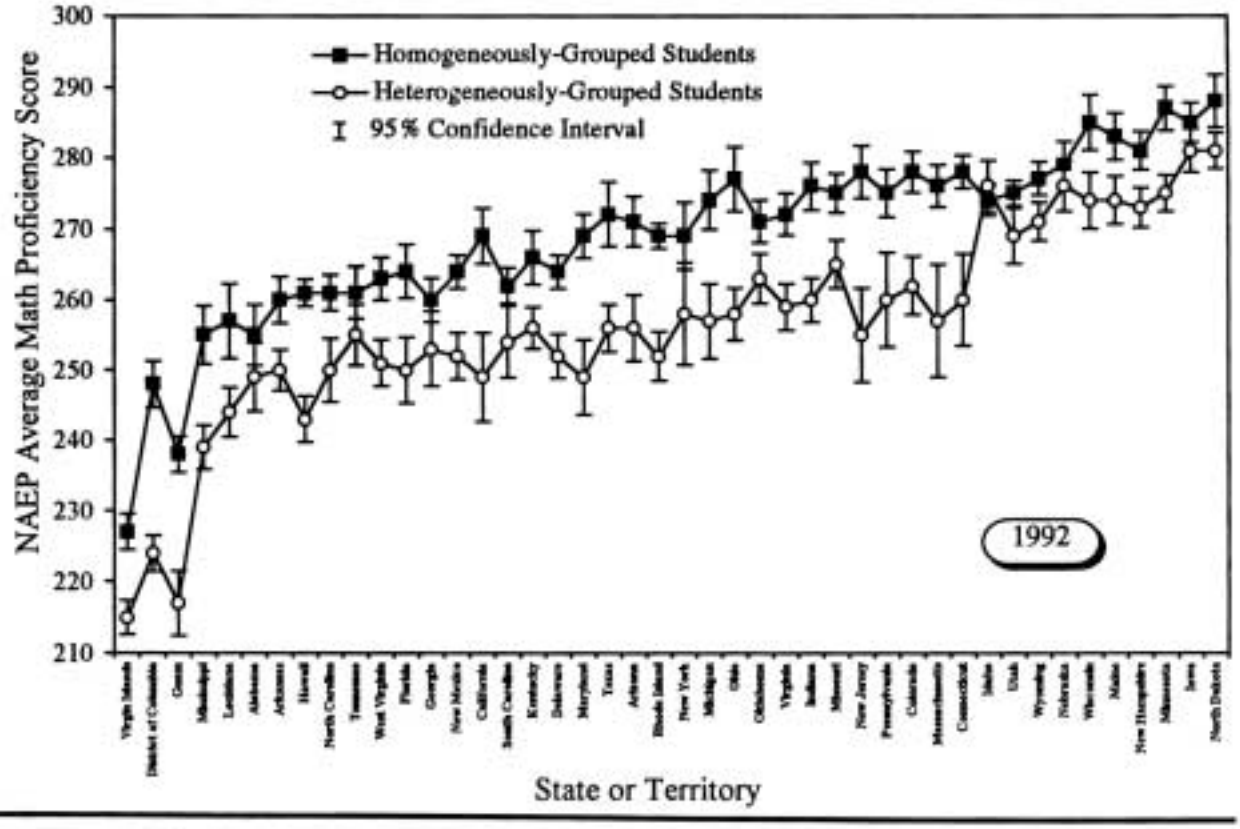
SOURCE: U.S. Department of Education, National Center for Education Statistics, 1993. *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States*, p. 463.

the average of the five effects, as in meta-analysis. This final estimate would be far more trustworthy than any of those that were displayed in Figures 3, 4, and 5.

Combining Contradictory Evidence: A Potential Problem

How one might *combine* estimates of effect from experiments and one or more NAEP analyses, particularly when the estimates are contradictory, is unclear. Although the GAO's introduction (1992) to cross-design synthesis discusses the problem and presents several options, it concludes that "many refinements are still to be developed" (GAO 1992, p. 96). The lone illustration (GAO 1995) of a cross-design synthesis, however, does not attempt to develop these refinements. The meta-analytic literature, which merited considerable consideration in the GAO's introduction (1992), also merits consideration here.

Figure 4—Estimates of math proficiency for grade 8 students by type of instructional Grouping (i.e., homogeneous or heterogeneous) for each state or territory in 1992



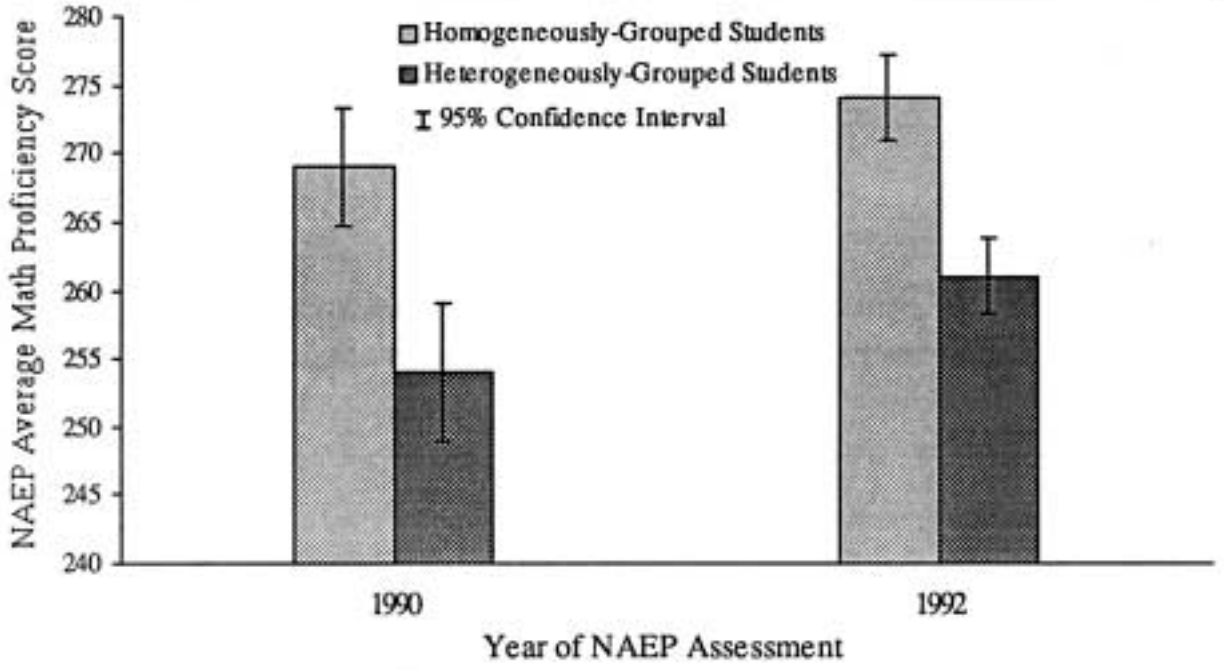
SOURCE: U.S. Department of Education, National Center for Education Statistic, 1993. *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and the States*, p. 463.

Meta-Analytic Strategy

In meta-analysis (see Hedges and Olkin 1985; Hunter and Schmidt 1990), the analyst computes one or more overall estimates of effect after 1) collecting, 2) coding, and then 3) weighting each study's effect size by its sample size; that is, the analyst computes a weighted average. This weighting scheme poses an analytic problem in the ability grouping example, however, on account of the size and nature of the sample of studies available for analysis. Put into question form: Does each adjusted NAEP state sample (of about 2,500 students) deserve to be weighted by 30 or so more times more than the smallest (Ford 1974, n=82) experimental study? The answer is probably not.

One approach to the problem is to divide the entire set of studies by design category (i.e., observational, randomized) prior to weighting studies within each category. After doing so, it then seems sensible to follow Hedges' and Olkin's advice. The general strategy that they recommend, as applied here, would be to do separate tests of homogeneity for the two sets—1) 44 state-level

Figure 5—National estimates of math proficiency for grade 8 students by type of instructional grouping (i.e., homogeneous or heterogeneous) in 1990 and 1992



SOURCE: U.S. Department of Education, National Center for Education Statistics. *NAEP Data on Disk: 1992 Almanac Viewer.*

NAEP analyses for 1992, and 2) two randomized experiments—of effect sizes. If, for the NAEP analyses, the null hypothesis of no difference is rejected (i.e., if significant random variation exists among the 44 effect size estimates), the analyst might then include additional covariates in the model to attempt to explain the variation. States with low teacher-student ratios, for example, may produce a small positive effect for ability grouping while those with high ratios may produce a large, negative effect. Teacher-student ratio (e.g., low or high) may therefore account for the variation beyond that expected from sampling error alone among all state-level effect sizes. The analysis, then, would generate two indices of effect for the NAEP analyses. Combining the two, however, would be inappropriate. The analysis would also produce (at least hypothetically because there are only two randomized studies) one or more estimates of effect for the experimental studies. This estimate or these estimates will be distinct from that, or those, produced by the experimental studies.

In this framework, it would also be possible to include in the analysis additional observational studies. Combining Hoffer's findings (1992) on the comparative effectiveness of homogeneous and heterogeneous grouping with those from the potential NAEP analyses, for instance, is one possibility. A reanalysis of Hoffer's data may be in order, however. Although Hoffer makes use of propensity scores, he does not use them to directly compare homogeneous and heterogeneous groups.

To summarize, *it may not be possible* to combine estimates of effect from experiments and one or more state-level NAEP analyses when the estimates contradict one another, particularly when there are very few experimental studies available for synthesis. As the sample of available experimental studies increases, however, the possibility of combining estimates across design categories also increases.

Implications of Research on Grouping

There are two broad implications of this illustration. First, there is an obvious need for more randomized experiments. Second, we will never know whether the apparent performance difference between homogeneous and heterogeneous groups is real without deeper analysis. It seems important to carry out the analysis, however, because the percentage of 8th grade math students who were grouped by ability decreased in 30 of 36 states (that participated in both NAEP Trial State Assessments) between 1990 and 1992 (USDE 1993)—a decline that may or may not prove wise, depending on the outcome of the proposed analysis.

HIERARCHICAL MODELS, MODELS MORE GENERALLY, AND THEORY: IMPLICATIONS FOR DESIGN OF NCES SURVEYS

Background

Survey samples sponsored by the NCES have often obtained data on institutions, such as schools, and simultaneously obtained data on individuals within the same schools, such as students. These include the National Longitudinal Study of the High School Class of 1972 (NLS−72), the National Education Longitudinal Study of 8th graders in 1988 (NELS:88), and High School and Beyond (HS&B), which focused on the high school class of 1980 and emulated parts of NLS−72.

The data on institutions have been combined in analysis with the data on individuals at times. Coleman, Hoffer, and Kilgore (1982) did do so, partly in the interest of discovering the relative effects of public versus private schools on student performance. Mosteller and Moynihan (1972) did so to understand the effectiveness of compensatory education programs. These illustrate early attempts to recognize the hierarchical nature of the data. More recent examples are not hard to identify, although Draper (1995) suggests that such analyses are the exception rather than the rule.

Despite the burst of recent attention to hierarchical data, technical advances in their analysis have been made for over 40 years (Draper et al. 1992). Work on the software that executes the analyses has been especially inventive and industrious over the last few years (Bryk et al. 1989). The fact that National Center for Educational Statistics has been collecting multilevel data for over 2 decades suggests that NCES anticipated, rather than lagged, advances in the software and analysis of such data, at least incrementally.

Draper (1995) argues that recent developments in hierarchical models (HM) have three clear advantages over earlier approaches to the statistical analysis of multilevel data:

1) "a natural environment within which to express and compare theories about structural relationships . . .

2) better calibrated uncertainty assessments in the presence of positive intracluster correlations . . .

3) an explicit framework in order to combine information across units . . . to produce accurate . . . predictions of observable outcomes."

Some readers are doubtless aware that Benefit #1 has been claimed for other analytic methods, such as LISREL. Boot strapping independent of HM arguably helps foster Virtue #2.

Hierarchical Scenarios, HM Models, and Analysis

The HM model we define here as a stochastic one that represents a setting in which units at the lowest level of measurement, for example, "A," are nested within units measured at a higher level, called "B," and these in turn may be nested in a still higher level of measurement unit called "C," and so on. Sampling and other random error is recognized at each level in the model. A variety of models and associated analyses can be regarded as special cases of a general hierarchical model.

So, for example, students (A) may be nested within classrooms (B) and classrooms may then be nested within schools (C). Variations among students, among classrooms, and among schools may be recognized in the random error terms and in other features of the model. Models that represent this scenario and the analyses are described in Bryk et al. (1989). An application to the data generated by the National Assessment of Educational Progress (NAEP) is given in Mullis, Jenkins, and Johnson (1994).

Or, time points of measurement (A) may be nested within students (B) who themselves are nested within classrooms. Some random coefficients models/analyses for longitudinal data fit this scenario. A related application to a sizable longitudinal study of participants in Boy's Town is given in Osgood and Smith (1995).

Or, one may conceive of a set of independent studies as a scenario in which individuals in (A) are nested within a given study (B) and various studies may be nested within (say) multiple geographic regions or institutions (level C). This scenario is similar to those encountered in attempts to combine evidence from different sources. Such a combination falls under the rubric of meta-analysis (Draper et al. 1992).

These scenarios and the associated analyses are considered in what follows. The emphasis here, as elsewhere in this report, is on what the advances in HM analyses imply for improving design of NCES surveys. The implications may concern: what units ought to be measured and how many, how, when and with what frequency, and at what level in a hierarchical setting (see Exhibit 1).

Hierarchical Models and Cross-Sectional Surveys

In principle, advances in hierarchical models (HM) invite one to analyze observations in contexts, e.g., students within classrooms, within school, within school districts, within states, and so on. An obvious abstract implication of the availability of the HM technology is that NCES might then collect data at these various levels. This data collection would be in the interest of exploiting a technology that purports to help understand, for example, how students' academic performance is influenced by classroom teachers, their schools, and the state education policies that influence them. At least, one might exploit the HM technology to understand where interesting statistical associations appear, even if one cannot be confident of where and how the influences are exercised, and even if one ignores time as a variable.

The implication just given is embarrassingly vague. It is also important. To get beyond the vagueness, we need to get to specific data sets, and to understand features of the models and the associated analyses and the data. Mullis, Jenkins, and Johnson (1994) did so. They tried out HM-based approaches to analyzing NAEP data on mathematics achievement. Their object was to identify "unusually effective schools" (outliers) and to determine how and why such schools differed from others. The bases for understanding were HM analyses that helped to arrange data at the student level within school and at the school level, so as to identify the predictable influences on student performance and school performance. Schools that departed from prediction in a positive direction could be regarded as unusually effective.

Here, the concern lies not with the substantive results of the Mullis et al. (1994) paper, which are interesting. Rather, the concern lies in what the authors say about better design of NCES surveys. The Mullis, Jenkins, and Johnson monograph, as one might expect from other sections of this report, contains no section on "implications." Drawing implications for better design of NAEP was not identified as an objective in the monograph.

Mullis, Jenkins, and Johnson did, however, construct a section entitled "Technical Issues in the Application of HLM to NAEP Data" (pp. 103–112). It is a springboard to implications. Their section taught us another small lesson: implicit in scholarly discussions of "issues" are possible implications. It invites us to encourage authors to write about issues rather than implications.

In short, what does the "technical issues" section of the conscientious HM analysis by Mullis, Jenkins, and Johnson imply for better design of NCES surveys?

*What Should Be Measured: Implication*s

First, NAEP measures of socioeconomic status (SES) are imperfect. Mullis et al. (1994) used what they could in a HM analysis based on NAEP. The imperfection in measuring SES are greater in NAEP than in other surveys. The implications are that

1) NAEP might measure SES more directly, e.g., asking questions about family size and income; or

2) NAEP might be linked to other information that gets at SES information more directly, e.g., SSA, IRS, and so on; or

3) NAEP might exploit imputation methods and/or indirect estimators to produce the SES information on individual students.

None of these options may be feasible for NCES. Still other options, not identified, may be more feasible. The raw implication is that analyses of NAEP data would be better if data on SES in the NAEP samples were better.

Who and How Many of Them: Implications

The Mullis, Jenkins, and Johnson (1994) report says plainly that the number of teachers within schools was not sufficient to sustain a HM analysis that could recognize the influence of classrooms and teachers (p. 104, first full paragraph). The implication here, as elsewhere, is conditional. If NCES and its clients want to learn about how teachers (classrooms) influence student behavior, having taken into account student-level variables such as family backgrounds and school-level variables, then NAEP should be designed so as to get at this level. That is, more teachers per school should be surveyed where multiple teachers per grade or class is the form.

The Mullis, Jenkins, and Johnson (1994) report also recognizes that the number of students within each school in NAEP may not be sufficient to estimate within school parameters (p. 104, last paragraph). Roughly speaking, they recognize, as others do, that relying on a random sample of 15–20 students within a school may not be sensible if the object is to understand average 8th grade students' performance within the school. But they also recognize that these data and estimates of average performance are aggregated up to regional and national levels that are arguably reliable because there are so many schools in the NAEP sample—1,500 schools in the aggregate.

We are aware of only one study of sample-size design based on HM that may be worth building on, by Magdalena Mok (1995). She chose a simulation scenario that is concrete, but it may not accord with scenarios in North America. Mok's simulation approach is at least promising, despite debatable relevance of the particular scenario.

This matter of numbers is controversial. The cautious implication is that NCES should support an investigation of sample size at all levels in the HM context. There appear to have been no comprehensive studies of statistical power/sample size issues or at least none sufficient to inform adjudicate decisions at the design stage of NAEP.

Longitudinal Data Analysis

Studies in education often explore how entities change over time. Analysis of such data has improved on account of analytical statistical advances in understanding growth curves, random coefficient models, event history, and so on. Analysis of longitudinal data on individuals is a special case.

Rogosa and Saner (1995) clarified approaches to analyzing such data and have compared different analytic methods. Breslow's paper (1989), unlike that of Rogosa and Saner (1995), stresses the benefit of empirical Bayes estimates over OLS in the context of longitudinal study in biometry. We depend here on the Rogosa work to lay out crude implications of the approach for the design and improvement of NCES surveys.

The presumption is that understanding individual growth is important, inasmuch as questions about growth precede and drive the exploitation of random coefficients (or other models) in analyzing the data. If NCES professionals, educational researchers, or other users of NCES data declare that questions about growth are unimportant, then the implications drawn here are unimportant.

The rudimentary individual growth model posits that the individual's state at time *t* is a simple linear function of time and random error. The individual's outcome state may be measured with error. It is common to characterize measures of this outcome using a classical measurement error model. Each individual in a group is characterized by the individual's base intercept and his or her growth parameter, i.e., a linear regression of outcomes on time. The model that describes this also recognizes random error. The group of individuals is then characterized by an overall mean and a mean growth parameter and some index of variability within the group over time.

This basic model is augmented, at times, by assuming that the individual's growth is a function of certain other variables. The individual's participation in a compensatory education program or the hours that the student spent studying are illustrations of such "control" variables.

Crude Implications

NCES should figure out *when* to measure each individual. Measure each individual's state at each of the time points, e.g., achievement, record each time point, and measure exogenous variables *z* that may influence growth parameters. These broad implications are obvious. Rogosa and Saner (1995) and others raised questions that bear on more interesting implications of analytic work on understanding growth.

Less Crude Implications: Sample Size

Empirical and simulation studies suggest that small samples lead to intolerably large standard errors in estimating growth parameters. Sample sizes above 200 seem acceptable to Rogosa and Saner, given the kinds of questions that they have explored. At the national and state level, NCES routinely depends on larger samples.

How big should the sample size be, under what conditions and particular growth models, and with what particular method of estimation? As yet, there seems to be no general answer to the question. This question can have no specific answer absent a specific question about what needs to be understood about a specific phenomenon. NCES may then choose to wait for others to address this question before going further. It may sponsor special studies to address the question so as to serve contemporary interest in growth curve analysis.

Less Crude Implications: How Often to Measure

Rogosa and Saner (1995) suggested that 4–6 time points for measurement is not uncommon. But we have seen no substantial analytic, empirical, or theoretical handling of the topic of how an agency such as NCES should decide.

If education would be served well by research on individual growth curves, then the study of when, how often, why, and how observations should be made is sensible. NCES might then commission studies that lay out the issues and support pilot work that addresses them. Or, NCES may wait for others to proceed further.

Hierarchical Models and Meta-Analysis

Draper (1995) considered briefly the link between hierarchical models and meta-analysis. He cited the hierarchical model's ability to detect between-study variation as the main reason why it is "a natural tool for implementing [a meta-analysis]" (p. 133). In the discussion, Draper described a six-study analysis (Goodman 1989) of the effect of aspirin on the survival rate of patients who had survived a heart attack. Although the results of the meta-analysis suggest that treatment is effective, there was substantial between-study variation. The researcher who initially implemented the meta-analysis, however, did not then "pose and [test] a series of linear models to explain the variation" (Bryk and Raudenbush 1992, p. 156).

From Draper's perspective, this constitutes a misuse of the hierarchical model. He contends, that "this can actually promote an antiscientific attitude of indifference to the cause of the study-level discrepancies" (p. 134). The implications of Draper's perspective "for allocation of research effort and resources" (p. 133) are to invest research time and money in discovering how and why the study level characteristics explain the between-study variation *before* recommending treatment. The implications of Draper's perspective on meta-analysis for the design of NCES surveys, however, are for the most part less clear.

Modeling and Analysis Generally

Clogg (1989) identified points of uncertainty in constructing models in the social and behavioral sciences and education. Each point engenders difficult choices in analysis. Each choice might be better informed through better survey design. Freedman (1985) assaulted conventional approaches to modeling in the social sciences, including those in education. Freedman's scientifically assaultive approach and Clogg's empirical approach have some of the same implications.

Universe

Clogg maintained that data analysts who depend on survey data produced by statistical agencies need better information about the universe that is sampled than they usually have. Because so much analysis is directed toward making generalizations about the nation based on national probability samples, he argues that the *census* must be improved. For instance, the census often is

used as a benchmark for checking the quality of other surveys, including NCES surveys. If the census universe is imperfectly specified, then the benchmark checks will be misleading. Similarly, if census figures are used to construct sampling frames but certain groups are undercounted, then the frames will produce results that differ from what they should be.

The implication is that insofar as NCES relies on census figures to design its surveys, improvements in the decennial census can help to improve NCES survey design and the analysis of NCES data.

Measurement

Clogg (1989) believed that good measurement is fundamental to good analysis and praised contemporary cognitive research on asking questions. Freedman (1995) argued that good measurement is not common enough in the social sciences. He told us that "good models are hard to build on bad data." But, aside from criticism of factor analysis, Freedman told us nothing new.

Cognitive approaches to understanding how people respond to questions can be regarded as a new approach to analysis and to designing better surveys. They may be employed at the survey design stage and, indeed, NCES does so. There is little published on the product of the effort however. The approach might also be productively employed at the stage of statistical analysis. For example, such research might reveal why nonresponse rate is relatively high for teachers' responses to questions about credentials in the Schools and Staffing Survey; these may then lead to redesign of the questions.

There are at least two implications with regard to cognitive research. First, publishing on the lessons learned from earlier NCES investments in predesign work on cognitive aspects of questions seems sensible. This might be done through NCES *Research and Development* reports or other means. The product is arguably of potential value for all scholars who seek to pattern local surveys after NCES efforts. The second implication is that the cognitive research might be undertaken after the survey is done in the interest of better understanding of the survey's results. NCES might encourage this at low cost to the agency through a variety of means—predoctoral and postdoctoral fellowship work, collaboration with university-based or institutional researchers, reliance on able and thoughtful graduate students, and so on.

Complex Sample Design

For Clogg (1989), "the failure to take account of uncertainty produced by complex sampling procedures is surely one of the most embarrassing problems we have at the moment. For at least some cases, reasonably tractable procedures are available, but the technology available now seems difficult to implement in the context of the formal models that we estimate routinely."

In some respects, NCES has already invested productively in addressing Clogg's concern. Scholars who seek to analyze data from the Schools and Staffing Survey (SASS), for example, are supported in effect by software (based on still other analyses) that characterize uncertainty in estimates of parameters *and* in formal statistical tests of a conventional variety.

The implication is that NCES ought to continue to build more user-friendly and accurate characterizations of uncertainty.

The Interface Between Theory and Models

For Clogg (1989, pp. 218–19), "the goal of analyzing social statistics is to explain how a system of variables works." The idea that theory is important is implicit in his remarks. Zellner (1989, p. 164), in discussing successful modeling of the sort that Clogg describes, says: "a good deal depends on whether good, relevant statistical theory and subject matter theory are available." That is, without good subject matter theory, modelers are forced to be content with description and exploratory work that may help to illuminate the structure underlying data and to make forecasts.

The immediate implication is that where subject matter theory is good, new analysis methods and models that generate the methods can be used to explore the theory. The products of this activity may have implications for better surveys.

The broader implication is that NCES should be aware of theories for which new analysis methods are useful. This awareness might be achieved, as it is at times, through advisory groups and consultants. It generally is achieved through contractors only when the contractors contribute to theory and to responses to an RFP (Coleman).

Subject matter theory in education in some areas is not sufficiently specific to determine which models ought to be used. This forces us to think in terms of description and forecasting. The less obvious and less certain implication is that designing better surveys rests heavily on deciding what to describe and how to describe it, rather than on new analytic methods and models.

The Roles of Models

Suppose we consider surveys of the kind that NCES runs as "nonexperimental social science." Suppose we then consider "new models" and the analyses they engender and ask: What the role of such models is in nonexperimental social science?" In fact, the question has been posed and addressed, in a special issue of the *Journal of Educational Statistics* (Shaffer 1992). The primary new models and methods reviewed in Shaffer (1992) include path analysis and structural models. David Freedman provided the main criticism. Rejoinders and reactions were developed by among others, David Rogosa, who is also not well disposed toward such models, and by Peter Bentler, Herman Wold, and others who have tried to develop such models.

Direct Implications

The volume contains no direct discussion of how structural models, as represented by LISREL or EQS, for instance, or of path models, should influence the design of observational studies.

Indirect and Very General Implications

The advocates of structural models argue that they are useful in developing parsimonious description. In effect, one is able to characterize a measure on any array of variables as unobserved measures on a far fewer number of latent variables. Then one could construe an implication as we should assure that the number of variables is sufficient to identify the latent trait well. If one believes that "home environment" is important as an "unobservable trait" of a child, there must be a sufficient number of questionnaire items to get at it.

Conversely, one may have many questionnaire items and reduce their number rationally through some approach related to structural models, e.g., factor analysis. How well one might do this depends heavily on substantive theory about how latent variables are related among themselves, and to the variables (questionnaire items) actually used. Freedman argues that good theory is absent. Further, LISREL and related approaches are not theory construction methods. He and Rogosa argue further that the scientific approaches are questionable at best.

More Direct Implications That Are Negative

Rogosa argued that if understanding individual growth or change is a main objective, then models/methods such as path analysis and structural models are inappropriate. His "message is that the between-wave covariance matrix provides little information about change or growth" (p. 89). More to the point for science, "covariance matrices arising from very different collections of growth curves can be indistinguishable" (p. 93).

The longitudinal studies of NCES are developed, in part, to understand individual change. Rogosa's position suggests that because structural models are inappropriate for analysis of data from observational studies, they are also an inappropriate resource for guiding the design of observational studies. Such methods have no implications for design because they are irrelevant to sensible analysis.

Rogosa argued further that more transparent and defensible approaches to growth curve analysis are at hand. In particular, the good scientist and statistical analyst can model each individual's trajectories, estimating parameters within each individual. One then models the differences among individuals.

Implication: Causal Structural and Path Models and Methods

Over the last decade, NCES has been advised not to undertake analyses that are causal in their orientation. The advice has been rendered by one Advisory Council on Education Statistics, notably by ACES Chairman Ellis Page during the 1980s. There was active opposition to such analyses by USDE Undersecretary Chester Finn to NCES' conducting such work in the late 1980s.

The models considered here and related ones are regarded as important in some quarters, e.g., among some education and sociological researchers. They are regarded as valueless, absent

stronger theory, by some educational and sociological researchers. The controversy is sufficient to justify educing that, exploiting the models and methods ought to be avoided, at least in official reports on the state of education.

A second implication bears on advisory committees' appointee to provide counsel on surveys. In particular, NCES may exclude from its survey advisory groups' scholars whose special interests lie in building structural models, path models, etc. on account of the controversiality of the latter. Or, such individuals might be included provided the advisory group is augmented by those who believe that such models are useless. Individuals with an interest in structural models arguably enhance the likelihood that variables regarded as important will be collected in an NCES survey. The inclusion of opponents will temper that influence.

The exclusion of structuralist fosters counsel based on a perspective that is more prescriptive, e.g., how children grow or change, or how districts grow and change and what are the covariates of growth. The models and methods then are arguably more transparent. The implications may be more obvious, e.g., focus on collecting data at more time points rather than fewer points and more subnational data.

The implication then is complicated. To the extent that NCES regards its role as the production of informative descriptive statistics (that are exploited occasionally for causal analyses), than relying on models, analyses, etc. that are not causal in their orientation is important. We can understand a lot about growth and change by seriously observing growth and change, not only through questionnaire/telephone surveys and administrative records, but through more direct observation.

To the extent that NCES regards its role as fostering the opportunity for structural, path, causal, analytical, *and* descriptive statistics, then relying on the more complex models is sensible. Society is complex and models must presumably be more complex.

There may, in fact, be no real conflict between these options in at least one sense. The data produced on the basis of an orientation toward good description (e.g., simple growth analysis) may "satisfice" for the structural models who use NCES data. Learning whether there is a satisfice and whether there are major differences in what each group regards as satisfactory suggests that NCES bring the groups together.

Implication: Vernacular and Causal Models

The Human Genome Project has had the benefit of great talent and considerable resources. Despite this, what a gene is, what "genome" means, and what the adjective "genetic" implies are still subject to some debate (*Science* 1994, 1995). Just as the scientists and statisticians struggle with vernacular differences and with remarkable efforts to understand what one means in the genetic arena, NCES and others must confront ambiguities in the model-building arena.

It does not seem unreasonable for NCES or scholars outside NCES to be confused about some statistical models and analytic methods. For example, "structural models" have been defined

at times as models whose parameters are invariant across some space, e.g., over time or geographic area. Structural models have *also* been defined, more or less, as models that represent the "as if by experiment." That is, they have come to be regarded as causal models.

Historically, "path models" have been characterized as "causal models." This characterization is important. These models, and structural models that are conceived of as causal models, are ways to develop a story. The story is one of plausible explanation of what influences what. Some scholars, such as Rogosa, have grouped the path models with structural models at times. Path and structural models have been lumped in with analytic methods whose underlying models were not originally explicit, e.g., cross-legged panel analysis.

At least a few scholars have tried to make distinctions plainer. Freedman distinguished between models that are helpful in summarizing data and those models that are born of more ambitious objectives, e.g., structural models. Roughly speaking, the idea is that observations on some "X"s are empirically related to observations on some "Y"s. Further and more important, Freedman's argument is that this is good description, but does not necessarily meet standards for a good structural model. The structural model is one that represents a good scientific theory.

The main point is that very able people, people who think, suffer the consequences of vernacular. "Structural models" for a fine economist may/can mean something different for a fine statistician or psychologists. "Causal models" means something to those of us who try to encourage controlled randomized experiments. It means the same, but it also means something different, to those of us who try to understand what variables (X) influence what variables (Y) in what is theory-based and arrangement based on observational data.

Exploiting Theory That Drives Survey Design and Model-Based Analyses

New analytic methods in statistics yield few *specific* implications for designing surveys. This is despite their usefulness for *interpreting* the data collected on the basis of an explicit design. Why is this? One can argue that we should not expect new analytic methods to imply anything about designing research. After all, the new methods have a certain objective, e.g., developing an estimate of a parameter that is better than its competitors. The first object is generally not to produce a better design.

Suppose we take a step back and imagine that advances in substantive theory (or policy), rather than the new analytic methods, are the drivers for improving both survey design and the analytic methods. Consider a simple example. Gender, in theory, is important. This theory then drives design of surveys that permit one to say something about gender differences. The data from such a survey permits one to analyze gender differences and to perhaps improve analysis.

This line of thinking may seem obviously true, at least, to a theoretician in the education arena. It may not be obvious to others. Even if the line of thinking seems sensible, how do we exploit this in the interest of better survey design? "Theories" are in ample supply in the education arena. Merely saying that we ought to rely more on theory to advance survey design is gratuitous.

It seems sensible to ask two questions: "Has a theory-based approach helped us to learn about how to improve design? If so, how do we exploit theory better, given the abundant supply?"

Consider one example by way of addressing the first question: How has theory-based analysis helped? Boe and his colleagues have depended on the NCES Schools and Staffing Survey (SASS) to understand the relation between teacher supply and demand, and the flow processes that underlie the relationship. Their analyses depend partly on being able to enumerate, from NCES data, teachers who are not credentialed to teach, i.e., anybody who is part of the actual supply of teachers. Their analyses cannot recognize a special source that is arguably important on theoretical and policy grounds. This source includes individuals who have been trained and employed in one professional arena but who then move to another. Engineers and scientists involved in the defense industry, for example, have moved to other occupations on account of the reduction in size of the industry. Some of these people find their way into the supply of teachers through federal programs that foster their transition. That is, thinking (theorizing) and finding out based on the thinking has an implication for SASS: the survey ought to obtain information about how certain people find their way into the teacher supply.

Suppose the reader is willing to grant that theory should have some influence on survey design and the models and analyses that exploit the resulting data in turn improve theory, design, and so forth. How might NCES be instrumental in tracking advances in theory so as to facilitate advances in design?

Implication #1

First, some framework for understanding advances can be invented. A simple one, based on the generic list given earlier for NCES' tracking advances in analytic methods (Exhibit 1), might simply list the things that new theory can address:

- What (new) variable ought to be measured?
- How and at what level ought the variable be measured?
- Who (or what entity) should be measured?
- When and with what frequency and periodicity should a variable be measured?
- What stratum (kinds of individuals, entities) should be observed?
- What statistical relationship need to be examined?

For example, some theorists have argued that we need to know more about family environment to understand the nature of family and school influences on the children's education. This implies that NCES can measure more related variables or measure them better in some NCES surveys that lend themselves to analysis of the topic. The NCES longitudinal surveys are an obvious option.

Considering the matter of statistical relationships, some analysts of NCES data have argued on scientific grounds that survey variables that are unrelated to others are candidates for abandonment. The argument is plausible.

In each of these cases, and others that can easily be constructed, theory plays a role. And a simple framework for understanding progress in thinking seems important for NCES and perhaps its sibling organizations. The mechanisms for tracking incremental advances on each front are fragmented.

Implication #2

Tracking advances in theory demands that NCES choose a target. Identifying self-declared theorists would result in a population of informants, of course. If these scholars depend directly or (more likely) indirectly on analyses of NCES data, then such an effort might be productive for NCES in the short or interim term. It then seems sensible to be able to locate and make use of individuals who actively capitalize on NCES data *and* individuals who depend on these data. The basic mechanisms available to NCES for doing so include those that NCES already depends on: using members of advisory groups for specific surveys, for projects undertaken by NCES contractors or by the NAS list of users of NCES data, and so on.

Implication #3

Assuming that relevant scholars can be identified and that frameworks for tracking the advancement in theory can be invented, then some method to facilitate the acquisition and sharing of information is still necessary. Conventional research journals, meetings of NCES data users and advisors are vehicles for doing so, that NCES already depends on. We might add to this the possibility that NCES can take better advantage of the World Wide Web. Such options are presented in the section on new technology.

COUNTING THE HARD TO COUNT AND MEASURING THE HARD TO MEASURE

Two topics are considered here. The first concerns eliciting information from respondents about a sensitive trait, state, or event. The second concerns the measurement of mathematics ability in NELS:88.

Background: Counting

At times, NCES has elicited information from students and others that can be regarded as sensitive. For example, NCES has asked students whether they have been victimized in an assault.

Asking a student member whether he or she provoked a fight or assaulted another student would be regarded as more sensitive. Asking about their having stolen property, engaged in unprotected sexual activity, and so forth may be regarded as extremely sensitive.

The Congress has attempted to limit the extent to which sensitive information can be required from students in surveys without the consent of their parents. Section 439(b) of the General Education Provisions Act (20 U.S.C. 1232g) for example, was amended in 1995 to say the following:

> No student shall be required . . . to submit to a survey, analysis, or evaluation that reveals information concerning . . . mental and psychological problems potentially embarrassing to the student or his family . . . sex, behavior and attitudes . . . illegal, anti-social, self incriminating behavior . . . income without prior consent of (adult or emancipated minor) or . . . of parent (minor).

That many surveys run by NCES are voluntary, rather than required of students, makes this statute a bit peculiar in its value. But recognize that the voluntariness of a survey may not be understood. And in any case, the mere posing of questions to a student about the student himself or herself may be offensive to some parents or teachers. Schools may decline to cooperate in surveys because the information being elicited from an individual about the individual's own behavior is regarded as sensitive.

Suppose that in many cases, NCES will not be able or willing to elicit sensitive information directly from students or others about their behavior. How then might one obtain information sufficient to estimate the incidence of a sensitive characteristic or behavior?

Network-Based Estimates

One approach to the problem of eliciting sensitive information has been developed by quantitative anthropologists and others with an interest in counting the hard to count: network-based estimators. Roughly speaking, individuals in a sample are asked *not* about their own behavior. Rather, they are asked about the behavior of unidentified acquaintances in their social network. For instance, we may ask a student: "How many students do you know provoked a fight in the last month?" This question is proffered instead of: "Did you provoke a fight during the last month?"

To estimate the total number of students who provoked fights, one also needs to elicit information from students about the size of students' social networks. The estimate of network size may be based on a survey question or a separate side study. Data on provoking fights elicited from students in the survey is combined with data on the average size of students' social network and on the size of the student population to produce an unbiased estimate of the total number of students who provoke fights.

Understanding how to estimate the average size of a personal network is no easy matter. It arguably depends on what kind of persons that one might encounter in a sample. For instance, a probability sample of adults might include priests and mail deliverers whose acquaintanceship network is larger than, e.g., a cloistered monk's. The efforts to understand personal network size in a variety of studies are reported in Bernard et al. (1987, 1989, 1990), Killworth et al. (1990), and others given in the reference list attached.

Contributions in this arena lie partly on the design side of surveys, e.g., learning how to elicit information. Part lies in analysis, including constructing estimates and understanding their quality.

Prior Analytic Work: Empirical Studies

The network-based approach has developed in research over the last decade or so. One of the more recent applications and a test of the approach is reported by Laumann, Gagnon, Michaels, Michael, and Coleman (1989). Their object was to estimate the prevalence of AIDS in the U.S. using the network approach. This was done partly in the interest of assessing the Centers for Disease Control's estimate of prevalence. The authors' vehicle for judging the quality of the network-based approach was a comparison of a network-based estimate of the distribution of homicide victims made against the distribution yielded by the FBI's Uniform Crime Reports and the PHS Vital Statistics. Network-based questions were embedded in the larger context of the 1988 General Social Survey, a survey that is independent of the FBI and PHS.

The results suggest that the network-based approach is trustworthy in producing homicide rates that are close to those yielded by official crime statistics. If one then chooses to trust the network-based estimates of AIDS prevalence, it appears then that the CDC data overestimate prevalence in some categories (e.g., in minority populations) and underestimate prevalence in others (e.g., in the Midwest).

The network-based approach has been the target for other interesting empirical research. For example, how to estimate average network size is crucial, and Bernard, Killsworth, and Johnsen (1994) have reviewed recent work.

A Broad Implication

The broad implication for NCES is this. If NCES wishes to estimate the number of people who have a sensitive characteristic in ways that avoid direct knowledge of the individual, then network-based estimators ought to be explored. In principle, the approach can be used in any NCES survey in which the size of the target population is known and a probability sample of the sample is drawn and the information about "knowing others who did X" and network size can be obtained. This then includes new waves of NELS:88, the National Household Education Survey, the birth cohort survey being considered, the Beginning Postsecondary Students Longitudinal Study, and at least some surveys mounted by the Fast Response Survey System.

Implication for the NCES National Household Education Survey

The National Household Education Survey presents some opportunities to exploit network-based estimators. NHES is based on a survey of a national probability sample of over 60,000 households. The sample and target population are well defined and over time (biennial roughly).

Assume that the target topic is sensitive, and that it would be difficult or impossible to get at the topic directly. Such topics might bear on the following:

- Indictment/conviction of school board members for wrongdoing;
- Indictment/conviction of teachers or staff who abuse children;
- Indictment/conviction of students;
- Parental abuse or neglect of students; and
- Who has been raped.

For instance, a survey of school board members that asks each individual whether he or she has been indicted or convicted of misuse of school funds, would arguably not be sensible. Obtaining information about the matter may nonetheless be desirable for *some* users of NCES data. Network-based estimators might be helpful to meet those users' demands. Moreover, they do it in a way that avoids privacy problems, embarrassment for the respondent, or intimidation. That is, asking each household respondent in a survey how many school board members they know have been convicted is likely to be more feasible than asking school board members whether they themselves have been convicted.

Asking the household respondent how many school board members or people they know also seems feasible. Similarly, understanding how many parents physically assault their children seems important. But the understanding cannot be gotten at directly. Instead, NCES might ask respondents in the Fast Response Survey or other vehicles whether the respondent *knows* about an assault, and about their network size (or independently, about network size in the target sample).

Background: Measuring the Hard to Measure

Kupermintz et al. (1995) analyzed the data from NCES' National Educational Longitudinal Study of 1988 (NELS:88) to understand how we might enhance the "validity and usefulness" of the NELS:88 measures of mathematics ability in the United States. Hamilton and colleagues (1995) examined the NELS:88 data on science also to understand how to enhance validity and reliability of testing on science.

The papers were generated as part of a seminar at Stanford. They are distinctive in that their objective was to enhance the quality of a periodic survey, notably NELS:88, based on conscientious analysis of data produced in earlier waves of the survey. Few published scholarly papers do so. There are still fewer that exploit "new analytic methods," regardless of how this phrase is defined, to do so.

Mathematics and Science in NELS:88

The work depended on a combination of small-scale cognitive research on the tests, conventional factor analysis, and new developments in full information-factor analysis. The latter involves employing a multidimensional item response model and a latent factor structure model that are, in conjunction, purported to yield estimates of item factor loadings on distinct abilities measured by the test that are better than estimates produced in other ways (Bock, Gibbons, and

Muraki 1988; Wilson, Wood, and Gibbons 1991). The implications of the latter are not obvious in the absence of its application to data such as mathematics tests in NELS:88. Even then, the relative contributions of specific analytic approaches used to yield the conclusions reached by the authors are unclear.

Broad Implications

What can NCES and we learn from this effort by our colleagues at Stanford? What are the implications of their work? The first lesson is that some university scholars indeed recognize that analyses can produce implications for better, not necessarily more, surveys. Further, and more important, they try to educe the implications. They are willing to embrace the challenge of doing so.

Second, the analyses are substantial, intriguing, and ecumenical. But the analyses occupy far more space in the published papers than do the papers' sections on conclusions. For readers who are interested in implications, this is not satisfying. The coverage is unbalanced, especially if the titles of the papers are taken seriously.

This perception of imbalance may be wrong, of course. An excellent implication based on ferociously difficult and time-consuming analyses, described in agonizing detail, may not take up much space. The idea that $E = MC^2$ is a conclusion of this sort. Despite considerable work, good conclusions that carry many implications can be astonishingly brief.

The analysts argued that the different dimensions of mathematical and science ability are influenced by different processes. Roughly speaking, the student's crystallized knowledge in mathematics is alleged to be influenced more by formal schooling; the fluid reasoning is influenced more by home factors. This argument is based on theory and on empirical regressions of mathematical factor scores on independent variables. The specific implication is that the theory and analyses ought to drive selection of variables and improvement of measures in NELS:88. For instance, one might focus more deeply on better measures of home education processes or characteristics that predict or explain fluid reasoning, keeping this initiative separate from attempts to develop measures of classroom processes that influence such reasoning. More generally, the implication that Kupermintz et al. (1995) draw is that large-scale assessments should certainly aim to represent the cognitive and educational distinctions being made by cognitive psychologists, math educators, and by the nation's education goals (p. 552).

That is, the theorists must be invited to contribute more to test development efforts and to the development of measures of potential influences on different abilities, especially higher order reasoning.

The Hamilton et al. (1995) results of analyzing science test scores from NELS:88 reveal different factor structures in 8th and 10th grades, and more factors than the math study revealed. The reasoning/knowledge distinction that appears clearly in analyzing the math scores does not appear in the science scores. Reasoning with knowledge appears as a factor distinct from science reasoning. The implication for the authors is that the multidimensional character of the tests ought

to be recognized in reporting and in comparisons, e.g., by state. Further, the authors imply that the way the science abilities are measured ought to be augmented by cognitive studies of the way students respond to items; they believe these can inform test design.

Finally, Hamilton et al. (1995) reiterate the idea that design of the NELS:88 survey should be "linked" with more direct investigations of the context in which instruction take place. This may involve asking more questions about instructional practice, e.g., emphasis on discovery learning or reciprocal tutoring, or more likely, out-of-school activity that has theoretic and empirical relation to factor scores.

Even Broader Implications

Consider first the possibility of improving the mathematics assessments in NELS:88. Kupermintz et al. (1995) suggest that the current test is "multidimensional and should be treated as such" (p. 550). The bottom line is that they provided good evidence to suggest that the mathematics assessments currently in use get at both factual knowledge and reasoning or crystallized and fluid knowledge. Further, they argue that the finding ought to be taken seriously in improving new NCES surveys. Their implication is a little unclear:

> In general, future survey testing efforts should rethink intradomain distinctions among such achievement dimensions and their links to theoretical formulations and empirical findings on the structure of cognitive abilities (p. 550).

That is, NCES is not measuring one "thing"—mathematics ability. NCES measures several things that are tied up in mathematics ability. The recognition can come about through improved design of the tests, reporting, or in other ways.

A second broad implication for NCES survey design and for university-based education hinges on the way Dr. Richard Snow and his colleagues appear to have approached their task. The Stanford seminar focused on a specific data set, NELS:88, and a reasonably specific implications topic, improving math and science measurement in NELS:88. Further, funding was available through a competitive peer review grant awarded by OERI to sustain the effort. This strategy is not common but has a good pedigree; recall that the Moynihan-Mosteller work on equality of educational opportunity depended on a Harvard seminar series that engaged very able people. Figuring out how to do this right in the interest of NCES, students and professors, and the public is not easy. But the example is sufficiently instructive to encourage taking the time to think about the strategy, its value for NCES and the public, and more interestingly, its value in advancing science at large.

INDIRECT ESTIMATES, INCLUDING SMALL AREA ESTIMATORS

Background

Agencies such as the National Center for Education Statistics (NCES) obtain databases on national probability samples and generate statistics pertaining to the national level based on those

data. For NCES and other statistical agencies, there has been episodic pressure to generate statistics at the subnational level, based on the national data. NCES also collects data at times at the subnational level, e.g., from the states. NCES has been encouraged at times to produce statistics at the substate level, once data users have found that the state-level data are instructive.

National samples, unless specially designed, do not usually yield results that are applicable to the state level. The national estimate of incidence of classroom disorder developed in NCES' Fast Response Survey of Public Schools on Safe, Disciplined, and Drug-Free Schools (1992) is not necessarily the incidence in New Jersey, for example. Similarly, the data on NAEP collected by NCES on students in Pennsylvania, for example, provides an estimate of mathematics ability for students in the state. The state-level estimate may not be an accurate characterization of abilities in local jurisdictions, such as Pittsburgh or Philadelphia.

The Question

Is it possible to exploit the data obtained at some aggregate-level data *and* other information so as to produce defensible estimates at the subaggregate level? Further, if it is possible, how can we understand the validity of these estimates?

Partial answers to the questions have been developed through recent work on Indirect Estimators. In what follows, we rely heavily on the Office of Management and Budget (OMB) (1993) *Statistical Policy Working Paper #21* and some other sources identified below.

The Approach: Indirect Estimators

Statistical agencies usually rely on *direct estimators* for reports. That is, the estimator, such as a mean number of assaults on students for the nation, is computed for a particular time and only from a sample of units in the population (or "domain") of primary interest, e.g., the students in the nation.

An *indirect estimator* is one that uses the design-based survey or a database for a direct estimator *and* auxiliary data from a sample or population (or "domain") or time period *other* than the one of initial primary interest. That is, auxiliary information is combined with information based on the data generated from the sample survey of the population and time that was the initial primary focus. The combination process usually depends on a statistical model that links the auxiliary information with information obtained on the population of initial primary interest.

For example, one might combine a national estimate of the incidence of student fights based on a NCES Fast Response Survey with state data on variables that may be related to fighting, such as urbanicity level, income, and so on. The combination would be based on a model of the purported relationship among the variables. The result is a *small area estimator* for the incidence of fighting at the state level. It is a special case of domain indirect estimators.

Or, one's interest may lie in updating a survey indirectly, using a *time indirect estimator*. For example, a survey run periodically, such as the Reading NAEP, might be combined with auxiliary data, based on a model of the relationship between the reading scores and auxiliary data, to estimate reading ability in a future time point between two points at which primary data are obtained and direct estimators are constructed. More specific definitions of domain indirect estimators (such as a small area estimator) and time indirect estimators are given in the introduction to OMB's *Statistical Policy Paper #21* (1993).

Precedent

NCES does not have a program to produce domain indirect estimates, time indirect estimates, or time/domain indirect estimates. Apparently, the National Center for Health Statistics (NCHS), Census, the Bureau of Labor Statistics (BLS), and the Department of Energy also have no special program as of this writing.

NCES, however, has been aggressive in building on and surpassing other statistical agencies as the need for a product appears and as resources change. The invention of licensing agreements, the use of CD-ROM for distributing data, and the use of videotapes, among other related NCES activity, illustrate the theme. Thus, it seems sensible to consider that NCES take advantage of recent developments in indirect estimation, so as to improve its surveys.

Empirical Examples Apart from Education

Breslow's paper (1989) in the *Sesquicentennial Proceedings* and the OMB (1993) report contain good illustrations of indirect estimators in various sectors, including health. Consider a small example: the NCHS national surveys of health are not directly generalizable to states. Nonetheless, there has been pressure on NCHS to produce state estimators; no resources have been provided for direct estimators. The NCHS has tried to develop reliable indirect estimators by doing the following.

NCHS obtains nationally reliable health statistics for certain subpopulations. e.g., gender, income level, and race. States per se are not in the subpopulations defined as important in the survey design although the demographic variables are.

The U.S. Census Bureau produces mid-decade estimators of the number of people within each state who belong to the subpopulations such as gender, income level, and race. The U.S. Census information is, in effect, auxiliary data that can be used to construct a NCHS indirect estimator of health characteristics for a state.

Combining the NCHS data with the auxiliary data from the Census Bureau's mid-decade data requires a model. The form of the model used to produce an indirect estimator can vary. The simplest form says that the state's health is a simple summation of the proportion of people in the state who are members of the subpopulation (i.e., female/male, high income/low income, and so on from the Census) times the mean health state of the respective subpopulation estimated from the

national-level data (from NCHS). This estimator, as described, is a basic synthetic estimator. Malec (1993) describes it and others that are more complicated. The latter try to exploit auxiliary information at higher or other levels of aggregation. Citations are given for PHS reports on indirect estimators for the states, on physician visits by the disabled, based on national and regional direct estimates.

The basic synthetic estimator is being tried out by Folsom and Liu (1994) to produce state-level substance used prevalence rates based on the National Household Survey of Drug Abuse (NHSDA). NHSDA is a national probability sample survey in which it is possible to link an individual's response to characteristics of the area in which the individual lives. These include census tract/block-level information within the county on, for example, median household income. They include county-level information on arrest rates from the Uniform Crime Reports. Within a state, then, individuals' dependence on illicit drugs, for example, is regarded as a function of the block/tract level within the county and county-level auxiliary statistics and with person-level data collected in the NHSDA. The model is based on Breslow and Clayton (1993).

The result is a predicted probability of dependency that takes into account the person, his or her block/tract characteristics, and county characteristics. A probability for each arrangement of characteristics (i.e., a person) is then computed. The prevalence rate is the sum of probabilities each being weighted by the number of individuals with those characteristics living in the block/tract, county, and state. This number is itself a forecast based on the 1990 Census updated to 1992.

Evaluating Indirect Estimators

There are several ways to evaluate indirect estimators, depending on the particular form and function of the estimation. None are perfect of course; some are less ambiguous than others.

The most straightforward of these involves 1) pretending that certain data are not available when in fact they exist, 2) building an indirect estimator, and then 3) comparing the indirect estimate to a direct estimate based on the actual data. For instance, the National Center for Health Statistics has tried to construct domain indirect small area estimates of state-level mortality rates for motor vehicle accidents. The validation is against actual rates computed directly from universe data at the state level. Similarly, indirect estimates of work disability have been compared to direct universe estimates from the 1970 Census (Malec 1993).

The basic idea is that the particular indirect estimator is judged against some known value of the direct estimator. If the results agree, this fosters confidence in the indirect estimator as a possible substitute for the direct estimator.

For instance, if NCES found that an indirect estimator can be shown to produce a good estimate for 1995, based on 1994 data, data collection for a subsequent cycle (e.g., 1996) might then be skipped. Resources could then be allocated elsewhere. Similarly, if a small area indirect estimator of, for example, state-level mathematics ability works well over a 3-year period, relative to the known value, one might then skip a 4th-year cycle of direct estimation and data collection at the state level, produce the indirect estimate, and reallocate resources.

Suppose that a straightforward direct estimator or known standard is not available. How then do we evaluate an indirect estimator? One may try to judge the latter's value in predicting some known estimator, which itself is predictable from direct estimators. One may also try to construct different indirect estimators and compare results. To judge from Malec's description (1993), it is not clear how to do this right if each of the different indirect estimators could be wrong in different ways. This warrants a bit more attention.

Malec (1993) also suggests that when the indirect estimator is model-based, then examining features of the model can help to inform a judgment about the quality of the indirect estimator. This is sensible. But developing coherent theory that leads to construction of a model whose elements also are testable is difficult. Effort in this direction nonetheless seems justified. Even small theories ought to be better developed and integrated with statistical models, in this arena and elsewhere. To the extent that different indirect estimators, based on different models and theories, invite deeper thinking about evaluating the indirect estimators, is to the good.

To summarize and extend Malec's treatment (1993), evaluating indirect estimators can involve the following:

- Comparing indirect estimators to direct estimators;
- Comparing the ability of indirect estimators to predict related direct estimators;
- Comparing different indirect estimators; and
- Examining the models and theories that underlie the models for the indirect estimators.

NCES might exploit the first two approaches. Examples are given elsewhere in this section. With a few exceptions, the last two approaches have not been examined deeply. These are where NCES might make a distinctive methodological contribution.

General Implication: Time Indirect Estimators

Some NCES data collection efforts occur annually. Examples include the yearly acquisition of school district fiscal data and the NCES investment in the Current Population Survey Enrollment Supplement. Often, NCES surveys are uniformly periodic, as is the Schools and Staffing Survey, in the sense of occurring every 3 years. Or, they are nonuniformly periodic in that the time intervals between surveys may vary, the National Assessment of Educational Progress being an example. The calendars for NCES data collection provided in the NCES publication *Programs and Plans* are particularly instructive.

Recall that a time indirect estimator exploits information from one time point to describe what we know about another. Estimates of reading ability in one year, for instance, might then be inferred from estimates of reading ability at another time. A time and domain indirect estimate of reading ability may rely on reading ability data from earlier times and on other auxiliary data.

In the abstract, time indirect estimation methods seem relevant to the NCES survey effort in several respects. One may imagine, for example, that the annual data on school district fiscal matters are not available. Rather, only data for every 2 years are available. Time indirect estimates for the imagined absent years can be developed. If the time indirect estimates accord well with the actual data, then one might consider eliminating the intervening year's data collection. That is, one reduces the burden on NCES and on the respondents. The reduction provides NCES with more opportunity to do other work.

The ability to validate the time indirect estimate is possible only because NCES now collects relevant data annually. If NCES adopts time indirect estimates so as to eliminate some data collection efforts at a given time, it would still be necessary to validate the indirect estimates periodically. Even if such estimators are deemed inappropriate now, knowing about their validity seems important for the future.

The NCES Education Assessment calendar shows that the surveys on reading have occurred every 2 years. That is, we understand reading ability in the United States from 4 assessments over an 8-year period. Imagine that in the future, NCES might conserve resources by doing the reading assessment every 4 years, instead of every 2 years. Would it be possible to produce estimates for the intermediate 3 years in which data were *not* collected? One way of thinking about this is to explore time indirect estimators. That is, having reliable direct estimates from NCES for 1990, 1992, and 1994, we imagine that the reading ability data are absent for 1990 and then exploit some time indirect estimation method to produce an estimator of reading ability for 1992. The time indirect estimator for 1992 might be based entirely on 1990 data alone or on auxiliary data in combination with the 1990 data.

More generally, of course, one might explore how time indirect estimators might be exploited in the interest of estimating and justifying an estimator 4 years or more out using current data. That is, if 2-year out data can be predicted well and the evidence for the quality of prediction is good, then one can eliminate burden to respondents and NCES.

The raw implication of all this is that the periodicity of some surveys can be altered because the results of such surveys can be characterized well. This may be possible because NCES has invested resources in data collection that permit one to understand whether the results can be characterized well.

The time indirect estimation methods might be exploited in any NCES program that obtains data annually. The annual efforts up to 1995 in the elementary and secondary arena include the following:

- Public School Universe;
- Local Education Agency Universe;
- State Aggregate Non-Fiscal Report;
- State Aggregate Fiscal Report;
- School District Fiscal Data; and
- CPS School Enrollment Supplement.

Where a universe sample or census is used to build a population frame, exploiting indirect estimates to eliminate the universe sample will be dysfunctional at worst. The functions of each of the annual efforts then need to be taken into account.

Time indirect estimation might also be exploited in reducing burden and expanding opportunities in the postsecondary education arena. The relevant annual NCES efforts include eliciting data on

- Institutional Characteristics;
- Fall Enrollments;
- Completions;
- Finances; and
- Doctorates.

Similarly, the NCES calendar for the Library Statistics Program involves annual data collection efforts that might be reduced to surveys every 2 years, e.g., on public libraries. The reporting burden and the burden to NCES might be reduced by exploiting time indirect estimators. Moreover, how well the indirect estimators perform can be evaluated because NCES has obtained data on public libraries annually since 1988, and has obtained data on academic libraries every 2 years since 1988. Again, if the survey is used to build a population frame for subsequent surveys, this would have to be taken into account.

The time indirect estimators might be also exploited in surveys and assessments that are undertaken every 2 (or 3) years, in the interest of reducing burden on NCES or respondents or both. The NCES data collection efforts that have occurred routinely every 2 or 3 years, and that then present an opportunity for evaluating the indirect estimators include, at the elementary and secondary level:

- Schools (SASS, 3-year cycle);
- School Administrators (SASS, 3-year cycle); and
- Teachers (SASS, 3-year cycle), among others.

Implication: Cross-Agency Efforts

The *OMB Statistical Report* #21 represents a small but noteworthy effort to pool expertise from different federal statistical agencies. NCES was represented in the work, as was BLS and others.

An implication of this NCES effort is that the cross-agency efforts can be important and productive. Beyond this, there is another implication. Education-related data are important in some efforts to produce dependable health statistics. County-level education data, for example, have been

exploited to produce more precise indirect estimators for state rates of physician visits. This use of education data reiterates the idea that, theoretically and empirically, education variables are important in characterizing health phenomena at national and state levels. NCES produces education data. NCES cooperated and assists in this production by other agencies, such as the Census Bureau. These roles are important to pursue in the interest of better design of surveys, i.e., surveys whose results can be used by other agencies.

Implications for Specific NCES Surveys

In the following, the purpose is to educe the implications of developments in indirect estimation, including small area estimation for specific NCES projects. However, the discussion is not always well informed. The NCES *Report on Programs and Plans* was a sturdy source of information. This section covers the following:

- National Household Education Survey;
- Schools and Staffing Survey;
- School District Mapping;
- Library Statistics Program; and
- Fast Response Survey.

National Household Education Survey

The National Household Education Survey (NHES) is undertaken (roughly) every 2 years on a probability sample of households with children. Special topical supplements or components have directed attention to adult's participation in adult literacy programs, school safety and discipline, and school readiness.

Time indirect estimators might be exploited to produce statistics on the years between the biennial surveys. It is not clear that there is immediate need for such estimates. Still, learning how to construct them and evaluate them can serve NCES well if 1) a biennial survey must be skipped and the indirect estimators suffice; 2) learning about the performance of these time indirect estimators is important for judging the quality of time in direct estimators in other contexts; and 3) enhancing opportunities for NCES to decrease coverage of some questions at little cost and increase coverage of others.

It is not clear whether and how small area estimators might be exploited, based on the NHES, auxiliary data, and models. The NCES *Programs and Plans* description, for instance, does not tell whether the statistics produced on the basis of the NHES are subnational. For reporting on some topics at the local or regional level, indirect estimators may be desirable. For instance, school safety and discipline matters are arguably most interesting at the city level. If resources are insufficient to provide direct estimates of, say, household encounters with school theft at the local level, then indirect estimators might be used.

Domain indirect estimates at the city level in the school safety and discipline arena might then be based on NHES, and on auxiliary data from the Common Core of Data, the NCES supplement to the Current Population Survey, or other information. It is not clear what models can be exploited to make the linkage between national (or state or regional) data and city-level data. Nor is it clear that linkage is possible.

How to validate the small area estimators produced for school safety and discipline is not clear. But there are interesting options. They include the NCES Fast Response Survey System (FRSS), the FBI's Uniform Crime Reports (UCR), the Bureau of Justice Statistics' National Victimization Surveys, among others.

Schools and Staffing Survey

The Schools and Staffing Survey (SASS) involves a periodic national probability sample of schools, interviews being directed toward school-based respondents, subsamples of teachers, and at times, subsamples of special interest. The latter included student records offices in 1993–94. The teacher survey has included a 1-year longitudinal followup of a subsample.

The NCES Common Core of Data provides the public school universe from which part of the SASS sample is drawn. The NCES 1989–90 Private School Universe File served as the population frame basis for the remainder of the sample.

The publication *SASS by State* provides statistics on public schools that are accurate at the aggregate level for each state and for the nation. The statistics are direct estimators, based on a sample design that permits their production. It seems reasonable to suppose that state-level *indirect* estimates might be produced using national-level data based on a smaller sample than is usually drawn, information available from the Common Core of Data's universe of public schools, and models to link the two. An alternative feature to exploring this strategy is that the indirect estimators can be validated because NCES does produce direct estimators against which the indirect estimators can be compared. Beyond this, the exploration may help us to learn why an indirect estimator cannot be produced or is inadequate in this instance and perhaps in others, and if it is adequate, NCES' flexibility in future SASS surveys.

Similarly, given that state-level data are available and assuming that some states have an interest in district or city-level estimates, NCES might explore the use of these estimators for this lower level of aggregation. Although evaluation of such estimates would be difficult, at least some validation is possible if in some states the district-level samples are sufficient for computing direct estimates.

The Common Core of Data (CCD) and perhaps other information supplied by the Decennial Census, for example, may be used in any time indirect or domain indirect estimator. For instance, the CCD might be exploited in attempts to build indirect estimates of supply and demand for teachers, teacher characteristics and opinions, and so on at the substate level. Insofar as municipalities, especially large cities, are concerned about the topics, the indirect indicators may be sufficient to meet their needs.

To judge from the 1995 NCES publication *Programs and Plans*, private school data generated by SASS is direct estimator based at the national and regional level, but not at the state level. The 1989–90 NCES Private Universe File may be capitalized as auxiliary variables in constructing state-level statistics for private schools or for regional statistics beyond the four censuses now used. The 1990 census data, or less likely NCES' supplements to the Current Population Survey, might be similarly capitalized. Insofar as voucher system or relative financial supports for education are state-based and that private schools have an interest in vouchers, the indirect estimators may be of value to the users of NCES data.

SASS has been planned for 5-year intervals. We then have direct estimates every 5 years on topics that are regarded as important by the users of NCES data. Suppose that estimates of statistics on the interviewing years would occasionally be valuable to states, municipalities, or the Congress or to other admiring users of NCES data.

Estimates of statistics on each year in the 5-year interval between SASS might be based on time indirect or domain indirect estimators. Suppose, for instance, that a new administration's interest lies in the best estimate possible of how many teachers will quit their jobs and form a new business, or how many teachers here engage in a new business while they maintain the teaching job. It is conceivable that indirect estimators can be constructed. Again, their defensibility is not clear.

Finally, consider the question: "Is it possible to generate indirect estimators for the state level?" Because direct estimators are available, the indirect estimates can be evaluated against them. The question may be worth addressing in a program of research on indirect estimators partly to better understand their general performance. It may be worth addressing so as to provide more flexibility in the design of SASS. That is, indirect estimators may be adequate for some parameters; and direct estimators may be essential for others. Learning which is which seems important.

School District Mapping

Regarding School District Mapping, the NCES (1995) publication *Programs and Plans* tells us that the U.S. Census identified blocks and has mapped them onto local and state education jurisdictions in 1970, 1980, and 1990. This mapping project is, in principle, important for indirect estimation at the local level. It provides a deterministic geographic link among jurisdictions—national, state, substate. The information however, may not be sufficient for small area estimation. *Programs and Plans* tells us little about what statistics at any geographic level are produced on the basis of the mapping project, despite the import in principle, of the project. It does tell us that "NCES will provide 200 tabulations of state and district totals to each of 16,000 education agencies and each state."

Library Statistics Program

The NCES Library Statistics Program is directed at public libraries, academic libraries, and school-based libraries and media centers. Public Library statistics are based on *annual* reports on a universe of nearly 9,000 libraries since 1988. Reports have been channeled through State Coordinators. The reporting is automated through the DECPLUS system.

NCES obtains Academic Library data on the universe of accredited institutions and unaccredited 4-year institutions every 2 years. The beginning year was 1988. The reporting, exploiting the software system IDEALS and state level coordinator, is done every 2 years in coordination with IPEDS. The school-based library statistics are gathered primarily through the national *probability sample-based* Schools and Staffing Survey. Data are available from 1991 and 1994; plans for the future are not yet accessible.

Consider first the annual public library statistics program. The implication of analytic developments in time indirect estimators is that the library statistics need not be obtained annually. That is, the library data might be obtained every 2 years, rather than each year. This reduces the reporting burden on respondents and the burden of processing reports for NCES staff. If time indirect estimators are adequate, the reduction in burden is not complete. A small burden would be shifted to those responsible for producing the time indirect estimators and for producing evidence that the estimators are credible relative to some standard.

Whether it is valuable to anyone to reduce the burden of reporting for public library statistics is debatable. NCES and the states have done a remarkable job to generate a low cost reporting system.

Despite this success, exploring the use of time indirect estimators may be warranted on accounts. First, the time dedicated to reporting each year for the annual public library data may be dedicated differently every other year without real damage to annual data that are based on direct and indirect methods every 2 years.

The alternative years might be dedicated to special surveys; for instance, we might like to understand the frequency of users of the public libraries, based on library card use. Now, we might like to understand periodically which books, in a list of 100, are most exploited based on check-out lists. The point is that the time dedicated to an annual report might be adjusted. The routine report may be made every 2 years. A report on special interest may be made for each intervening year. The statistics on intervening years may be generated through the year in direct estimators.

Fast Response Survey

The Fast Response Survey System (FRSS) is usually based on a national probability sample. The members of the sample are asked for information by mail or telephone.

The sample design, in some cases, is such that regional-level estimators (i.e., subnational) are possible. Further, the surveys, at times, obtain substantial demographic information on individuals or the institutions on whose behalf individuals report.

The FRSS data collected at the national level can, in theory, be combined with other NCES data obtained at state or regional level in the interest of producing local area statistics that are of interest. For example, FRSS does not produce state or substate statistics on criminal and noncriminal disorder in schools. It produces national and regional statistics. Statistics on cities are often deemed important by Congress and by the states. Statistics at the state level are often deemed

important. The subtechnology of small area estimators might be brought to bear to provide those statistics. The objective seems feasible for state rather than substate levels, but both seem worth exploring.

ADJOINING RANDOMIZED EXPERIMENTS TO OBSERVATIONAL SURVEYS: SATELLITE POLICY

Introduction

There are a variety of ways to enhance the usefulness of surveys. In this essay, one such strategy is considered. The idea is to attach controlled randomized field experiments periodically to ongoing NCES surveys.

Research policy that encourages coupling the two approaches, experiments and surveys, will make both survey data and experimental data more useful for social science and public policy, and decrease the artificial separation of the sample survey and experimentation traditions. The expectation is that linkage will occasionally reduce unnecessary debates over policy-relevant data analyses. In short, a policy that invites coupling of surveys and experiments would combine the strengths of each approach while compensating for their respective analytical and administrative weaknesses.

The following section provides excerpts of work by Blumstein et al. (1986), Farrington (1988), Fienberg and Tanur (1986), Boruch and Pearson (1988), Boruch (1975), and others. It also presents some new ideas.

Definitions

Longitudinal surveys are defined here as repeated observations of the same persons or organizations or other entities in the interest of documenting growth and change. A major purpose of such studies is to understand how individuals (or organizations, and so on) change over time. Interest may, for example, lie in the growth of children's intellective achievement and how that growth accelerates rapidly during some periods (e.g., early childhood) and accelerates less rapidly in other periods. Or, the interest may lie in variations in level of delinquent activity over some period. When based on well-designed national probability samples, such surveys are the best possible approach to statistical characterization of individuals' growth, development, and engagement in various educational and social systems. Compendia of national longitudinal surveys are given in Taeuber and Rockwell (1982) and in Verdonik and Sherrod (1984).

Randomized experiments are defined as settings in which individuals (or organizations, or other units of study) are randomly assigned to one of two alternative regimens. The object of the experiment is to estimate the relative differences among regimens in a way that is unbiased, and that permits formal probabilistic statements to be made about one's confidence about the estimates. Interest in long-term differences between what are frequently referred to as "treatment" and "control" groups are often of interest and may engender the repeated observations that characterize longitudinal or panel research designs. Collections of field experiments are listed by Boruch et al. (1978), Riecken et al. (1974), and others.

The statistical models used to analyze each kind of data usually differ. Heckman and Singer's edited monograph (1985), for instance, reviews methods of analysis but not the design of such studies. But one can develop analyses that simultaneously exploit contemporary experimental design models and models designed for common panel or longitudinal data (e.g., Boruch 1975; Fienberg and Tanur 1986, 1987b).

A Proposal for Satellite Policy

The proposal for joining experimental studies to ongoing NCES surveys may be stated as follows (amended from Boruch and Pearson 1988).

Any NCES survey study should be designed so that independently designed experimental studies can be adjoined to the survey so long as 1) the experiment is compatible with the mission of the NCES longitudinal survey; 2) the risks of disruption to the NCES survey can be managed, especially in regard to the time frame, respondent's burden, and institutional cooperation; 3) designated contractors are responsible for oversight of the process; and 4) the experiment involves no appreciable cost to NCES.

This proposal is analogous to the policies on satellite use that have been used in astrophysics. The satellite, like a longitudinal survey, has a primary monitoring mission and requires considerable resources to place and maintain. Further, scientists can obtain access to part of the satellite periodically for limited, temporary investigation of important scientific questions (i.e., experimentations).

The strategy proposed here allows the research to depend on the infrastructure of the ongoing survey as a vehicle for conducting prospective experimental studies. The proposal also extends a scientific tradition of "data sharing" in the social and behavioral sciences and education research (Fienberg et al. 1985). In particular, it requires that resources be shared: population listings and sampling frames, the organizational vehicles for longitudinal surveys, and so on, not just data.

Adjoining experiments to ongoing longitudinal surveys is likely to be feasible. However, this may occur for only for a few projects, perhaps only one every year or two, because of the difficulty of coupling a special study to an already complex survey.

Justifications for a Satellite Policy

There are several kinds of justification for adding controlled tests to such a study design, as described next.

Scientific and Statistical Rationale

The mathematical conditions under which longitudinal (nonrandomized) study will fail to yield an unbiased estimate of relative program effects are well understood. Rubin (1987) provides a basic description in the context; Campbell's and Boruch's treatment (1975) is more rudimentary.

Heckman and Robb (1985) provide elaborate description for analysis of both longitudinal and cross-sectional data in an economic context.

Despite advances in the mathematical aspects of the topic, the problem of assuring that mathematical assumptions are tenable remains. Even determining whether assumptions are met can be difficult or often impossible, especially where theory is not adequate. All approaches to estimating the effects of intervention based on longitudinal nonrandomized data depend heavily on the assumption that performance of individuals in the absence of the intervention can be estimated accurately.

The assumption is patently suspect to judge from empirical comparisons of evaluations based on longitudinal against evaluations based on randomized evaluations. LaLonde (1986), Fraker and Maynard (1987), and Maynard (1987) show how estimates of program effect based on the former have been demonstrably wide of the mark in evaluation of manpower programs.

The economist's work is recent. Early research on nonrandomized clinical tests in medicine and on randomized clinical trials showed differences in results between the two. Boruch and Riecken (1975) gave relevant illustrations.

Work by Gray-Donald and Kramer (1988), for instance, reiterates the point for research in pediatrics. Observational studies have typically shown a definite association between infant formula supplementation in hospital settings and lower subsequent breast-feeding by mothers. The inference has been that supplementation then has an important potentially negative effect. Controlled randomized tests show no such difference, reducing pediatricians' concerns about supplemental feedings in hospitals.

The point of this and other illustrations is though longitudinal studies may be useful for description of growth and change, they cannot be relied on for accurate estimates of the effects of new intervention programs, at least not in the absence of strong theory.

The implications for Chapter I evaluations based solely on longitudinal study are direct and have identified been by Smith (1988). The law's demand that Chapter I effects be estimated using only longitudinal study cannot be met without heroic assumptions about children's behavior in the absence of such programs. Such assumptions may be tolerable politically. But they are often indefensible scientifically. The implications for longitudinal study of the Program on Human Development and Criminal Behavior are related if indeed the program seeks to determine how onset of delinquent behavior and resistance can be affected by intervention. They are reiterated by Farrington (1988) and Farrington, Ohlin, and Wilson (1986) among others.

A second justification for adjoining experiments to longitudinal study is that the science and technology of randomized field tests of projects has developed more or less independent of the technology of longitudinal surveys. The intellective separation is often sufficient to prevent researchers from thinking about both in designing tests of new programs or in designing longitudinal studies of important topics. There are good scientific reasons to avoid intellective parochialism here and to understand the union of approaches when the opportunity arises.

A third scientific justification stems from the observation of Fienberg et al. (1985) that although major experiments involve collecting longitudinal data, their analysis is often based on dynamic models that were not incorporated into the design of the experiment. The failure to involve these models in design of the survey, they suggest, ultimately leads to less defensible analyses of experimental results. The argument seems sensible. But little formal research on the relative gains and costs of basing designs on analytical models appears to have been undertaken.

The scientific justification for coupling experiments and longitudinal surveys is then to capitalize on the strongest merits of each. That is, one obtains both the information produced by national probability samples—often conducted over a considerable length of time—and the information produced by smaller comparative experiments in which causal inferences are more appropriately deduced. Insofar as the experiments can be adjoined systematically to surveys, their generalizability will be enhanced.

Economic Rationale: Less Costly Policy Experiments

It takes considerable effort to mount high-quality surveys. It also takes considerable effort to mount randomized tests of policy relevant programs, more effort if we recognize the difficulty of maintaining control over selection of individuals into programs and over program operations. To the extent that an experiment can capitalize on the resources and data of a survey, the experiment becomes a less costly enterprise.

Experiments undertaken by the Broward County School Board's Department of Research (1987) are a case in point. Their experimental tests of the AIM project for youth at high academic risk capitalizes heavily on a regular system of standardized testing using Iowa Achievement Tests (i.e., a periodic survey) and the infrastructure to which regular testing was based to execute the experiments. The infrastructure was especially useful in tracking the large number of children who migrated from the original 6 schools to 18 schools (Carey Sutton, Personal Communication, November 11, 1988). In a longitudinal study, for instance, we might reasonably expect the adjoined experiment to exploit one or more of the following elements of the basic survey-based study:

- Interviewer cadre, the investments in their training, supervision, and quality control;
- Questionnaire and interview design;
- Information generated in the longitudinal study about local institutional, political, and managerial constraints and stakeholders; and
- Knowledge emanating from the survey about the structure and quality of administrative records, e.g., police records, education records.

Two kinds of local statistical data generated in surveys are often crucial to a well-executed experiment: estimates of the number of individuals relevant to a particular experimental project and estimates of the temporal flow of such individuals through various systems. So, for instance, a longitudinal study that included attention to youthful co-offenders might generate good information

on their number, their geographic stability and their general geographic location or locatability. Such pipeline studies could arguably help to avoid the problem of some experimental tests in police handling of domestic violence and others (Project Review Team 1988). Such information is basic to a pipeline study that would inform the design of a experiment dedicated to preventing illegal activity by co-offenders.

It would, of course, be a mistake to depend on a survey system to inform all aspects of the design of experiments. It usually cannot help much, if at all, in understanding the ethical or legal propriety of experimental tests, for instance. Nor would a survey help to understand the obstacles to implementing a new regimen in the experiment.

The implication is that field experiments can exploit surveys done in the areas in which the experiment will be emplaced, to decrease the cost of experiments. The reduction in cost stems from capitalization on human and statistical resources and savings in time.

Prophylactic Rationale

Cross-sectional and longitudinal surveys are often used to produce evidence that they often cannot support as, for example, in addressing questions in the social sciences and public policy about the impact of social programs. The Continuous Longitudinal Manpower Survey, for instance, has been and is supported primarily on grounds that it is important for understanding the changing nature of the pool of human resources available to society.

A second justification for the Continuous Longitudinal Manpower Survey is that it is useful to understand the effect of special programs in youth employment and job training. The second justification may be useful for rhetorical purposes (e.g., to gain political and fiscal support for the survey). But it is not always appropriate and is counterproductive insofar as the claim is exaggerated. That is, longitudinal surveys alone are usually not sufficient to estimate the effects of programs designed, say, to affect the earnings of individuals, some of whom happen to participate in the survey. Nor are these designs sufficient for making causal statements about the effects of programs in the health, criminal justice, and other areas. See the earlier remarks on scientific justifications and the reference to the Fraker and Maynard (1985; 1987) and LaLonde (1987) comparisons of program effects based on randomized experiments against effects based on data, notably the CLMS and the Current Population Survey (CPS).

In the case of evaluating Chapter I programs or others in elementary and secondary education, relying on a longitudinal study will merely continue a practice that is known to be risky. The estimates of program effect, if one follows the instruction of law, will be ambiguous at best and misleading at worst. To the extent that randomized experiments are a prophylactic to such results, and have been recognized as such in medicine and education since the early 1970s (Campbell and Boruch 1975), then such experiments ought to be considered seriously.

The Program in Human Development and Criminal Behavior has grappled with this issue (Farrington, Ohlin, and Wilson 1986) and continues to do so.

Calibration Rationale

An engineering justification for joining experiments to ongoing longitudinal surveys is that one may use the experiments to calibrate estimates of program effects that are derived entirely from the longitudinal survey (Boruch 1976). That is, the biases in estimates of program intervention that are based on longitudinal data can be assessed, and periodically corrected, through controlled experiments. Longitudinal studies are then likely to be more policy-relevant and less ambiguous with respect to biases in estimating program effects. Experiments are likely to benefit from their greater generalizability, lower costs, and more manageable administration.

As a practical matter, systematic calibration is a couple of decades in the future. Nonetheless, one can develop rude comparisons of results from both kinds of study. In the work on comparing estimates in supported-work manpower training programs, for instance, the biases engendered by relying on a longitudinal survey differ depending on whether one considers youth or recipients of Aid to Families with Dependent Children. For instance, the estimates for the impact on youth in 1979 was near zero for the experiments and minus $1,200 for the nonrandomized study. Estimates for AFDC women do not differ appreciably.

It is especially appealing to consider calibration in the case of Chapter I programs because the better parts of the Chapter I Reporting and Information System and infrastructure might be exploited. (See Reisner et al. [1982] for work up to 1981.) The comparison of estimates of program effect based on grade equivalents against estimates based on randomized tests may reveal that the former does well under certain conditions, e.g., for 2nd graders. The accumulation of experience about when each type of estimate is in accord can help us to understand when experiments are not needed.

Methodological Rationale: Better Methods and Data

Some of the methodological reasons for joining experiments to longitudinal studies are implicit in the earlier remarks. The economic rationale, for instance, carries the implication that experiments can be better designed so they cost less. The statistical and calibration justifications also accord with methodological interests.

The methodological rationale for joining experiments to longitudinal study can be narrowly construed, and often is, to understanding how to reduce measurement error in tests and interviews. Understanding how to elicit accurate information from people in the face of poor memory, difficulty in understanding questions, and reluctance to provide responses seems important. The problem has, at times, prompted the design of experiments in the general context of longitudinal studies.

Malvin and Moskowitz (1983), for example, undertook randomized experiments to understand how to better elicit information from junior high school students on their drug use and attitudes. The work involved comparing completely anonymous responses to ones in which identification was elicited but privacy assured by the substitute teachers responsible for administration of questionnaires. The biases reported in identified questionnaires appear to the authors to be very small except for current use of drugs.

The Weis (1987) review of research on reliability of reports on delinquent and criminal behavior suggests that new methods of eliciting information do often not work better than high quality conventional ones. The paper is persuasive on this account. Still, need to improve quality invites attention to better controlled tests. Some of the tests can be adjoined to longitudinal study.

Mathiowetz (1987), for instance, mounted studies to understand how to better ask questions about the unemployment spells of employees of a large company partly to improve quality of data in the Panel Study of Income Dynamics (Mathiowetz and Duncan 1984). Her object was to ask questions in two different ways to determine which yielded more valid results: validity standard and available company employee records. Although in this case the same sample was asked both kinds of questions, an experiment could have been designed to achieve related ends.

Policy and Political Rationale

A longitudinal study's usefulness to policy lies partly in its capacity to show change. A national shift in school truancy level may, for instance, direct attention to the problem.

Consider then that the scholarly and policy use of longitudinal data is high soon after a first wave of measurement. The use tapers off rapidly until the next wave. Consider further, several waves of measurement may be characterized by little change in the phenomenon of interest.

The implication is that "surprises" in the sense of new understanding over time will be infrequent and will decay. If they occur at all, they will be tied to frequency of measurement and frequent change. To the extent that this is true, one might choose to measure frequently. This may make possible results that show, for instance, that only 10 percent of the individuals involved in high crime commission rates in one year are involved in low or zero rate in a subsequent year. This finding has implications for policy: the high rate individuals are not durable in their enterprise and so perhaps one ought to invest in prevention rather than punishment.

It is safe to assume that such surprises will be infrequent. And the longitudinal study may have to be refreshed, in the interest of generating understanding that is not obvious.

To refresh and invigorate the longitudinal study, it seems intellectually justified to consider joining policy experiments to the enterprise. That is, one guarantees surprises—new understanding of a policy-relevant kind—by doing controlled experiments that are designed to inform policy. The regimens tested are, of course, unknown with respect to their effectiveness. On this account they also assure new understanding.

Consider, for example, Chapter I program evaluations. The expectation of some observers, to judge by P.L. 100-297, is that such programs will indeed affect truancy. A national longitudinal study may detect no effect of a program on truancy simply because a national study cannot measure as specifically, frequently, and reliably as is desirable; nor is it reasonable to expect that despite the enormous variation in such programs all will be directed toward truancy. Controlled tests of programs that replicate what appear to be the best of the *existing* programs might then be undertaken in sites that do not have such programs.

In the case of the Program in Human Development and Criminal Behavior, one might also refresh the longitudinal study periodically by undertaking experiments. For instance, handling of students at risk of further truancy varies a great deal. Ethnographic studies of the sort implied by Cooley (1988) may help to identify how most schools handle the matter and how the most conscientious do so. Designing formal programs based on what *appears* to be the best and testing these in a variety of settings is likely to be at least as important, more important perhaps, and as newsworthy as a longitudinal finding that "truancy is associated with delinquency and subsequent crime."

Related Research Policies and Origins

Precedents exist for coupling prospective methodological experiments to ongoing surveys. The Bureau of the Census, the Social Security Administration's Office of Research and Statistics, and other agencies have undertaken experiments to assess the validity of information reported to them. Measurement error and validity studies have, for example, preceded or been adjoined to the National Longitudinal Study of the Class of 1972 and the Adult Literacy Survey. In the social scientific community, the general Social Survey, which regularly employs split-half designs to study such phenomenon as the effects of question ordering.

More pertinent here is a recent effort to evaluate the USDE-sponsored Even Start programs. The program directs attention to family literacy and support services for preschoolers. Robert St. Pierre (1993) and his colleagues executed randomized experiments in five purposely selected sites to assess the relative effects of the programs. Alongside its effort, a National Evaluation Information System (NEIS) was exploited to provide information on each of a much larger number of Even Start sites to provide another estimate of program effects. Estimates, incidentally, differ on outcome measure and reasons for the differences are being explored.

Earlier precedents exist. Fraker and Maynard (1987), for example, reported on comparisons between controlled experiments and selection model-based analyses of survey data in evaluations of manpower training programs.

The proposal adjoining experiments to longitudinal surveys is related, of course, to piggybacking in observational surveys, i.e., adding questions to a questionnaire to meet the special needs of sponsors or the public. It is related also to the common practice of augmenting samples to investigate special groups that cannot be explored in a conventional national probability sample. The sample augmentation procedure of the National Assessment of Educational Progress, for example, permits states to add respondents within their states so that confident statements can be made about the state's students' achievement test scores, statements that would have not been possible with the survey's national sample design.

The satellite policy proposed here differs from earlier policies and precedents in that it suggests that the studies adjoined to the survey be *prospective randomized tests* of programs, substantive program variations, or their components. Such studies are not designed primarily to inform the methodologist; that aim is important but secondary here. Rather, they are designed to help understand what works better. The distinction is an important one insofar as social experiments

engender problems that are not encountered (or are encountered in less extreme forms) in methodological experiments.

The proposal for joining experiments to ongoing longitudinal surveys has origins in the debate among scholars and bureaucrat-scholars about how much one can depend on longitudinal data. It shares an interest with those who have discussed the issue of combining experimental and sampling structures (Fienberg and Tanur 1986; 1987b). There is no doubt about the need for such data for understanding change. The debate lies in whether these data can be used sensibly to understand the causes of change.

Making comparisons between results of controlled tests is sufficiently important to evaluation policy in AIDS prevention that the National Academy of Sciences urged that agencies such as the National Science Foundation sponsor research on the topic (Coyle et al. 1991).

The National Research Council's Panel on Criminal Careers makes longitudinal study paramount in its proposed research agenda (Blumstein, Cohen, Roth, and Visher 1986). Randomized field experiments are considered generally in the context of longitudinal study as a device to test hypotheses emerging from such study and to test projects in prevention, criminal career modification, and selective incapacitation. Specific linkages between each approach to understanding are implied but not discussed in detail.

Similarly, the National Academy of Sciences' Committee on Youth Employment Programs examined major studies to understand whether one could draw firm conclusions about program effects from earlier research (Betsey et al. 1985). The committee concluded, among other things, that longitudinal surveys are no substitute for randomized experiments when the object is to estimate the effectiveness of new youth employment programs. Moreover, the committee urged the use of randomized experiments for this purpose; a satellite policy is discussed in an appendix to its report.

The proposed guideline for coupling randomized design to longitudinal surveys can also be traced to a technical advisory committee for employment program evaluation appointed by the Department of Labor. The DOL sought to learn whether analyses of manpower programs based on conventional longitudinal surveys against estimates based on randomized trials. The conclusion of this exercise was that the two estimates are not always in accord. Indeed, they differ remarkably.

The justification for the coupling of longitudinal, cross-sectional and other surveys with randomized experiments appeared in the early 1970s. In particular, the Social Science Research Council's Committee on Experimentation as a Method for Planning and Evaluating Social Interventions devoted considerable attention to the problem of generalizing from experiments.

The Committee produced two state-of-the-art monographs: Riecken et al. (1974) and Boruch and Riecken (1975), as well as a variety of papers. One of these papers concerned the coupling of randomized experiments to "approximations to experiments," such as longitudinal surveys and the models used to underpin their analyses (Boruch 1975).

Proposals for adjoining experiments to longitudinal and some cross-sectional studies have since this early work been presented formally to policy boards responsible for enhancing databases and their utility. The groups include the Policy Advisory Board of the National Center for Educational Statistics (1982), the Policy Advisory Board of the National Assessment of Educational Progress (Boruch and Sebring 1983), the National Science Foundation's Human Resources Division (1982), and others.

Examples of the Contexts to Which the Satellite Policy Is Relevant

To illustrate the kinds of setting to which the proposal is pertinent consider some examples. In what follows, different longitudinal studies and different experiments are considered. The settings bear on out-of-school youth and young adults, high school students, and children in early grades who are at risk.

Chapter I Evaluation

Consider Broward County's AIM project as a possible model. The project was targeted at 2nd graders at risk of academic failure. Risk was determined by the students' performance below the 26th percentile on the Iowa Test of Basic Skills. The AIM program involved random selection and assignment of these students to all-day programs in small classrooms, with an emphasis on basic skills; the classes were being taught by specially selected teachers.

The project was undertaken in a district that has considerable standardized testing and a Research Department that is active. The experimental field test of the AIM project exploited the testing and research infrastructure in several ways that can be emulated in evaluating Chapter I programs.

- Candidates for the program were identified on the basis of regular testing, i.e., low ITBS scores;
- Impact of the program was based on the ITBS administered to project participants and comparison students;
- Routinely collected administrative records on absences and behavior problems were used to understand implementation and outcome;
- Specialty tests were developed to capture localized differences between the randomized AIM and non-AIM students; and
- The administrative system for tracking students was used too.

Not all school districts are interested in improving programs in ways that are testable, of course. Not all schools have sufficient numbers of students at risk to justify the investment in either program innovation or formal test. Broward County School District is, for instance, the largest in the country.

The implication is that not all districts with Chapter I programs are capable, much less willing, to emulate such tests. Nonetheless, the Broward experience can help to inform the work

of others, and to inform the way we think about coupling experiments to surveys and to routine administrative and academic information systems.

Multicohort-Multicity Longitudinal Studies of Delinquent Behavior

Consider surveys currently being designed by the Program on Human Development and Criminal Behavior. These surveys are relevant to proposals for Chapter I evaluation in the sense that both studies are longitudinal in character, are likely to focus on at least some common outcome variables such as truancy, and will be national in scope.

It is not hard to identify potentially interesting experiments that might effectively exploit a longitudinal study infrastructure and be worth doing. In fact, the number of options is sufficiently great to make choice difficult. The feasibility of any option may then be the determining factor, e.g., willingness of the site's public service agencies, such as police departments or courts or community-based organizations to cooperate.

For example, relatively innocuous and small but useful side experiments might be adjoined in all longitudinal studies to determine which methods are most effective locally in eliciting cooperation in the main longitudinal study or in improving the accuracy of reporting on delinquent or criminal activity. A strategy that comports with this aim might simply replicate and improve earlier experimental tests of such methods, such as the following:

- Malvin and Moskowitz (1983) on drug attitudes and use among junior high school students;
- Goodstadt and Grusen and others on the use of randomized response and other methods for eliciting sensitive information (Boruch and Cecil 1979);
- Bradburn and Sudman (1981) and others on alternative methods of interviewing and questionnaire design to improve data quality; and
- Potentially useful experimental tests are implicit in Weis (1987).

For adolescent or in-school cohorts, it may be desirable and feasible to design and test programs based on a variety of theoretical perspectives. Differential association theory (Ohlin 1988), for instance, suggests that association of target adolescents with others who are more or less delinquent will affect the targets' delinquent behavior. To the extent that school-based programs (e.g., that focus on unacceptable social behavior) or programs that attract individuals who are out of school into employment or other programs are worth testing, the longitudinal infrastructure will facilitate such testing. The extent to which shifts in association can be controlled at all seems worth testing in a controlled education, sociological, and training context.

Taking this idea further, Reiss (1986) reviewed available research on co-offenders generally. He endorses the idea put forward by Klein and Crawford that external sources of cohesiveness of gangs, if eliminated, would lead to gang dissolution or degraded cohesion. He recognizes that

conventional approaches, e.g., incapacitation and social work attention, do not reduce internal cohesion and, on the contrary, may increase it. The options that are explicit in the Reiss paper and that lend themselves to experimentation include the following:

- Court-oriented efforts to sanction co-offenders in ways that are different from sanctioning individuals (to increase sense of risk), e.g., early sanctions to all co-offenders;

- Interventions designed to reduce external sources of cohesiveness (e.g., threats from gangs, revenues from drug sales); and

- Intervention designed to disrupt recruitment of co-offenders.

Consider now a different kind of coupling, one that involves a randomized test, a time-series analysis, and longitudinal study. The idea of combining these has precedent in at least one major economic effort: the Experimental Housing Allowance Program. In EHAP, poor families with certain cities were randomly assigned to various kinds and levels of housing allowance (e.g., for home repairs). In other cities involved in so-called saturation experiments, the providers of housing were given federally subsidized support to understand how to enlarge the supply of quality housing for the poor; the estimated effect in these projects was based on time-series analyses.

Related kinds of couplings have been planned but not executed in Wisconsin. Irv Garfinkle and his colleagues have begun randomized experiments on better ways to extract child support from delinquent fathers. And to understand how communitywide interventions affect such payment, saturation tests have been designed for county-level implementation. It is conceivable that similar randomized tests and nonrandomized time series or panel analyses can be executed in other areas, in the interest of understanding how to assure that young, out-of-home fathers provide financial support to their children.

Alex Weiss (1988) has considered the merits and shortcomings of randomized experiments on police handling of crime. His stress on the use of time-series approaches suggests a coupling of the approaches. So, for instance, if the general effects of delinquency deterrence are plausible at all they ought to emerge from communitywide programs that focus on norms, associations, handlers, sanctions, and so on. And in some geographic areas, pertinent saturation experiments that exploit time-series *or* longitudinal data may be feasible. Elsewhere, deterrent efforts that focus on offenders and co-offenders might be designed and tested in randomized experiments that also include long-term (longitudinal) followup.

Consider the NLS−72, HS&B, and NELS:88. These surveys are costly and widely used by the educational research and policy community. They are sponsored by NCES and have led to a variety of provocative reports, e.g., Coleman et al. (1982).

There are a variety of reasons why such studies are relevant to proposals for a Program on Human Development and Criminal Behavior. To the extent that the Program or Chapter I evaluation will involve study of the onset and resistance of delinquency among in-school children, the NCES longitudinal studies might be augmented to focus on the high risk geographic areas and people that

are of primary interest. Questions might be added to ordinary questionnaires to add to the fund of knowledge.

More to the point, consider that the Program in Human Development and Criminal Behavior may be in a position to augment not its own longitudinal survey, but future NCES surveys or waves of measurement. That is, if the program invents, extends, or facilitates the invention of programs that reduce delinquency among high school students, then the Program's interest in testing them could drive the tests beyond its own borders. The drive may stem from inadequacy or irrelevance of its own target samples, or from simple interest in better use of institutional resources.

For instance, differential association theory explored by Ohlin suggests that an individual's resistance to crime results in part from a change in associations, notably a change from criminal associations to noncriminal. Inducing and maintaining such a change may involve jobs, military service, or other special handling methods. Programs designed to do the job should take account of history in locations, number of those at risk, level of risk, and so on. Information about these are available or can be collected at marginal cost from target areas in a national NCES survey. Further, the relations between the survey and local sites are sufficiently good to consider providing opportunities to do side experiments on effectiveness of such programs.

The example implies a link between delinquency research and educational research. Why would a federal office of educational research and statistics benefit from an explicit satellite policy more generally? There are several reasons. First, issues of data and resource sharing have emerged often during meetings of advisory committees for HS&B and NLS, and it seems reasonable to expect their reoccurrence. It then seems sensible to develop a program of joining experimental studies to these surveys that would help such committees and their staff understand how to respond to these issues equitably and efficiently.

Beyond this, it is not difficult to identify major survey-based studies and related multisite controlled field experiments. For instance, NELS:88 is being used to try to assess the effects of precollege programs, such as Upward Bound, on persistence in school, college applications, and so on. A series of controlled experiments on Upward Bound (Trio more generally) are being run independent of this. NLS−72 and NELS:88 have been used to study the effect of various factors on dropping out. There are over a dozen controlled experiments in the field designed to understand whether USDE dropout demonstration projects work.

Employment and Training

Let us suppose that randomized trials of employment and training programs are not always appropriate or feasible. Suppose further that there is some interest in learning from such trials, especially through using longitudinal surveys as a vehicle for their implementation. How might such experiments be carried out? Several strategies may be appropriate, and were reflected, for example, in early plans to evaluate programs of the Job Training and Partnership Act (Bloom et al. 1987). All of the following discussion assumes that experiments can be conducted in a way that permits one to take advantage of the longitudinal data and the organization structure used for its collection without disrupting that process.

Specific components of full programs may warrant testing. For example, we know very little about when, why, and how different varieties of job counseling "work." Mounting experiments in a selection of sites to assess the effects of the components of an employment and training program will often be more feasible and perhaps more appropriate than national trials on full-blown programs. See, for example, Bickman (1985) on assessing preschool programs for children in Tennessee.

Augmenting the existing employment and training regimens may be feasible in some sites. For example, how "residential" does residential training have to be? We know that some residential programs work (e.g., the Job Corps). We do not know how brief the residential experience can be while continuing to be effective (see, for example, Betsey et al. [1985] on such programs).

There is little good evidence to help answer the question "Does it 'pay' to treat the most needy, rather than the least needy?" The most "trainable" people (i.e., those most likely to benefit from training) often lie at the margin of need. And this margin often defines a population for which randomized trials are likely to be most feasible. Randomization at the margin can be coupled with other designs as well, e.g., regression-discontinuity (Riecken et al. 1974).

Selecting only the best of an array of research sites that are capable and willing to conduct experiments will not give fair estimates of the impact of programs. But such sites will demonstrate the best that can be done, thus providing evidence that may be sufficient for purposes of making policy and producing research that is heuristically rich for the social sciences.

Probable Issues and Options

The idea of adjoining field experiments periodically to longitudinal surveys is not new. But it has not emerged often and this accounts perhaps for the scarcity of thoughtful papers on the topic. Another reason for this scarcity may be the difficulties of executing the idea.

Some of the difficulties are resolvable given the current ability of research managers and manager researchers. Others require more thinking and perhaps pilot tests.

The following considers issues and options that are general, i.e., not depending on whether the experiments are adjoined to an existing longitudinal study or to a proposed study. Respondent burden is important regardless of design for example. It also treats issues that depend on whether the experiment is adjoined to an existing study, e.g., proprietary interests, or to a proposed one.

Standards for Joining Field Experiments to Ongoing Surveys

The proposal put forward earlier suggested that adjoining experiments to a longitudinal study be regarded as a legitimate research as long as

1) the experiment is compatible with the mission of the longitudinal survey;

2) the risks of disruption to the survey can be managed;

3) designated contractors are responsible for oversight of the process; and

4) the experiment engenders no appreciable cost to the agency supporting the longitudinal research.

Adhering to these standards is likely to reduce or eliminate obvious problems.

Still, one must decide which of a variety of potential experiments should and can be adjoined to the longitudinal study. Greenwood's draft paper (1988) lays out five criteria that help in making a choice. Paraphrased, the criteria include the following:

1) theoretical importance of the program(s) proposed for experimentation;

2) empirical evidence for the worth of the program(s);

3) "amount of difference" between proposed regimens and current practice;

4) compatibility with the longitudinal design; and

5) political feasibility.

The fourth item of course is part of the Boruch-Pearson (1988) proposals. Discussions and criteria for understanding political and managerial feasibility are important and have been given in, among others, Chelimsky's edited volume (1985) on evaluation at local, regional, and federal levels of government, and in Riecken et al. (1974) on managerial, ethical, and institutional and political issues, engendered by social experiments.

Greenwood's second criterion implies that evidence ought to be available from quasi-experimental or other randomized experiments. It seems sensible, given the likely cost of mounting new experiments, the need to anticipate outcomes, and the need in most field experiments to rely on earlier pilot testing of randomization procedures, measures, and negotiation strategies (Boruch and Wothke 1985).

Criterion number three is interesting in part because one can easily argue two sides. To the extent a difference between proposed regimens and existing control regimen is small, then detecting a difference in outcome will probably be difficult and perhaps not worth the effort. On the other hand, a small change is likely to be politically and managerially more feasible than a large one.

Similarly, to the extent that the difference between proposed regimen and existing control regimen is large, differences in outcome are likely to be more detectable and the product may be useful on policy and theory ground. But the managerial problems may be difficult. The handling of this matter by Riecken et al. (1978) is to encourage some testing of extreme program levels, the reasoning being that most interventions are weaker than they are predicted to be and that effects are, if the variation is effective, more detectable (pp. 33–34).

Adjoining Experiments to Existing Surveys

Proprietary interests of researchers are important, of course. The principal investigators in a longitudinal study such as a Chapter I evaluation may be disinclined to permit another research

group, such as the Program on Human Development and Criminal Behavior, to augment Chapter I samples or questionnaires because this would capitalize on the Chapter I infrastructure, expertise or ideas. It would yield no obvious benefit to the Chapter I researchers. Similarly, the major sponsor for a Chapter I evaluation, the U.S. Department of Education, may see no benefit in sharing credit for an important survey by cooperating with another federal agency, e.g., the National Institute of Justice.

Some ways, quid pro quos, to meet proprietary interests then must be developed to make satellite policy possible. The National Opinion Research Center, for instance, operates HS&B and is under no obligation to cooperate with organizations responsible for surveys or experiments in another area. Moreover, developing such an obligation through contract and negotiated agreements may be difficult. There are few precedents for interorganizational cooperative research in policy and social science research. There are none for the satellite research of the kind proposed here.

Adjoining Experiments Regardless of Longitudinal Study Type

Respondent burden is and will continue to be important. For example, if an experiment on effects of Chapter I program variations asks a substantial fraction of children in early grades in a set of school districts to respond to a questionnaire and a separate study of delinquent behavior directs other questions to the same individuals, the burden on the respondents and their guardians (who must provide consent) may be increased and be notable.

Monetary payments may offset the burden. Indeed, the experience in at least some studies of adolescents suggests that payment leads to not only good cooperation of the target sample members but to requests to cooperate from those outside the sample (Howard et al. 1988).

Monetary incentives are irrelevant if there is competition for respondents in any real sense. That is, if local rule or custom dictates that the respondent can participate in only one study, then payment by a second aspiring researcher will not be relevant.

Further, monetary payments to respondents ought not be relevant if the experiment adjoined to the ongoing survey can disrupt the survey. In this case, augmenting the basic sample targeted for survey may be the only way to obtain additional information for the experiment.

Similarly, and more important, an experiment adjoined to a survey will disrupt the results of a survey in a special sense. For example, the survey researcher requires that members of the sample encounter "ordinary" conditions. The experiment will perforce introduce an extraordinary condition, albeit for a small fraction of the sample. The experimental regimen will, if effective, then affect the estimates of prevalence for incidence that are important to the longitudinal study. Again, the only resolution to this problem appears to be augmenting the sample targeted in the longitudinal study.

Augmentation of a targeted sample to reduce individual respondents' burden then may help to resolve one problem but it generates another. If a central federal, state, or local agency dictates the permissible total number of respondents, then the tactic does not help. Paying additional respondents may do so, as might other tactics.

Feasibility and Appropriateness of Experiments

Conducting controlled experiments to plan and evaluate new programs, program variation, or components is no easy matter. This is regardless of whether the experiment is coupled to a longitudinal study.

The standards for judging their appropriateness and feasibility have been laid out elsewhere, e.g., Boruch (1985). Put briefly, appropriateness hinges on answers to questions such as the following:

- Does current practice need improvement?
- Is there important uncertainty about the proposed innovation?
- Will methods other than randomized experiments yield good estimates of relative effectiveness?
- Will results of the experiment be used?

These are closely linked to standards for ethical propriety of experiments.

The standards for feasibility hinge on answers to the following questions:

- Have standards for appropriateness and propriety been met?
- Are technical and financial and human resources sufficient?
- Is the process of the new program or variation understood, described, and capable of replication?
- Is the target group and context well understood?

Methods for addressing these questions and enhancing feasibility are discussed in Bloom et al. (1987), Betsey et al. (1985), Boruch and Wothke (1985), Riecken et al. (1974), Boruch and Riecken (1975), among others.

The human resources are perhaps most important in assuring quality and feasibility of controlled experiments. For Chapter I evaluations, it seems clear from precedent that some school districts have relevant capacity, e.g., Broward County, Florida; and Austin, Texas. Some, not all, of the Chapter Technical Assistance Centers are likely to have the expertise necessary to provide counsel to school districts on the use of randomized tests for program improvements (Reisner, Turnbull, and David 1988). Indeed, directors of TACs, such as Echternacht, constitute a resource that can be capitalized nicely in this arena.

Summary

- Longitudinal surveys based on well-designed probability samples are the best possible approach available to describing growth of individuals and change at the national level. Such surveys often do not yield defensible estimates of the effect of intervention, e.g., Chapter I programs.

- Controlled randomized experiments are the best possible approach to estimating relative effects of interventions, program variations, and so on. They are often not feasible at the national level, however.

- Coupling controlled randomized tests to longitudinal study can provide both understandings of growth or change and unbiased estimates of what works better in more local contexts.

- A formal policy for coupling experiments to longitudinal study then seems sensible. Such a policy is analogous to research policy in satellite use. The major vehicle for generating information, the satellite, is periodically reoriented and partly dedicated to special experiments and is analogous to the longitudinal study system.

- The main justification for the proposed satellite policy for Chapter I is scientific and policy relevant: better data to inform policy about how to improve programs. The secondary reasons include: economic ones, e.g., local experiments capitalize well on longitudinal infrastructure; methodological reasons, e.g., learning about how to improve data quality generally; political reasons, notably permitting answers to several questions.

- Selection of interventions for experimentation should be guided by several criteria: theoretical import of the intervention, empirical support for its promise, propriety of a test, feasibility of implementing both the interventions and the randomized experiment.

- In Chapter I, replication of exemplary projects may meet all these criteria. The experiments may for example test new ways of sustaining parental involvement, reducing dropout rates, decreasing low grades and failures, tutoring, and so on.

- Executing controlled experiments in Chapter I projects requires resources: well- trained researchers and practitioners and support for both. Failure of some projects is likely because learning how to improve and generating evidence on it is difficult. Assuming a failure rate of 20 percent for executing the experiment (regardless of program success) is reasonable.

- Statistical characterization of the target groups (who is eligible, who gets service, and so on) is essential for design of the experiments, as is careful literal and statistical description of the processes engendered by the program, e.g., time in Chapter I variation, nature of variation. Both can be generated at least crudely by longitudinal study.

- Theory will be important in the longitudinal study to estimate effects at the macro-level. The experimental programs will, if based on similar theory, help to adjust statistical vulnerability of the longitudinal work.

- A major legislative implication of this perspective is that mandates for longitudinal study must also authorize demonstrations, i.e., implementations of new programs, variations, and components.

LINKING NCES SURVEYS AND DATA FROM OTHER SOURCES

This essay concerns linking different data sets. The main vehicles for understanding in what follows are a volume edited by Hilton (1992); *Using National Databases in Educational Research*; a paper on the analysis of multiple surveys by Hedges and Nowell (1995); material generated by NCES for the NCES Advisory Council on Education Statistics; reports generated by scholarly groups such as Boe and Gilford (1992); Board on Children and Families (1995); and others. The purpose is to educe the implications of analyses undertaken on multiple data sets, in the interest of improving the design of NCES surveys.

The minutes of the NCES Advisory Council on Education Statistics (ACES) have reflected periodic interest in linking or integrating NCES-sponsored surveys. Recall, for instance, Griffith's presentation (1992) to the ACES. The agency has also sponsored scholarly work that depends implicitly on a capacity to link data in a variety of senses. Scheuren (1995), for instance, developed a variety of provocative ideas whose value hinges on linking records, record sets, or statistical data sets. The presumption here is that the general topic will continue to be of continuing interest to NCES.

Background

The Hilton (1992) book's origin lies in a project undertaken by the Educational Testing Service to understand whether different sources of statistical information, each based on national samples, could be combined to produce a "comprehensive unified database" of science indicators for the United States. Sponsored by the National Science Foundation, the project's general goal was to improve the way we capitalize on data that bear on educating scientists, mathematicians, and engineers. The book's implications, inadvertent and otherwise, are arguably important for designing NCES surveys.

Twenty-four education databases were reviewed by the project. They included the Survey of Doctoral Recipients, National Teacher Examinations, and at least four massive longitudinal studies of high school students undertaken with NCES support. Of the 24 ostensibly related databases, only 8 were deemed worthy of deeper examination. That is, they could be "linked," in some sense, given the resources available. They included the NCES NLS−72 and NELS:88, the Equality of Opportunity Surveys (1960s), cross-sectional efforts such as the SAT, and the NCES National Assessment of Educational Progress (NAEP).

As Hilton made plain in the book's preface, the project was "not feasible." Put more bluntly, the ETS effort to combine data sets was a flop despite competent and thoughtful efforts. The databases that were chosen for examination could not be used for the purpose considered, i.e., to produce a comprehensive science database. It was nonetheless a project noble in aspiration and diligent in execution.

The questions that were posed about the available databases and which are relevant to linking any data sets, seem important for designing new NCES surveys. Put in modified term, the questions are as follows:

1) What *variables* are common to various databases?

2) What *ways of measuring* the variables, *ways of sampling*, and administration are common, making comparison (or linkage) among data sets easy?

3) What *differences* in ways of measuring, administration, and sampling make comparison (or linkages) dubious or difficult?

4) What can be done to *fix* different data sets so they are "comparable" (or linkable) in some way and therefore make it sensible to put them together?

The Hilton book contained no detailed catalog of why the databases failed to meet one or more of the criteria implied by the questions.

Hedges and Nowell (1995) attacked a different but related topic, understanding sex differences in tests of mental activities of various kinds based on disparate surveys. They chose to depend only on studies based on samples of roughly the same target populations and that purportedly measured the same abilities, e.g., reading. They selected only studies that approached questions 1) and 2) above in similar ways. Their group of studies included NCES-sponsored work, notably NELS:88, NLS−72, HS&B, and NAEP (trend data only), and Project Talent and the National Longitudinal Youth Survey sponsored by the Department of Labor.

There was sufficient commonality in what was measured on whom in the Hedges-Nowell ambit to produce an informative analysis. It is a fine illustration of combining data sets in the interest of how males and females differ on mental abilities. Moreover, the dependence on well-defined national probability samples avoided the inferential problems in earlier studies, notably depending on self-selected samples (as in SAT/ACT testing), idiosyncratic samples (for example, in test norming), and distributional assumptions (to get at characteristics of extreme scores).

Questions That Have Been Addressed

What plausibly accounts for a decline over a decade in Scholastic Aptitude Test (SAT) scores? In the Hilton's (1992) book, Beaton, Hilton, and Scharder take the 1960−72 decline seriously, based on combining SAT cross-sectional data. Their analyses suggest that a decline on account of real reduction in student ability alone is unlikely. Over the period, the number and heterogeneity of youth who took the test (from 2 to 3 million) increased. There were increases in the number of youth at the bottom of the test score distribution (from 2,000 to 54,000).

What might account for cross-sectional declines in the mean visual-spatial test scores achieved by high school seniors in 1960 and seniors in 1980? Hilton argues that a portion of the decline is not attributable to any real change in ability. Rather, he maintains that it is attributable at least partly to increases in high school completion rates during the period (from 67 percent to 74 percent) and related demographic changes. The available data evidently were insufficient to illuminate competing explanations such as changes in curriculum. Oddly, Hilton ends all this with a *non sequitur*. He said that differences in sampling method and administration are such that "what

the net effect of all these may have been impossible to say. The conservative position is that they balanced each other."

Are the tests given to large numbers of students measuring roughly the same thing over long stretches of time? Based on factor analyses of test scores of 1972 and 1980 senior high school cohorts, and of scores from longitudinal testing, Rock (1992) maintains that there has been no real change in factor structure despite (unspecified) changes in ways that tests were administered and characteristics of students.

Is it possible to say much about the persistence of a youth cohort's interest in science over a 2-year period and about whether cohorts born a decade apart are similar in their persistence? Valerie Lee's (1992) analyses were based on NLS—72 and a followup of them, and on HS&B, a longitudinal study that includes a cohort of 1982 seniors. There were radical changes across the cohorts: both above-average and below-average students, in more recent years, leaned toward science and mathematics. Within the cohorts, the rate of declaring science, math, and engineering as a course of study dropped remarkably regardless of racial/ethnic category.

The analyses in the Hilton book dedicate much attention to the methodological problems of exploiting two or more databases in combination rather than to substantive research results. Consider the following:

Even in studies designed as longitudinal efforts, the structure of a question's bearing on a particular topic may change dramatically over time. This means that the longitudinal changes in the trait that is targeted by the question will be difficult or impossible to discern. Lee's paper in the Hilton book was instructive on this account.

Lee suggested that one could in principle construct a question addressed to high school students about their planned major course of study and a parallel question addressed to the same students when they reach college level about their actual major course of study. To judge from Lee's work, the investigation of persistence of students' interest in science is thwarted by remarkable differences in the way the relevant question has been handled. Multiple-response categories in one round of a longitudinal survey have been followed by open-ended questions with not clearly related coding categories in the next.

Similar changes in question format, in repeated rounds of a longitudinal survey or across different surveys, affect measures of achievement across time (unless special provisions are made for equating tests that are not comparable), attitudes toward science, and so on.

Less obviously, skip and detour patterns in otherwise similar questionnaires may differ. The result can be (and for certain studies has been) the elimination of information from one target sample/database and the production of information in another. For instance, students who said that they were oriented toward vocational education in a high school level survey were then asked to skip a block of questions bearing on college. Some of these students changed in their interest and went on to college. The loss of the block of questionnaire items on those who changed their orientation is important in its own right. Further, information available on them from a later survey round differs from that on college students who did express an early interest in college.

Whether to survey individuals who dropped out of school has differed across longitudinal surveys. Following dropouts is more common now. But the noncomparability means that some data must then be ignored, i.e., on dropouts. This means that some analyses cannot be done, for instance, on what happened to dropouts from high school in the 1960s versus the dropouts of the 1970s or 1980s.

The Hedges and Nowell study (1995) was less ambitious in some sense than the Hilton project, but no less instructive. Their focus on national probability samples helped greatly to "simplify" the task of summarizing the results of multiple studies in order to learn where men differ from women in mental abilities. Further, focusing on certain abilities that were measured in each study, regardless of how they were measured, advanced our understanding. Exhibit 1 illustrates the simplification.

There was sufficient commonality in what was measured to produce comparisons. Reading ability was assessed in all six studies in the Hedges-Nowell compass, for example. This permitted the authors to recognize that differences in mean performance level between men and women are reliable but small (women surpass men) and variance across gender differs at a low level (men are more variable than women). Mathematics ability was measured in four of the six studies. Results suggest a reliable but small mean difference favoring males and again, more variance among males than females. Despite "small" differences in mean ability and variance, of course, large percentage differences can appear between the sexes. That is, remarkably more males relative to females appear in the upper tails of distributions. Further, NAEP trend data suggest that the ratio of male-to-female variance has not changed appreciably over time.

Such results run counter to small-scale studies reporting declining difference between the sexes in ability level. Independent research show high male-to-female ratios among selected "very talented" samples. The Hedges-Nowell work suggests that the ratios are plausibly attributable to small mean and variance differences, apart from "differential selection by sex" (p. 45).

The Pedigree of Efforts to Put Different Databases Together

The idea underlying any linkage study undertaken by NCES or by others is that putting together data from different sources can help us to learn something new. The combination can help to learn something that cannot be learned from individual sources.

The idea has a fine pedigree. Alexander Graham Bell, for instance, exploited the notion in his study of genetic transmission of deafness. He depended, in the late 1880s, on completed Census Bureau interview forms found strewn in a government building basement and on genealogical records from other sources (Bruce 1973).

The pedigree of linkage studies is also reflected in contemporary efforts to evaluate social programs. In studies of manpower training, it has become common to link the employment records on specified individuals to their program records, and to link these data to research records on the individuals (Rosen 1974). In agriculture, health, and taxation, there have been fine studies of why and how one ought to couple data from different sources in a variety of ways (Kilss and Alvey 1985). From papers by Scheuren (1985) and others, we may learn about contemporary history of record linkage algorithms (e.g., developed by Tepping and Felligi-Sunter), the construction of

matching rules and the information exploited in matches, the idea of linkage documentation, and various approaches to adjusting for mismatches. We can learn about the role of privacy issues and statistical analysis implications from a related body of work, e.g., Cox and Boruch (1988). We learn about appraising the benefits and costs of linkage of administrative records, or the difficulty of doing so on account of sloppy practice, from aggressive investigatory agencies such as the U.S. General Accounting Office (1986a and 1986b).

Scheuren's paper (1995) for the NCES Conference on the Future of Data Collection has a different but related pedigree line. It focused on better exploitation of administrative records in NCES survey contexts, and conscientiously exploited such records more generally. One can trace the theme to John Graunt's efforts in the 17th century to learn how to use records in the Crown's interest. Graunt exhorted the Crown to learn about the kingdom through a lens consisting of compilations of records in statistical form, on the counts of soldiers-at-arms, for instance, and the numbers of births, deaths, and so on. Scheuren, similarly thoughtful and exhortative, generates ideas and reiterates others' ideas about how to augment the administrative records and understand them better through surveys.

The title of Hilton's book, *Using National Databases*, may suggest to some readers that they can learn something about whether, why, and how massive studies are combined and used. This belief will be born of recognizing recent work on how to enhance the usefulness of statistical data. Such work has been economically oriented, e.g., Spencer's work (1980) on benefit-cost analysis of data used to allocate resources and the follow-up papers by Moses, Spencer, and others. It has been based on scholarly interest in why and how social research data, including educational and health research data, are used; Kruskal's volume (1982) is a gem on this account. The work has been deepened by serious attention to how statistical data and results are misused.

The analyses contained in the Hilton book are not burdened by this knowledge. They failed to put the ETS linkage studies into the larger context of such studies or the still larger context of design and exploitation of databases and survey. We learn about attempts to link the Armed Forces Aptitude Battery to tests given in the longitudinal HS&B survey and to SATs. But we are not told about how this would enhance science indicators or inform decisions or, most important, improve the design of surveys.

Similarly, the Hedges and Nowell paper does not consider the implication of the work for the design of better surveys that can be linked in any sense. This is despite the fact that the authors are sensitive to the implications of their work on other accounts.

Building on Efforts to Put Data Sets Together

Despite the Hilton project's considerable investment in figuring out how to put different databases together, and despite the conclusion, that the databases at hand could not be put together sensibly in the interest of science-related knowledge, the book offered little counsel on how matters might be improved. Hedges and Nowell (1995) offered no counsel either, despite what can be regarded as a successful attempt to put different data sets together to advance our understanding. Scheuren's work (1995) bears naturally on linkage, but the word and synonyms for it do not appear in this paper as it does in other products of thinking. Despite the fact that the Board on Children and Families (1995) focused on "integrating federal statistics," there is no substantial examination of

what integration means and its relationship with coupling, merging, pooling, and so forth. This presents something of a challenge.

Vernacular and Definitions

The Hilton book's vernacular is sufficiently different from technical parlance in related areas to confuse some readers. For instance, there are repeated references to "linking" and "merging" of different databases. But these terms are undefined. The reader should be aware that the terms have not been defined here either. Further, the book's use of them is, at times, *not* the same as is customary in contemporary statistical work of the sort, e.g., linkage being defined as combining micro-records based on a single common identifier. At times, the book's use of the word "link" is to imply an intention to "put together." At other times, the word "link" means to stratify the units in each database in the same way (e.g., high ability, Hispanic, and so on) in order to look at how frequencies in these strata change over time on a dimension such as persistence in studying science. The word "merge" is used to describe putting different records together that may or may not have a common source.

The phrase "pooling data" was used by Hilton and has been used by others, in the sense of doing a side-by-side comparison of statistical results from each of several different data sets. This use of the phrase is not as some readers would expect. Pooling data for some analysts means combining the data from two or more samples of the same population into one that can be analyzed as a complete sample. For others, pooling means combining the results from samples of different populations.

One of the implications of this vernacular problem for NCES is that discussion, analysis, and agreement on terminology are in order. Because there has been little standardization in educational research, NCES has, in recent years, played a leadership role in getting state education agencies to agree to common standards and definitions in statistical reporting. NCES can play a related role here, and to refresh the roles taken by the IRS Statistics of Income division, the Census Bureau's methods division, and others. That is, NCES can help to make plain what we mean by

- "Combining" data sets or surveys;
- "Linking" data sets or surveys;
- "Merging" data sets or surveys;
- "Pooling" data sets or surveys; and
- "Integrating" surveys.

Absent explicit definitions, reaching mutual understandings in the statistical community will be difficult or impossible. And most important, designing surveys so they can be linked, compared, merged, and so on will be impossible. NCES can be a leading agency in this effort.

Questions

The Hilton book provides ample evidence that questions about economic status or race/ethnicity or other important topics are asked differently across surveys and data sets. Differences prevent straightforward comparison. There are, however, no recommendations about whether and how to standardize such questions. There is no discussion of how directing two or more varieties of the "same" question to respondents in a survey can help to equate or calibrate the different questions across surveys. There is no serious exploration of whether and how imputation methods can help in doing so. Yet, we know that embedding different forms of the same question in a questionnaire, for a subsample at least, is a decent vehicle for learning about relations among questions. More general tactics might be invented, based perhaps on the test-equating strategies that have been explored by Holland and Rubin (1992), among others. Certainly the matter is pertinent to NCES' investments in learning how to integrate (and in what senses to integrate) the longitudinal and cross-sectional surveys that it sponsors (Griffith 1992).

An implication of this is that survey questions need to be designed with linkage in mind. NCES often does this implicitly, and in an ad hoc fashion. We are unaware of an explicitly written standard for doing so as part of NCES survey design strategy. Nor does there appear to be a systematic program of empirical side studies or pilot work by NCES that regularly takes linkage seriously.

Analyses

In the Hilton book, there are few substantial references to multiple independent analyses of the same data sets. Hedges and Nowell are more conscientious on this account. For example, there are no references in the Hilton work to other analyses that are suspect or arguably wrong. This can be regarded as a shortcoming. It is also symptomatic of the lack of good registry for tracking who analyzed what data set.

No federal agency or private foundation, including NCES, has an excellent system for tracking the research uses to which its data sets are put. This makes the evaluation and improvement of any given survey difficult. It makes development of better statistical design very difficult.

An implication is that constructing registries of analyses is one option that NCES might consider in the interest of improving NCES surveys. More conscientious efforts by authors and journal editors to assure the proper citation of data sets is another. A third option, related to the first two, involves better exploitation of contemporary Internet capabilities to build an informative registry of analyses of NCES data sets in the interest of improving survey design. It is described later under the topic of new technology.

It is important to maintain a sense of history in this. Three of the Hilton (1992) chapters were excerpted from reports produced in 1975, 1977, and 1983. The chapters contained no discernible updating. One concerns the declines in mean reading test score based on data generated in 1960 and 1972. There was no attempt to relate the data or the analyses to more recent arguments about test score declines. This lacuna is astounding given that President Bush and President Clinton

stressed an education agenda based on what were claimed to be declines in student performance, declines found to be misleading by these analysts.

Documentation

The Hilton book recognizes the investment that statistical analysts must make in learning the "ponderous user's manuals" for complex data files. But the book presents no deep thinking or data on the matter. Hedges and Nowell are also silent on the matter. This is a general and nontrivial issue. Learning how to learn easily about complex files and how to teach well about complex data files seems important.

Some attention is being dedicated to the topic, if we interpret properly the current efforts of NCES. The NCES has generated and issued Read Only Memory diskettes (CD-ROM) that introduce complex data less formidably than the way public use tapes have been introduced. Beyond this, it is not clear whether and how NCES invests resources in making data file documentation less difficult to deal with.

It seems sensible to expect those who have made distinctive contributions to the quality of documentation (for instance, ICPSR) to collaborate with statisticians in this task. At least one major contractor to NCES, the American Institutes for Research, actually does research on the topic of "readability" of documents. Work of this sort might be exploited by NCES to enhance the ease of use of documentation on its data files.

Naming Surveys

It may not seem difficult for some readers to keep in mind the eight studies that are used in the Hilton book. But it is for this writer. The difficulty lies partly in the disconnectedness of the book's chapters. The difficulty goes well beyond the book, and is partly numerical. The multiple pieces of any given survey must be kept in mind. One or more of five points in time in the NLS−72 may be a focus of study. Any one or more of three points might be exploited in the NCES HS&B surveys.

Part of the difficulty may also lie in our predilection to name rather than to number. It is more pleasing, perhaps, to talk about "High School and Beyond" or HS&B than about survey #8.2, just as it is for our Chinese colleagues to refer informally to the "Red Flower" factory instead of Factory #26.

The implications of this "naming" problem for NCES are not clear. There is sufficient opportunity for confusion or difficulty to argue that a name such as "NLS−72" is more informative to many potential users of data than "High School and Beyond." It seems reasonable to argue that "Wave 2" is an important amendment to study, e.g., NLS−72: Wave 2. Perhaps this is as far as we can go.

Missing Data

Missing data are ignored by analysts in Hilton's book, chapter one, by Valerie Lee. Nor was the topic mentioned in works that are at least as important, by Hedges and Nowell (1995), Board on Children and Families (1995), and Boe and Gilford (1995).

This is despite the fact that reasons for missing data and the models that might be used to impute the missing data can differ across databases, just as definitions, sampling methods, survey conditions, and so forth differ across databases. More to the point of some analyses, missing data or differences in the reasons for it are not considered in understanding whether data from different sources can be sensibly compared. See, for instance, Little and Rubin (1987) and Rubin (1987) on imputation. At bottom, this suggests that another criterion be used by NCES to make judgments about the possibility of linkages among databases: missing data.

Major Factors

Various chapters of the Hilton book remind the reader to take into account both obvious and subtle factors in using the results from different surveys that might be thought comparable: differences in the definition of the target population, sampling frame, selection of organizational units, selection of individuals within units, cooperation rates, conditions of administration, coding of open-ended responses, multiple response categories, and timing of measures. Three major multimillion dollar surveys, arguably more, have differed notably in all respects, making comparison very difficult. Yet the book offers no advice on how to better structure the portfolio of longitudinal or cross-sectional surveys sponsored by the federal government.

Understanding how to design a portfolio of longitudinal and cross-sectional studies so that they *can* be put together (compared, linked, coupled, yoked, or otherwise used) goes well beyond what the book's authors tried to accomplish. With the exception of a chapter by William Turnbull, a statesman in the educational measurement arena, and one by the editor, Hilton, they confined themselves to the tasks at hand. Since the time that they engaged in the enterprise, NCES appears to have tried to make progress along related lines.

NCES sponsors an astonishing variety of longitudinal and cross-sectional surveys, at least four of which are exploited by the ETS Project. The agency began to collect longitudinal data in 1972, initiated six longitudinal studies afterward, has been asked by the Congress for more, and has supported a large number of cross-sectional surveys. The problems of how to develop an integrated portfolio of studies, and what integration means, how to integrate, in the face of disparate demands from Congress and the educational research community, and others under the influence of severe limitations on staff size, and other factors, are formidable.

In a sense, the Hilton book helps to understand and to justify what NCES has done to integrate studies (in the NCES vernacular) if not to create "unified databases" (the ETS parlance). One NCES initiative, for instance, focused on identifying rationales and settling on a rationale for integration, and for shaping the relations among longitudinal surveys and the relation between these and cross-sectional surveys (Griffith 1992). This does not differ in spirit from the book's focus on

longitudinal study as a vehicle for a unified database but gets well beyond it. The NCES focus, to judge from Griffith (1992), is on universe and sampling frames and on how to develop agreement on each, in the interest of integration, for the design of new surveys. Hilton and his colleagues make plain that their difficulty in developing a unified database on science indications was attributable to differences in each factor.

The Hilton book alludes to factors beyond sampling that may influence the construction of unified or integrated databases. But there is no pursuit. For contemporary work at NCES and perhaps other statistical agencies, the questions are numerous and the search for answers serious. Which particular surveys are sensible targets for integration out of the portfolio of all surveys that have or might be done? How do we decide? For education surveys undertaken by NCES and others, what should be the grade span of surveys, the time between rounds, the survey's lifetime, the time between initiating new cohorts, the starting grades of cohorts (Boe and Gilford 1992) What rationale based on integration can inform the choices? How can an integration standard influence surveys on the allegedly crucial transition periods from kindergarten to preschool, middle school, and so on and durable policy issues such as the supply of science-oriented students and teachers?

These questions deserve wider attention from the statistical methods and policy communities and the disciplinary communities with which they collaborate. Here again, there appears to be fine opportunity for thinking at the National Center for Education Statistics and other federal agencies (not just statistical ones) and groups that advise them, such as the National Academy of Sciences and the Social Science Research Council.

EXHIBIT FROM HEDGES AND NOWELL (1995)

Table 1—Summary of the characteristics of the six data sets

Characteristics	NLS–72	NLSY	HS&B	NELS:88		NAEP
Year of assessment	1960	1972	1980	1980	1992	1971–92
Sample size	73,425	16,860	11,914	25,069	24,599	Varies
Population	All 15-year-olds	12th grade students	Non-institution-alized 15- to 22-year-olds	12th grade students	8th grade students as of 1988	17-year-olds in school
Abilities measured						
Reading comprehension	◆	◆	◆	◆	◆	◆
Vocabulary	◆	◆	◆	◆		
Mathematics	◆	◆	◆	◆	◆	◆
Perceptual	◆	◆	◆	◆		
Science	◆		◆		◆	◆
Social studies	◆				◆	
Nonverbal reasoning	◆	◆				
Associative memory	◆	◆		◆		
Spatial ability	◆			◆		
Mechanical reasoning	◆		◆			
Electronics information	◆		◆			
Auto and shop information			◆			
Writing						◆

NEW TECHNOLOGY

Introduction

The object here is to describe how NCES might use the Internet and the World Wide Web (Web), the Internet's graphical component, to improve the design of surveys. The main vehicle of illustration is George Terhanian's Home Page (HTTP://www.dolphin.upenn.edu/~terhania/). It relies on subtechnologies available to NCES that can in turn be exploited to improve NCES survey design.

Definitions: What Does It All Mean?

The Internet and the Web have spawned a large, somewhat confusing, vocabulary. It is necessary, therefore, to first provide definitions of several terms before describing how NCES might better exploit the Internet and the Web. Providing definitions that are precise is a challenge, however, because new terms continue to emerge, and the meanings of old terms continue to evolve, as the Internet and Web expand. Consider, for example, how the meaning of "server" has changed. A few years ago, "information and file provider" would have sufficed; e.g., a server provides information and files to clients. Today, this definition seems too narrow—it does not account for the capacity of a server to receive, process, and store information (e.g., responses to questionnaire items) that clients might send.

The lack of an official Internet dictionary, no matter how inchoate some terms may seem, also makes providing definitions difficult. "Electronic mail," "bulletin board," "discussion group," "listserv," and "newsgroup," for example, all refer to slightly different methods of sharing information. But discovering how these methods differ requires perseverance: a call to a computer-literate friend, a trip to the library or bookstore, an on-line database search, and so forth. These qualifications aside, the definitions (e.g., see Howe 1995; Raisch 1994) are as follows:

Electronic Mail: A system of sending information and files to anyone who has access to the Internet through an e-mail account. Messages are automatically passed from one computer user to another, often through computer networks and/or via modems over telephone lines.

Bulletin Board: A message database where any user may submit or read any message in public areas. It is also possible to post (i.e., to place for public perusal) other types of files (e.g., statistical software) on bulletin boards.

Discussion Group: A mail system through which members exchange messages. Membership in particular groups is often based on a common interest (e.g., hierarchical models) or affiliation. Separate messages are sent individually to each member.

Listserv: A mailing list server on **Bitnet**, an academic and research computer network. Listserv is Bitnet's version of a discussion group.

Newsgroup: A combination bulletin board/discussion group. Messages are placed in a central location, for example, like a bulletin board. However, like a discussion group, access to these messages is generally restricted to the particular newsgroup's members.

Protocol: A standard, or set of formal rules, that defines the method of communication (i.e., how to transmit data across a network) among computers. There are a variety of protocols. The more popular ones include Gopher, FTP, Telnet, and HTTP.

Gopher: A user-friendly protocol that relies on hierarchically linked menus. One limitation of Gopher systems is that the client may have to work through several layers of menus before locating a desired file.

File transfer protocol (FTP): A protocol that allows for the transfer of files from server to client. Although menus may exist, those that do generally lack the detail of Gopher menus.

Anonymous FTP: A variation of FTP. An interactive service provided by many Internet servers allowing any user (i.e., those who do not possess accounts) to transfer files.

Telnet: A protocol that may permit a remote client to log on to another server. This method does not permit the client to retrieve actual files, however.

Network: Computers that use the same protocol to exchange information.

Internet: The network of networks.

World Wide Web (WWW or **Web)**: Computers that communicate via the **Hypertext Transfer.**

Protocol (HTTP): HTTP differs from other protocols in two important respects: 1) it enables clients to view graphics, and 2) it relies on hypertext links. **Hypertext** or "text that is not constrained to be linear" (Magid, Matthews, and Jones 1995, p. 8) indicates a reference to another document or file type located elsewhere. To retrieve the referenced document, one need only "click" on boldfaced and/or underlined hypertext.

Uniform Resource Locator (URL): A unique address that specifies the target (i.e., a referenced document or file type) of a hypertext link.

Hypertext Markup Language (HTML): The language of HTTP and the Web. HTML requires authors to insert a variety of formatting information or "tags" on a page of text to indicate, for example, italics, underlining, new paragraphs, links to other documents, and electronic mail addresses.

Graphical User Interface (GUI): The use of pictures rather than just words to represent the input and output of a computer program. Popular Web browsers (e.g., Netscape and Mosaic) and popular computer operating systems (e.g., Microsoft Windows) make use of GUIs.

Multipurpose Internet Mail Extension (MIME): A systematic method of categorizing transportable (via the Internet) file types. A file extension (e.g., .au for sound, .xls for Excel spreadsheet file, and so on) indicates the specific file type. Transportable types of files include images, sounds, motion pictures, word processing documents, and so forth.

How Does NCES Now Use the Internet?

Aside from sending and receiving electronic mail, NCES now uses the Internet primarily to disseminate general information, reports, and raw data. It is possible, for example, to retrieve any number of NCES-produced items through the ED Gopher server. NCES asks that users not access its servers through the "somewhat cryptic" (Davis and Sonnenberg 1995, p. 136) File Transfer Protocol (FTP) method, and denies access to those who use the Telnet protocol to access its site. Until recently, NCES used the World Wide Web only to display and describe several publications (e.g., *The Condition of Education*, *The Digest of Education Statistics*, and so on) available through the ED Gopher. Since mid-November, however, hypertext versions of some of these publications have also been made available on the Web.

How Might NCES Use the Internet to Improve Survey Design?

NCES might want to consider exploiting the flexibility of the Internet, particularly the Web, to create and strengthen ties in a variety of ways with those who analyze NCES data. The rationale is that a deeper understanding of the experiences of those who analyze survey data might help NCES to design better surveys. In addition, NCES might also use the Internet and the Web to elicit, exchange, and access information from numerous sources in order to educe the implications, or at least track the development, of new analytic methods for the design of surveys.

Why Focus on the World Wide Web?

The Web possesses at least five attributes that make it an attractive vehicle for eliciting, exchanging, accessing, and distributing information. First, it enables different types of computers (e.g., IBM, Macintosh) to communicate through a common protocol (HTTP). Second, Web graphical browsers (e.g., Netscape, Mosaic, and so on) are available at no or low cost for most popular operating systems. Third, these browsers are relatively easy to use (because of their graphical interface), flexible, and powerful. They can interpret documents written in HTML, for instance, as well as several types of graphics files. Moreover, in many senses, browsers transcend protocols through their ability to access HTTP, Telnet, Gopher, and FTP servers. Fourth, the latest release of HTML allows authors to create fill-out forms (e.g., questionnaires). Fill-out forms, in particular, exploit the capacity of Web servers to receive, process, and store responses. Finally, the Web is growing rapidly—by more than 500 percent in the past year (WebCrawler 1995). There are now more than 40,000 Web servers and about 10 million daily Web users (Netscape Communications Corporation 1995). The estimates are crude, however, because the Web, for the most part, is unregulated. No official registry of servers exists and many server administrators choose not to track usage (e.g., the number of visits to a home page or the number of downloads of a particular document), although they could do so easily.

What Might NCES Do? Strategies to Elicit, Exchange, Access, and Distribute Information

This section describes several strategies, some of which are related, that NCES might implement to elicit, exchange, and access information from numerous sources. It also describes strategies to disseminate information. For an illustration, readers are again encouraged to visit Terhanian's home page at: HTTP://www.dolphin.upenn.edu/~terhania/.

Strategy 1: Elicit Information Through Fill-Out Forms and Electronic Mail

NCES might want to consider eliciting information through graphical fill-out forms and e-mail from those who are actually analyzing NCES data (e.g., licensed users). The implication is that data users are an underexploited, though valuable, resource. Questions that NCES might ask include the following:

- What methods do you employ when analyzing survey data?
- What problems pertaining to the design of NCES surveys have arisen?
- Have any journals published your work?

Analysts are not the only ones from whom NCES might elicit information. NCES is obliged, at times, to ask questions of the general public that bear on data use. The commissioner of education statistics, for example, is "responsible for providing continuing reviews including validation studies and solicitation of public comment on NAEP's conduct and usefulness" (White 1994). NCES might therefore provide a Web window (e.g., fill-out form) through which the public might either ask or answer questions about NAEP and other surveys.

Although the ability to post questions on Web pages and the capacity of Web servers to collect, process, and store responses may have direct implications for the administration of future NAEP surveys, we have focused here, and throughout, on strategies that might influence the content of surveys no matter how they are administered.

Strategy 2: Distribute Spreadsheet Files Through the Web

NCES generally distributes raw data and finished products via the Internet; that is, seeds and mature trees. There is an opportunity for NCES to distribute saplings as well. This strategy recognizes and relies on the ability of spreadsheet software, notably the most recent versions of Lotus, Quattro Pro, and Excel, to hold alphanumeric data, and graphical displays based on this data, in different sections or pages of one file. By using a mouse to click on reference tabs (i.e., links) within a spreadsheet file, the user can move from page to page.

This strategy also recognizes and relies on the ability of Web browsers to configure helper applications to interpret spreadsheet files. For instance, after the user clicks on a Microsoft Excel spreadsheet file (.xls extension) located on a Web, Gopher, or FTP server, the Web browser (e.g., Netscape), because it does not recognize the .xls file extension, will ask the user how he or she

wishes to handle the file. The user may instruct the browser either to save the file or to open a local viewer, i.e., the particular application (e.g., Microsoft Excel). The user may also instruct the browser to thereafter open the particular helper application automatically whenever a file with an .xls extension is selected.

Spreadsheet files are a natural home for information that NCES might receive from data analysts. Depending on the questions that NCES decides to ask, the file might reveal what analytic methods researchers have applied to the NCES data, names of journals that have published articles, titles of published articles, years of publication, and the like. NCES, through this method, can then count publications, create displays, for instance, of NELS:88 publications by year, and sort the information however it chooses. Periodically, NCES might also post the updated file on a Web page to provide others with access. This information might help current and potential researchers to shape their analyses and it might also lead to the exchange of information. "Why isn't my article there?" or "Here's another," researchers might say to themselves. And they might then send NCES a reference for their own particular article or for others of which they are aware. Or they might send an updated spreadsheet file to NCES, thereby eliminating much of NCES's data entry work.

NCES will have to choose which type or types of spreadsheet files to distribute. Although we recommend any of the most recent versions of Windows software because of their widespread use and hypertext-like tab features, it is not possible at this time to open, say, an Excel file with Lotus software because the tabs pose conversion problems. Nor is it possible to use Macintosh software to open a Windows spreadsheet file. NCES might therefore consider distributing a more generic type of spreadsheet file (e.g., an Excel 4.0 file) as well.

Strategy 3: Track the Emergence and Development of New Analytic Methods Through the Web

It is not always clear how advances in statistical theory or technology might affect the design of future NAEP surveys. But the question is important enough to warrant attention. NCES might then also use the Web to track such advances. NCES might post references on a Web page, or links when appropriate, to journal articles and books that describe or use new analytic methods, including multilevel modeling, meta-analysis and cross-design synthesis. NCES might also provide a Web window through which Web users report additional references and the Web addresses of informative home pages. The home page (HTTP://www.ioe.ac.uk/hgoldstn/home.html) of the *Multilevel Models Project* (MMP) that is based in the United Kingdom, for instance, is an example of one type of free information source upon which NCES might rely. Among the many resources that the MMP provide are a description of multilevel models and their applications, an invitation to join a discussion group (i.e., listserv list), and references to recent articles that use multilevel models.

Strategy 4: Create Electronic Discussion Groups (or Listserv Lists)

It seems sensible for NCES to use the Internet to connect through discussion groups or listserv lists those who share common interests (e.g., licensed data users of SASS) or constitute particular technical review panels (e.g., the SASS data user's group). Whatever communications transpire among members of such data analysis groups might then be made available to those who design NCES surveys. Providing the designers with this information exploits the capability of

electronic mailing list servers to permanently record all messages. There is precedent for creating discussion groups at NCES as well. The Advisory Council on Education Statistics (ACES), for example, makes use of a listserv, one form of a discussion group.

Strategy 5: Most Frequently Asked Questions and Relevant Literature on the Web

NCES data users may frequently ask NCES numerous questions about the data and its analysis. Posting these questions (and their answers) on the Web then seems sensible inasmuch as it may prevent those who respond to the questions from repeating themselves incessantly. The strategy complements NCES's emplaced effort to provide instruction to researchers who aspire to analyze NCES data. NCES, for example, holds seminars during the summers "to provide young scholars and researchers with opportunities to gain access" to NCES surveys (NCES 1994). Knowledge of the types of questions that data users frequently ask, moreover, might prove useful to those who design NCES surveys. For example, if a preponderance of questions were to pertain to the techniques required to model the measurement error that results from NAEP's use of plausible values, then survey designers might want to consider a variety of options for the design of future assessments, including use of a different method of estimating proficiency.

At times, NCES might also post entire documents "to make things easier for interested parties in terms of their hunt for relevant literature" (Maline 1993, p. iii) It may be, for instance, that data analysts frequently request a particular document, say, the annotated bibliography of NLS−72 studies. To accommodate such interested parties, NCES might convert this document either to an HTML or .pdf file, then make it available through the Web.

Strategy 6: Consider Using Adobe Acrobat to Disseminate Information

Posting Adobe's portable document files (.pdf) on the Web is a particularly attractive alternative for organizations that wish to disseminate information through the Web but resist the intensive editing that HTML requires. The US General Accounting Office, for example, makes available .pdf files for dissemination via the Web. NCES might do the same. If the original NCES document were a WordPerfect of Microsoft Word file that included several graphical figures (e.g., a data user's manual), NCES, after purchasing the reasonably priced Adobe Acrobat, would only have to issue a print command to create a .pdf file. The software has additional features that NCES might find attractive, as well. It is possible, for example, to include hypertext links (e.g., from the table of contents to the conclusion) within .pdf documents. Future versions of the software, moreover, will enable authors to include hypertext links from within .pdf documents to other Web locations (e.g., NCES's home page). Finally the Adobe Acrobat Reader, the application required to read .pdf files, operates almost seamlessly with Web browsers, particularly Netscape, and is available through the Internet at no cost for many operating systems.

Implementing the Strategies: How Difficult Is It?

Making judgments about which strategies to implement and in what order boils down to a cost-benefit analysis that NCES will have to do. What we can provide, however, are some final

thoughts. We base our thoughts in large part on the effort required to create this document's illustrative Web home page at HTTP://www.dolphin.upenn.edu/~terhania/.

Strategy 1: Elicit Information Through Fill-Out Forms and Electronic Mail

Setting up a Web server to receive, process, and store information (e.g., responses to questionnaires) is straightforward, although it does require some tinkering on the server end (e.g., see Magid, Matthews, and Jones 1995). The necessary resources are in place, however, because NCES has already begun to use the Web.

Developing Web pages through HTML requires some expertise. Nevertheless, it is fairly easy to capitalize on the work of others. The HTML code that underlies the creation of each file posted on the Web is available at no cost; that is, prototypical Web pages are readily available.

Strategy 2: Distribute Spreadsheet Files Through the Web

Someone at NCES must do the work. It is possible to capitalize on the work of others, however. This is one object of creating Web windows.

Strategy 3: Track the Emergence and Development of New Analytic Methods and Their Implications Through the Web

A template for acquiring and consolidating such information might be based on the list of "implication" categories given earlier in this report. The categories, put into question form, are: What are the implications of the new analysis method or its application for deciding

1) What variables to measure;

2) How to measure;

3) Whom to measure;

4) How many respondents to sample;

5) When to measure;

6) With what sample design characteristics;

7) In connection with what other data collection;

8) Why; and

9) With what reporting strategy (e.g., CD-ROM, and so on).

Strategy 4: Create Electronic Discussion Groups (or Listserv Lists)

NCES has a list of all licensed data users. Membership of NCES's technical review panels, moreover, is public information. Further, there is precedent for using mailing list servers to connect members of particular panels or groups. Creating additional discussion groups, therefore, is a logical next step. The listserv of the Advisory Council on Education Statistics is a prototype.

Strategy 5: Post Frequently Asked Questions and Relevant Literature on the Web

Posting those questions that data users frequently ask and relevant literature on the Web requires some effort. Nevertheless, many may benefit through the work of few. Moreover, NCES might capitalize on the work of others here as well. One question that NCES and its contractors may frequently field, for example, relates to the appropriate statistical procedures that must be applied to obtain accurate variance estimates with NCES surveys, e.g., SASS. These, we presume, are the "reasonably tractable procedures" to which Clogg (1989) refers. We know, for example, that SASS analysts use, among other software, a package developed at Westat called WesVarPC to apply the procedures. We know, as well, that Westat provides documentation that includes an introduction to replication methods in portable document format (.pdf). NCES, with Westat's permission, might make this file available to analysts.

Strategy 6: Consider Using Adobe Acrobat to Disseminate Information

Using Adobe's portable document format (.pdf) does not force NCES to decide among software packages. Creating a .pdf file is as simple as issuing a print command.

NOTES

1. All data from the National Center for Education Statistics that are used here apply only to public schools (and public school students).

2. “Significantly,” as used here and throughout the paper, refers to a mean difference of at least two standard deviations.

3. Alternatively, the analyst might simply use the propensity score as a covariate in an analysis of covariance—e.g., see Rosenbaum and Rubin (1983).

REFERENCES

American Statistical Association. 1993. *Proceedings of the American Statistical Association: Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Bailar, B. A., and Lanphier, C. M. 1978. *Development of Survey Methods to Assist Survey Practices*. Washington, D.C.: American Statistical Association.

Barnes, R. E., and Ginsburg, A. L. 1979. "Relevance of the RMC Models for Title I Policy Concerns." *Educational Evaluation and Policy* 1 (2): 7–14.

Barro, S. M. 1992. "Models for Projecting Teacher Supply, Demand, and Quality: An Assessment of the State of the Art." In E. E. Boe and D. M. Gilford (Eds.) *Teacher Supply, Demand, and Quality*. Washington, D.C.: National Academy Press, 129–209.

Berk, R. A. et al. 1985. "Social Policy Experimentation." *Evaluation Review* 9: 387–429.

Bernard, H. R., and Killworth, P. D. 1973. "On the Social Structure of an Ocean-Going Research Vessel and Other Important Things." *Social Science Research* 2: 145–184.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1987. "Estimating the Number of People in an Average Personal Network and an Event Subpopulation." *Proceedings of the American Statistical Association: Survey Research Methods Section*. Washington, D.C.: American Statistical Association, 17–25.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1989. "Estimating the Size of an Average Personal Network and of an Event Subpopulation." In M. Kochen (Ed.) *The Small World*. Norwood, N.J.: Ablex.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., McCarty, C., Shelly, G. A., and Robinson, S. 1990. "Comparing Four Different Methods for Measuring Personal Social Networks." *Social Networks* 12 (3): 179–215.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1990. "Estimating the Size of An Average Personal Network and of An Event Subpopulation: Some Empirical Results." *Social Science Research* 20: 109–121.

Bernard, H. R., Killworth, P. D., Johnsen, E. C., Shelley, G. A., and McCarty, C. 1994. *Estimating the Size of Uncountable Populations: A Summary of Research*. Gainsville, Florida: University of Florida.

Bernard, H. R., Johnsen, E. C., Killworth, P. D. and Robinson, S. 1994. *How Many People Died in the Mexico City Earthquake?* Gainsville, FL: University of Florida.

Betsey, C., Hollister, R., and Papgiorgiou, M. (Eds.). 1985. *The YEDPA Years: Report of the Committee on Youth Employment Programs*. Washington, D.C.: National Research Council.

Bickman, L. 1985. "Improving Established Statewide Programs." *Evaluation Review* 9: 189–208.

Bloom, H. S., Borus, M. E., and Orr, L. L. 1987. *Using Random Assignment to Evaluate an Ongoing Program: The National JTPA Evaluation*. Presented at the Annual Meeting of the Statistical Association, San Francisco, August 17–20.

Blumstein, A., Cohen, J., Roth, J., and Visher, C. A. (Eds.). 1986. *Criminal Careers and "Career Criminals."* Washington, D.C.: National Academy Press.

Board on Children and Families and Committee on National Statistics. 1995. *Integrating Federal Statistics on Children*. Washington, D.C.: National Academy Press.

Bock, D., Gibson, R., and Muraki, E. 1988. "Full Information Item Factor Analysis." *Applied Psychological Measurement* 12: 261–280.

Boe, E. E., and Gilford, D. M. (Eds.). 1992. *Teacher Supply, Demand, and Quality: Policy Issues, Models, and Databases*. Washington, D.C.: National Academy of Sciences Press.

Boruch, R. F. 1975. "Coupling Randomized Experiments and Approximations to Experiments in Social Program Evaluation." *Social Methods and Research* 4: 31–53.

Boruch, R. F. 1994. "The Future of Controlled Experiments." *Evaluation Practice* 15 (3): 265–274.

Boruch, R. F. 1995. "Comments on Droitcour and Chelimsky's *Cross-Design Synthesis*." Presented at the Annual Meeting of the American Evaluation Association and the Canadian Evaluation Association, Vancouver, B.C. (Author: University of Pennsylvania, Philadelphia, PA).

Boruch, R. F., and Riecken, H. W. (Eds.). 1975. *Experimental Testing of Public Policy*. Boulder, CO: Westview.

Boruch, R. F., and Cecil, J. S. 1979. *Assuring Confidentiality of Social Research Data*. Philadelphia, PA: University of Pennsylvania Press.

Boruch, R. F., McSweeny, A. J., and Soderstrom, J. 1978. "Bibliography: Illustrative Randomized Experiments." *Evaluation Quarterly*.

Boruch, R. F., and Wothke, W. 1985. "Seven Kinds of Randomization Plans for Designing Field Experiments." *New Directions for Program Evaluation* 28: 95–118. San Francisco: Jossey-Bass.

Boruch, R. F., and Pearson, R. W. 1988. "Assessing the Quality of Longitudinal Surveys." *Evaluation Review* 12 (1): 3–59.

Bradburn, N., and Sudman, S. 1981. *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

Breslow, N. 1989. "Biostatistics and Bayes." In M. Gail and N. Johnson, Coordinators (Sesquicentennial Invited Paper Sessions: Proceedings of the American Statistical Association.) Alexandria, VA: American Statistical Association, 51–69.

Breslow, N. E., and Clayton, D. G. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88: 9–25.

Brooks-Gunn, J., Brown, B., Duncan, B., and Moore, K. A. 1995. "Child Development in the Context of Family and Community Resources." In Board on Children and Families. *Integrating Federal Statistics on Children: Report on a Workshop*. Washington, D.C.: National Academy Press, 27–97.

Broward County School Board. Department of Research. 1987. *Achievement Through Instruction and Motivation: Program Evaluation Report for 1986–87*. Fort Lauderdale, FL: School Board of Broward County, Research Department.

Bruce, R. V. 1973. *Bell: Alexander Graham Bell and the Conquest of Solitude*. New York, NY: Little Brown.

Bryk, A. S., and Raudenbush, S. W. 1992. *Hierarchical Linear Models: Applications and Data Analysis*. Newbury Park, CA: Sage.

Bryk, A., Raudenbush, S., Seltzer, M. and Conger, R. 1989. *An Introduction to HLM: Computer Program and User's Guide*. Chicago: University of Chicago, Department of Education.

Campbell, D. T., and Boruch, R. F. 1975. "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives." In C. A. Bennett and A. A. Lumsdaine (Eds.) *Central Issues in Social Program Evaluation.* New York: Academic Press, 195–297.

Chelimsky, E. (Ed.). 1985. *Program Evaluation: Patterns and Directions*. Washington, D.C.: American Society for Public Administration (PAR Classics Series).

Coleman, J. S., Hoffer, T., and Kilgore, S. 1982. *High School Achievement: Public, Catholic, and Private Schools Compared*. New York: Basic Books.

Clogg, C. C. 1989. "Modeling Social Statistics: Current Issues." In M. H. Gail and N. L. Johnson, Coordinators. (Sesquicentennial Invited Paper Sessions: Proceedings of the American Statistical Association. Alexandria, VA: American Statistical Association, 214–225.

Cohen, M. 1994. "Intergrated Sampling of Education Institutions." (Proceedings of the American Statistical Association Survey Research Methods Section.) Alexandria, VA: American Statistical Association, 638–640.

Cooley, W. W. 1988. *Design for a Longitudinal Study of Chapter I*. Briefing to the U.S. Department of Education, Washington, D.C.

Cottingham, P., and Rodriguez, A. 1987. *The Experimental Testing of the Minority Female Single Parents Program.* Presented at the Annual Meeting of the American Statistical Association, San Francisco, CA, August 17–20.

Coyle, S., Boruch, R. F., and Turner, C. F. (Eds.). 1991. *Evaluating AIDS Prevention Programs*. Washington, D.C. National Academy Press.

Cox, L. H., and Boruch, R. F. 1988. "Record Linkage, Privacy, and Statistical Policy." *Journal of Official Statistics* 4 (1): 3–16.

Cronbach, L. J. et al. 1980. *Toward Reform of Program Evaluation*. San Francisco, CA: Jossey-Bass.

Davis, C., and Sonnenberg, B. (Eds.). 1993. *Programs and Plans of the National Center for Education Statistics: 1993 Edition*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Davis, C. and Sonnenberg, B. (Eds.). 1995. *Programs and Plans of the National Center for Education Statistics: 1995 Edition*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Draper, D. et al. 1992. *Combining Information for Research.* Washington, D.C.: National Academy of Sciences. (Also in: *Contemporary Statistics*. #1, Alexandria, VA: American Statistical Association, Undated).

Draper, D. 1995. "Inference and Hierarchical Modeling in the Social Sciences." *Journal of Educational and Behavioral Statistics* 20 (2): 115–148.

Droitcour, J., and Chelimsky, E. 1995. *Cross-Design Synthesis*. Paper presented at the Annual Meeting of the American Evaluation Association and the Canadian Evaluation Association, Vancouver, B.C. (Authors: U.S. General Accounting Office, Program Evaluation and Methodology Division, Washington, D.C.)

Droitcour, J. A., Silberman, G., and Chelimsky, E. 1993. "Design Synthesis." *International Journal of Technology Assessment in Healthcare* 9 (3): 440–449.

Duncan, G. J., and G. Kalton. 1985. *Issues of Design and Analysis of Surveys Across Time*. Presented at the centenary session of the International Statistical Institute, Amsterdam.

Duncan, G. J., Juster, F. T., and Morgan, J. N. 1984. "The Role of Panel Studies in a World of Scarce Research Resources." In S. Sudman and M.A. Spaeth (Eds.) *The Collection and Analysis of Economic and Consumer Behavior Data: In Memory of Robert Ferber.* Champaign, IL: Bureau of Economics and Business Research.

Elmore, R.F. 1993. "What Knowledge Base?" *Review of Educational Research* 63: 314–318.

Farrington, D. P. 1988. "Advancing Knowledge About Delinquency and Crime: The Need for a Coordinated Program of Longitudinal Research." *Behavioral Sciences and Law* 6 (3): 307–331.

Farrington, D. P., Ohlin, L. E., and Wilson, J. Q. 1986. *Understanding and Controlling Crime: Toward a New Research Strategy*. New York: Springer-Verlag.

Fienberg, S. B., and Tanur, J. 1986. "From the Inside Out and the Outside In: Combining Experimental and Sampling Structures." *Technical Report 373*, Carnegie-Mellon University (December).

Fienberg, S. B., and Tanur, J. 1987a. "The Design and Analysis of Longitudinal Surveys: Controversies and Issues of Cost and Continuity." In R. F. Boruch and R. W. Pearson (Eds.) *Designing Research with Scarce Resources*. New York: Springer-Verlag, 60–93.

Fienberg, S. B., and Tanur, J. 1987b. "Experimental and Sampling Structures: Parallels Diverging and Meeting." *International Statistics Review* 55: 75–96.

Fienberg, S. B., Singer, B., and Tanur, J. 1985. "Large Scale Social Experimentation in the United States." In A. C. Atkinson and S. E. Fienberg (Eds.) *A Celebration of Statistics: The ISI Centenary Volume.* New York: Springer-Verlag, 287–326.

Fienberg, S. B., Martin, M. E., and Straf, M. L. 1985. *Sharing Research Data*. Washington, D.C.: National Academy of Sciences.

Folsom, R. E. and Liu, J. 1994. "Small Area Estimator for the National Household Survey of Drug Abuse." *Proceedings of the Section on Survey Research Methods: Annual Meeting of the American Statistical Association*. Alexandria, VA: ASA, 565–570.

Fraker, T., and Maynard, R. 1987. *The Use of Comparison Group Designs in Evaluation of Employment Related Programs*. Princeton, NJ: Mathematica Policy Research.

Fraker, T., and Maynard, R. 1987. "The Use of Comparison Group Designs for Evaluations of Employment-Related Programs." *The Journal of Human Resources* 22: 194–227.

Frederikson, C. H., and Rotondo, J. A. 1979. "Time Series Models and the Study of Longitudinal Change." In J. R. Nesselroade and P. B. Baltes (Eds.) *Longitudinal Research in the Study of Behavior and Development*, pp. 111–154. New York: Academic Press.

Freedman, D. A. 1985. "Statistics and the Scientific Method." In W. M. Mason and S. E. Fienberg (Eds.) *Cohort Analysis in Social Research*, pp. 343–36. New York: Springer-Verlag, 334–36.

Gail, M. H., and Johnson, N. L. (Eds.) 1989. Sesquicentennial Invited Paper Sessions Proceedings of the American Statistical Association. Alexandria, VA: ASA.

Gerald, D., and Hussar, W. J. 1992. *Projections of Education Statistics to 2000.* Washington, D.C.: National Center for Educational Statistics (NCES 92–218).

Gray-Donald, K., and Kramer, M. S. 1988. "Causality Inference in Observations Versus Experimental Studies." *American Journal of Epidemiology* 127: 885–892.

Greenwood, P. July 1988. *The Role of Planned Interventions in Studying the Assistance of Criminal Behavior in a Longitudinal Study*. Concept Paper Developed for the Resistance Group of the Program on Human Development. Santa Monica: RAND Corporation.

Griffith, J. 1992. Presentation to the National Advisory Council on Education Statistics March 12–13, 1992: *Draft Paper on a Proposal for an Integrated Longitudinal Studies Program.* Washington, D.C.: National Center for Education Statistics.

Gueron, J. M. 1985. "The Demonstration of State Work/Welfare Initiatives." *New Directions for Program Evaluation* 28: 5–13.

Hamburg, B. 1993. *New Futures for the Forgotten Half: Realizing Unused Potential for Learning and Productivity*: *William T. Grant Foundation Annual Report*. New York: The William T. Grant Foundation.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I .M., and Snow, R. E. 1995. "Enhancing the Validity and Usefulness of Large Scale Assessments: II. NELS:88. Science Achievement." *American Educational Research Journal* 32 (3): 555–582.

Heckman, J., and Singer, B. (Eds.). 1985. *Longitudinal Analysis of Labor Market Data*. Chicago, IL: University of Chicago Press.

Heckman, J. J., and Robb, Jr., R. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In J. J. Heckman and B. Singer (Eds.) *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press, 156–246.

Hedges, L. V., and Nowell, A. 1995. "Sex Differences in Mental Test Scores, Variability,and Numbers of High Scoring Individuals." *Science* 269: 41–45.

Hedges, L. V., and Olkin, I. 1985. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Heubert, J. 1992. *Personal Communication: Class Notes from Course in Law and Education*. Cambridge, MA: Graduate School of Education, Harvard University.

Hilton, T. (Ed.). 1992. *Using National Databases in Educational Research.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoffer, T. B. 1992. “Middle School Ability Grouping and Student Achievement in Science and Mathematics.” *Educational Evaluation and Policy Analysis* 14 (3): 205–227.

Hoffreth, S. 1995. “Children’s Transition to School.” In Board on Children and Families and Committee on National Statistics. *Integrating Federal Statistics on Children: Report of A Workshop.* Washington, D.C.: National Academy Press, 98–121.

Holland, P. W., and Rubin, D. B. (Eds.). 1982. *Test Equating*. New York: Academic.

Howard, K. et al. 1988. *A Survey of Adolescents and Their Access to Mental Health Services*. Evanston, IL: Psychology Department, Northwestern University.

Howe, D. 1995. *The Free On-Line Dictionary*. (HTTP://wombat.doc.ic.ac.uk/foldoc?Free+On-line+Dictionary).

Hunter, J. E., and Schmidt, F. L. 1990. *Methods of Meta-Analysis*. Newbury Park, CA.: Sage Publications.

Johnsen, E. C., Bernard, H. R., Killworth, P. D., Shelley, G. A., and McCarty, C. 1994. “A Social Network Approach to Corroborating the Number of AIDS/HIV+ Victims in the U.S.” *NERC Oceanography Unit.* Parks Road, Oxford, England: Clarendon Laboratory.

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., and McCarty, C. 1990. “Estimating the Size of Personal Networks.” *Social Networks* 12 (4): 289–312.

Killworth, P. D., McCarty, C., Johnsen, E. C., Shelly, G. A., and Bernard, H. R. 1994. “A Social Network Approach to Estimating Seroprevalence in the United States.” *NERC Oceanography Unit*. Parks Road, Oxford, England: Clarendon Laboratory.

Killworth, P. D. et al. Undated. “Estimation of Seroprevalence, Rape, and Homelessness in the U.S. Using a Social Network Approach.” *NERC Oceanography Unit*. Parks Road, Oxford, England: Clarendon Laboratory.

Kilss, W., and Alvey, W. (Eds.). 1985. *Record Linkage Techniques: Proceedings of the Workshop on Exact Matching Methodologies*. Washington, D.C.: U.S. Department of Treasury, Statistics of Income Division, IRS.

Kreft, I. G. (Ed.). 1995. “Hierarchical Models.” *Journal of Educational and Behavioral Statistics* 20 (22): 109–240.

Kruskal, W. H. (Ed.). 1982. *The Social Sciences: Their Nature and Use*. Chicago: University of Chicago Press.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., and Snow, R. E. 1995. “Enhancing the Validity and Usefulness of Large Scale Educational Assessments: 1. NELS:88 Mathematics Achievement.” *American Educational Research Journal* 32 (3): 525–554.

LaLonde, R. 1986. "Evaluating the Econometrics Evaluations of Training Programs with Experiments." *American Economic Review* 76 (4): 604–620.

Laumann, E. O., Gagnon, J. H., Michaels, M. S., Michael, R. T., and Coleman, J. S. 1989. "Monitoring the AIDS Epidemic in the United States: A Network Approach." *Science* 244: 1186–1189.

Lee, V. E. 1992. "Pooling Data from Two Longitudinal Cohorts." In T.L. Hilton (Ed.) *Using National Data Bases in Educational Research.* Hillsdale, NJ: Lawrence Erlbaum, 246–258.

Linn, R. L. 1979. "Validity of Inferences Based on the Proposed Title I Evaluation Models." *Educational Evaluation and Policy Analysis* 1 (2): 23–32.

Little, R. J. A., and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York, NY: Wiley.

Magid, J., Matthews, R. D., and Jones, P. 1995. *The Web Server Book: Tools & Techniques for Building Your Own Information Site*. Chapel Hill, NC: Ventana Press.

Malec, D. 1993. Chapter 8. "Model Based State Estimates from the National Health Interview Survey." In Subcommittee on Small Area Estimation, Federal Committee on Statistical Policy." *Statistical Policy Working Paper 21. Indirect Estimators in Federal Programs.* Washington, D.C.: U.S. Office of Management and Budget.

Maline, M. S. 1993. *The National Longitudinal Study of the High School Class of 1972: Annotated Bibliography of Studies*. Washington, D.C.: Office of Research, U.S. Department of Education.

Malvin, J. H., and Moskowitz, J. M. 1983. "Anonymous Versus Identifiable Self Reports of Adolescent Drug Attitudes, Intention and Use." *Public Opinion Ouarterly* 47: 557–566.

Manski, C. R. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Mason, W. M., and Fienberg, S. E. (Eds.). 1985. *Cohort Analysis in Social Research: Beyond the Identification Problem.* New York: Springer-Verlag.

Mathiowetz, N. A., and Duncan, G. J. 1984. "Temporal Patterns of Response Errors on Retrospective Reports of Unemployment and Occupation." *Proceedings of the American Statistical Association: Section Q Survey Research Methods*, pp. 652–654. Washington, D.C.: American Statistical Association.

Mathiowetz, N. A. 1987. "Response Error: Correlation Between Estimation and Episodic Recall Tasks." *Proceedings of the American Statistical Association: Survey Research Methods Section*, pp. 430–435. Washington, D.C.: American Statistical Association.

Maynard, R. A. 1987. "The Role of Randomized Experiments in Employment Training Evaluations." *Proceedings of the American Statistical Association: Survey Research Methods Section*, pp. 109–113. Washington, D.C.: American Statistical Association.

Mazur, A., and Boyko, E. 1981. "Large-Scale Ocean Research Projects: What Makes Them Succeed or Fail?" *Social Studies of Science* 11: 425–449.

Messick, S. 1984. "A New Design for the National Assessment of Education Progress." *Proceedings of the American Statistical Association: Survey Research Methods Section*. Washington, D.C.: American Statistical Association.

McCarty, C. et al. 1995. *Eliciting Representative Samples of Personal Network*. Gainsville, FL: University of Florida.

Mok, M. June 1995. "Sample Size Requirements for 2-Level Designs in Educational Research." *Multilevel Modeling Newsletter*. June 1995.

Mosteller, F., Light, R., and Sachs, J. 1995. "Sustained Inquiry in Education: Lessons from Ability Grouping and Class Size." *Center for Evaluation of the Program on Initiatives for Children*. Cambridge, MA: Harvard University.

Mosteller, F., and Moynihan, D. (Eds.). 1972. *On Equality of Educational Opportunity*. New York: Vintage Books.

Mullis, I. V. S., Jenkins, F., and Johnson, E. G. 1994. *Effective Schools in Mathematics: Perspectives from the 1992 NAEP Assessment*. Research and Development Report. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Mundel, D. 1979. "Memo to Franklin Zweig" (November 15). Congressional Budget Office.

Murnane, R. J. 1992. "Who Will Teach?" In E. E. Boe and D. M. Gilford (Eds.) *Teacher Supply, Demand, and Quality*, pp. 262–270. Washington, D.C.: National Academy Press.

National Center for Education Statistics. 1995. *Agenda for a Meeting on the Future of, Education Statistics*. Berkeley, CA: MPR Associates.

National Center for Education Statistics. 1994. *Announcement: Advanced Studies Seminar on the Use of NELS:88 and SASS Data for Research and Policy Discussion*. Washington, DC: NCES, U.S. Department of Education.

National Center for Education Statistics. 1993. *National Adult Literacy Survey*. Washington, D.C.: NCES.

Netscape Communications Corporation. 1995. *White paper*. (HTTP://home.netscape.com/comprod/at_work/white_paper/index.HTML).

Oakes, J. 1990. *Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Sciences.* Santa Monica, CA: RAND.

Office of Management and Budget. 1993. *Statistical Policy Paper 21. Indirect Estimators in Federal Programs*. Washington D.C.: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget (Subcommittee on Small Area Estimation/Federal Committee on Statistical Metrology).

Ohlin, L. E. May 12, 1988. "Memo to Working Group on Desistance Regarding Policy Statement as Program Objective." *Program on Human Development and Criminal Behavior*. Castine, Maine.

Osgood, D. W., and Smith, E. L. 1995. "Applying Hierarchical Linear Modeling to Extended Longitudinal Surveys." *Evaluation Review* 19 (1): 3—30.

Pallas, A. 1995. "Federal Data on Educational Attainment and the Transition to Work." In Board on Children and Families and Committee on Federal Statistics. *Integrating Federal Statistics on Children: Proceedings of a Workshop*, pp. 122–155. Washington, D.C.: National Academy Press.

Pearson, R. W. 1987. *Researchers' Access to U.S. Federal Statistics*. Items 41: 6—11.

Pearson, R. F. and Boruch, R. F. (Eds.). 1986. *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits*. (Lecture Notes in Statistics, N. 38.) New York: Springer-Verlag.

Project Review Team. 1988. "Report on the Spouse Assault Replication Project to the National Institute of Justice." *Department of Statistics and Psychology.* Northwestern University, Evanston, IL.

Raish, M. 1994. *Network Knowledge for the Neophyte*. Binghamton, NY: Binghamton University Libraries.

Reisner, E. R., Alkin, M. C., Boruch, R. F., Linn, R. L., and Millman, J. 1982. *Assessment of the Title I Evaluation and Reporting System*. Washington, D.C.: U.S. Department of Education.

Reisner, E. R., Turnbull, B. J., and David, J. L. 1988. *Evaluation of the ECIA Chapter I Technical Assistance Centers*. Washington, D.C.: Policy Studies Associates, Inc.

Reiss, A. J. 1986. "Co-Offending Influences on Criminal Careers." In A. Blumstein, J. Cohen, R. Roth, and C. Visher (Eds.) *Criminal Careers and "Criminal Careers" Volume I.* Washington, D.C.: National Academy of Sciences.

Reiss, A. et al. 1988. "Pipeline Studies in the Spouse Assault Replication Project." In Report of the Program Review Team, Spouse Assault Replication Project, to the National Institute of Justice. Departments of Statistics and Psychology. Northwestern University, Evanston, IL.

Riecken, H. W. et al. 1974. *Social Experimentation*. New York: Academic Press.

Rock, D. A. 1992. "Pooling Results from Two Cohorts Taking Similar Tests." In T.L. Hilton (Ed.) *Using National Data Bases in Educational Research*. Hillsdale, NJ: Lawrence Erlbaum, 192–213.

Rogosa, D., and Saner, H. 1995. "Longitudinal Data Analysis Examples with Random Coefficient Models." *Journal of Educational and Behavioral Statistics* 20 (2): 149–170.

Rosen, S. (Ed.). 1974. *Final Report of the Panel on Manpower Training Evaluation: The Use of Social Security Earnings Data for Assessing the Impact of Manpower Training Programs*. Washington, D.C.: National Academy of Sciences.

Rosenbaum, P. R. 1986. "Dropping Out of High School in the United States: An Observational Study." *Journal of Educational Statistics* 11 (3): 207–224.

Rosenbaum, P. R. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science* 2 (3): 92–316.

Rosenbaum, P. R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84 (408): 104–1032.

Rosenbaum, P. R. 1991. "A Characterization of Optimal Designs for Observational Studies,"*Journal of the Royal Statistical Society* B53 (3): 597–610.

Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika* 70 (1): 41–55.

Rosenbaum, P. R., and Rubin, D. B. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–524.

Rosenbaum, P. R., and Rubin, D. B. 1982. "Comparing Effect Sizes of Independent Studies." *Psychological Bulletin* 92: 500–504.

Rubin, D. B. 1974. "Estimating Causal Effect of Treatment In Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688–701.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Scheuren, F. 1985. "Methodological Issues in Linkage of Multiple Databases." In B. Kilss and W. Alvey (Eds.) *Record Linkage Techniques.* Washington, D.C.: U. S. Department of Treasury, Statistics of Income Division, Internal Revenue Service.

Scheuren, F. 1985. "Methodologic Issues in Linkage of Multiple Databases." Prepared for the Panel on Statistics for the Aged Population, National Academy of Sciences. Washington, D.C.: National Academy of Sciences.

Scheuren, F. 1995. *Administrative Record Opportunities in Educational Survey Research*. Report prepared for the National Center on Educational Statistics. Washington, D.C.: The George Washington University.

Shaffer, J. P. (Ed.). 1992. "The Role of Models in Nonexperimental Social Science: Two Debates." *Journal of Educational Statistics* (Special Issue).

Slavin, R. E. 1993. "Ability Grouping in the Middle Grades: Achievement Effects and Alternatives." *Elementary School Journal* 93 (5): 535–552.

Smith, M. 1988. *Thoughts on the Chapter I Longitudinal Evaluation Design*, Briefing to the U.S. Department of Education. Washington, D.C.

Spencer, B. D. 1980. *Benefit-Cost to Allocate Funds*. New York: Springer-Verlag.

St. Pierre, R., Schwartz, J., Murray, S., Deck, D., and Nickel, P. 1993. *National Evaluation of Even Start Family Literacy Program* (Contract 9006–2001). Washington, D.C.: U.S. Department of Education.

Stafford, F. 1985. "Forestalling the Demise of Empirical Economics: The Role of Microdata in Labor Economics Research." In O. Ahsenfelter and R. Layard (Eds.) *Handbook of Labor Economics.* New York: North-Holland.

Stone, E. F., Gardner, D. G., Gueutal, H. G., and McClure, S. 1983. "A Field Experiment Comparing Information Privacy Values, Beliefs, and Attitudes Across Several Types of Organizations." *Journal of Applied Psychology* 68: 459–468.

Taeuber, R., and Rockwell, R. C. 1982. "National Social Data Series: A Compendium of Brief Descriptions." *Review of Public Data Use* 10: 23–111.

U.S. Department of Education, National Center for Education Statistics. 1993 *Data Compendium for the NAEP 1992 Mathematics Assessment of the Nation and States*.

U.S. General Accounting Office. 1986. *Computer Matching: Assessing Its Costs and Benefits* (PEMD-87–2) Washington, D.C.: USGAO.

U.S. General Accounting Office. 1986. *Computer Matching: Factors Influencing the Agency Decision Making Process* (PEMD-87–3 BR) Washington, D.C.: USGAO.

U.S. General Accounting Office. 1992. *Cross-Design Synthesis: A New Strategy for Medical Effectiveness Research* (GAO/PEMD-92–18) Washington, D.C.: USGAO.

U.S. General Accounting Office 1995 *Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies* (GAO/PEMD-95−9) Washington, D.C.: USGAO.

Verdonik, F., and Sherrod, L. R. 1984. *An Inventory of Longitudinal Research on Childhood and Adolescence.* New York: Social Science Research Council.

WebCrawler. 1995. *WebCrawler News*. (HTTP://Webcrawler.com/WebCrawler/Facts/Size.HTML).

Weis, J. G. 1987. "Issues In the Measurement of Criminal Careers." In A. Blumstein, J. Cohen, J. A. Roth, and C. A. Visher (Eds.) *Criminal Careers and "Career Criminals*." Washington, D.C.: National Academy Press, 1986.

Weiss, A. 1988. *Randomized Experiments and Time Series Analysis in Police Research.* Evanston, IL: Department of Political Science, Northwestern University.

White, S. (Ed.). 1994. *Overview of NAEP Assessment Frameworks*. Washington, D.C.: NCES, U.S. Department of Education National Center for Education Statistics.

Wilson, D., Wood, R., and Gibbons, R. D. 1991. *ESTFACT: Test Scoring, Item Statistics,and Item Factor Analysis*. Chicago, IL: Scientific Software.

Zellner, A. 1989. "Discussion." In Mitchell H. Gail and N.L. Johnson (Coords.) *Proceedings ofthe American Statistical Association, Sesquicentennial Invited Papers*. Alexandria: American Statistical Association, 162-166.

Discussant Comments

FREDERICK MOSTELLER

I am extremely impressed with the paper by Bob Boruch and George Terhanian, partly because in several instances they address matters totally new to me. I anticipate that their paper will repay study by the staff of NCES for a long time. I shall comment on only a few of the many issues they treat.

What makes their paper especially effective is the way it appreciates problems of methodology as well as substance and, as Emerson Elliott recommended in his opening speech, how it blurs the distinction between statistics and research and between retrieval and dissemination. Their ability to make connections between different fields and to suggest enterprises that have interest for multiple agencies enhances the opportunities to serve the public by informing Americans about the state of various problems in education. And Boruch and Terhanian have also a beneficial, insightful capacity to see what sorts of activities will engage the attention of an administration, a Congress, or a public proceeding down the information highway.

For example, encouraging people to interpret their own analytic contributions and those of others in order to improve the design of sample surveys is certainly good advice. I had not formerly thought about such a move.

Droitcour and Silberman of the GAO have given us a great challenge in developing the idea of cross-design synthesis. It is especially appropriate to think of its possibilities as a way of using sample surveys to strengthen inferences from experimentation. Their general idea is to let weaknesses from one form of investigation be buttressed by strength from another method, for example, by balancing biases. This good idea needs extensive development.

In order to achieve this goal, we need many investigators to carry out practical examples. From a collection of such examples, we may be able to sieve out principles that can be used in other circumstances. So far, we do not have many examples.

With so few examples of applications, we cannot yet speak of cross-design synthesis as a working method, but when the examples grow in number, we will have a new technique. It will be useful to have NCES encouraging the use of sample surveys to broaden the variety of devices available for cross-design synthesis.

NCES can also develop the ideas mentioned by Boruch and Terhanian that would additionally help link ideas between experiments and surveys. This requires knowing what experiments are being carried out and which ones might be usefully linked to one another by suitable future surveys. For example, can local experimental treatments increase performance in a region? Can surveys measure improvements in a region flowing from local programs of education or of health, such as disease prevention?

Boruch and Terhanian discuss what I like to call "skill grouping," rather than "ability grouping," where classes with different skill levels are put into homogeneous classes rather than heterogeneous ones. Presumably the hope is that students in homogeneous classes will learn more than those in heterogeneous groupings. They present NAEP data that implies that students in homogeneous classes perform better in 8th grade arithmetic than those in heterogeneous classes in most jurisdictions in 1990 and 1992.

In the 10 randomized (or nearly randomized) experiments my colleagues found to review (not restricted to arithmetic), the average performance over different subjects was about equal for the skill-grouped (homogeneous) and the whole-class (heterogeneous) instruction. Although the variation in outcome from study to study was substantial, the reporting was often inadequate. There was also no real way to appreciate whether the students in the studies represented the nation in any reasonable sense. Moreover, no experiment lasted more than 1 year (at least in the most popular form of skill grouping), and each experiment represented only one school.

The contrast between the outcomes in the sample survey and the experiment deserves more explanation. This is an example of an issue whose study might be aided by a compilation by *topic of investigation* of experiments, surveys, and demonstration programs. I do not mean to include analysis, however, which is merely a map of the territory. Most investigators would like their studies included in such a list. Consequently, making such a collection may be feasible. Investigators such as myself would find such information very useful.

The only substantial educational experiment I have come across like this has been the Tennessee Class Size experiment. I have concluded that we need more such experiments.

One might hope that even though schooling is primarily run by the states that some organization could bring together groups of districts regionally or even in a national sample to carry out experiments that would have more than a single state participating. A compendium of surveys, experiments, and demonstrations might help school districts and states think of opportunities to cooperate in such ventures.

In the fourth section of their paper, Boruch and Terhanian discuss work on people who are hard to count and on measurements that are hard to make. They suggest special methods of questioning. With respect to guessing unknown numbers, I have discussed the possibility of trying to estimate the unknown numbers by independently using several different approaches. I call this process "triangulation." To accomplish this, essentially one sets up several different models, and by guessing or knowing parameters of the models, one tries to construct estimates from each model. If the models differ in structure but produce similar outcomes, this seems to give some evidence favoring the resulting estimates. I suggest that adding the idea of several approaches to these difficult measurement and counting problems may help to develop new methods of assessment.

When faced with such a plethora of suggestions as Boruch and Terhanian supply, one is tempted to try to prioritize the list. But as Elliott suggested, much of what will be feasible in the near future will depend on the *accidents* of perceived joint interests of otherwise independent organizations, and so trying to prioritize these suggestions would not be very profitable, as compared with having it done by someone who is more familiar with the current goals of the Center

and its interactions with other organizations. It would be valuable to have individuals at the Center who are well prepared to work cooperatively, and this paper and others presented here certainly are making major contributions to that end.

5 New Data Collection Methodologies, Part II: Experimental Design

THIS PAGE INTENTIONALLY LEFT BLANK

Incorporating Experimental Designs Into New NCES Data Collection Methodologies

Charles E. Metcalf

ABSTRACT

This paper considers some potential methods of accommodating policy evaluations using a formal experimental design—that is, with randomized treatment and control groups—within NCES national data collection efforts. The paper first addresses some limitations in using national data sets for selecting comparison groups for policy evaluations, and then explores the following approaches to integrating experimental designs into ongoing longitudinal databases:

- Designing a specific experiment for implementation at the initiation of the longitudinal survey, using a within-survey treatment group that receives the policy intervention;
- Designing a longitudinal survey to accommodate as-yet-unspecified future experiments;
- Augmenting a survey with supplemental sampling units that receive an experimental intervention, or expanding the longitudinal sample to incorporate separately defined demonstration treatment and control groups for common data collection efforts; and
- Providing a sample frame for the random selection of schools to test school-based innovations.

The paper draws the following five conclusions:

1) Because the descriptive value of NCES national data sets for framing policy issues ought not to be minimized, precautions should be taken so that efforts to accommodate policy experiments do not dilute this value.

2) Attempts to improve the attractiveness of national data sets as general-purpose (nonrandomized) comparison groups would not be warranted, because the intrinsic weakness of comparison groups relative to randomized control groups makes this effort an unpromising investment.

3) Experiments are difficult to incorporate into a national longitudinal sample, unless the timing of a demonstration implementation converges fortuitously with the initiation of a longitudinal survey that has a compatible age cohort. There is potential for improving the efficiency and comparability of longitudinal data collection, however, and for moving such experiments in the direction of using representative sample frames compatible with national data frames.

4) Attempts to append an experiment to a longitudinal survey after the survey's initiation point would be fraught with difficulties, unless supplemental samples are drawn for the demonstration treatment and control groups.

5) The potential for implementing demonstrations with across-school random assignment appears to have been severely underestimated, both for student- and school-targeted initiatives.

BACKGROUND ON THE USE OF EXPERIMENTAL DESIGNS

Since the first income maintenance experiments in the late 1960s, experimental methods that involve the random assignment of a target population to treatment and control groups have proved to be both feasible and extremely valuable for evaluating social programs and policy interventions. This approach has been established as the most defensible method for determining the extent to which *specific* policy interventions affect behavior or outcomes of interest.

Randomized experiments have been used to test interventions in such areas as welfare reform, employment and training, food stamp benefit cashout, health care delivery, long-term care, medical treatment, offender rehabilitation, domestic violence, and family preservation services. Evaluations of the Upward Bound program (funded by the U.S. Department of Education [ED]) and the Job Corps program (funded by the U.S. Department of Labor [DOL]) have broken new ground in measuring the impacts of *existing* broad-based programs by diverting nationally representative samples of program applicants into randomized control groups.[1] In addition to Upward Bound, ED has funded other recent randomized studies, including evaluations of the Dropout Demonstration Assistance Program, Dropout Prevention and Reentry projects in vocational education, the Even Start program, and Workplace Literacy programs.

While program evaluation methodology was evolving, the National Center for Education Statistics (NCES) initiated a series of longitudinal studies "to provide ongoing, descriptive information about what is occurring at the various levels of education and the major transition phases of students' lives," beginning with the National Longitudinal Study of 1972 (NLS–72) (NCES 1995). Similarly, other large-scale data sets, such as the Survey of Income and Program Participation (SIPP), the Panel Study of Income Dynamics (PSID), and the National Longitudinal Survey–Youth Cohort (NLSY) track representative samples of some of the same populations targeted by programs subject to demonstration evaluations.[2]

Yet almost invariably, program evaluations based on control group or nonrandomized comparison group methodologies have involved independent data collection efforts, usually using samples and demonstrations sites that are not nationally representative.[3] These evaluations do not take advantage of the existing array of continuing large-scale data collection efforts that might include representative samples of the potentially relevant target population of interest, except possibly for limited benchmark purposes.

Why is this? Are there deficiencies in national data sets that can be remedied, from a policy impact evaluation perspective, without compromising the primary focus of these data sets? If

program evaluations could use general-purpose databases effectively, this seemingly inefficient use of data collection resources could be rectified.

In judging the efficacy of national data sets for the evaluation of education policy, I should stress that the NCES national data sets are used for both descriptive and evaluation purposes. They are available to a wide variety of potential users as data sets for evaluating both education policy and the dynamics of educational processes and student behavior. Accurate, representative descriptions are essential for understanding an economic or policy sector: providing incontrovertible evidence of what is happening is a legitimate and primary focus of national data collection efforts that should not be compromised. In my experience, some of the greatest revelations of research projects have involved description and documentation of facts that turned out to be controversial, rather than sophisticated evaluation of policy demonstrations or experiments.

This paper considers some potential methods of accommodating policy evaluations using a formal experimental design—that is, with randomized treatment and control groups—within NCES national data collection efforts. Examples of these methods might include the following:

- Designing a specific experiment for implementation at the initiation of the longitudinal survey, using a within-survey treatment group that receives the policy intervention;

- Designing a longitudinal survey to accommodate as-yet-unspecified experiments in the future;

- Augmenting a survey with supplemental sampling units that receive an experimental intervention, or expanding the longitudinal sample to incorporate separately defined demonstration treatment and control groups for common data collection efforts; and

- Providing a sample frame for the random selection of schools to test school-based innovations.[4]

The next section of this paper addresses some limitations in using national data sets for selecting comparison groups for policy evaluations. The third section explores methods for adapting national data sets to accommodate formal policy experiments. The paper concludes with a brief reality assessment of approaches showing the most promise.

Using National Data Sets to Select Comparison Groups for Policy Evaluations

Longitudinal and repeated cross-sectional data sets permit many insightful analyses of causal relationships and policy impacts, but their use falls short of conventional experimental standards for measuring program impacts. They are often proposed, and sometimes used, to create comparison groups for demonstrations of a policy implemented with a separate sample of students and/or schools. But repeatedly these data sets are rejected in favor of independently collected data sets for control or comparison groups.[5]

When considering the use of an existing data set as a comparison group, designers of demonstrations and policy evaluations are confronted with a major cost advantage over the use of

an independent control or comparison group and its associated data collection costs. They are also confronted with two major classes of disadvantages from a methodological standpoint. These disadvantages are associated with 1) characteristics of specific data sets relative to those of an independently tailored comparison group, and 2) general deficiencies of comparison groups relative to randomly selected control groups.

Criteria for Evaluating Existing Data Sets as Comparison Groups

Aside from general problems associated with nonrandomized comparison groups, an existing data collection vehicle would have to meet several basic requirements to be a suitable substitute for an independently defined comparison group:

- The sample must contain an identifiable subgroup that is comparable to the group receiving the demonstration treatment;
- The subsample meeting target group requirements must be large enough to meet the statistical precision requirements of the planned evaluation;
- The survey should have a longitudinal structure for tracking individual outcomes for a period comparable in length to that used for tracking the demonstration treatment group, ideally for the same period in chronological time;[6] and
- The survey database must contain comparable data elements, both for measuring background characteristics of sample members and for defining outcome measures.

To provide a concrete example of how these criteria were applied, the following describes the process by which existing longitudinal surveys were considered for use as a comparison group for the Job Corps evaluation that is currently under way. I chose this example because of my firsthand involvement in the design effort, even though Job Corps is funded by DOL rather than by ED.

The Job Corps program provides a range of education, vocational training, and support services in a predominantly residential setting to disadvantaged youths between the ages of 16 and 24.[7] Approximately 60,000 new enrollees are served each year. In 1993, DOL initiated an evaluation of the program that eventually adopted a randomized design in which approximately 8 percent of all eligible Job Corps applicants were assigned to a control group. Sample intake began in November 1994 and is scheduled to end in early 1996.

Before adopting a randomized design, we considered using an independently constructed comparison group (not discussed here) and several existing surveys—the National Education Longitudinal Study (NELS), SIPP, PSID, NLSY, and the Current Population Survey (CPS). To fulfill the criteria discussed earlier for the requirements of the Job Corps evaluation, an existing survey would have to provide a comparison group with the following characteristics:

- A representative sample of youths aged 16 to 24 in 1995, who meet specific definitions of being disadvantaged and having limited employment opportunities;

- A longitudinal structure providing outcome data for 36 to 48 months after the 1995 enrollment window; and

- Outcome measures of employment and education experience, transfer receipts, and criminal activities.[8]

None of the considered data sets could have identified eligible youths in a manner strictly comparable to the criteria applied in the Job Corps recruitment process, but all could have provided acceptable approximations of the relevant population. The NELS sample, which started as a cohort of 1988 8th graders, would have provided a sample of about 2,000 Job Corps eligibles aged 20 to 21 in 1994, but it would not have covered the full age span of eligibles. The planned 1998 NELS survey would have provided detailed education and training outcome measures 36 months after the enrollment window for the Job Corps sample, but incomplete information on labor market experience and no information on criminal activities or transfer receipts. Finally, the baseline data would have been defined for 1994 (1 year before the data collected for the treatment sample) for a sample that had already experienced 6 years of attrition, thus threatening the representativeness of the sample.

The other data sets under consideration also had disadvantages sufficient for their disqualification. The SIPP and CPS data sets included the full age span of Job Corps eligibles, but the sample sizes were inadequate (fewer than 1,000 each). Furthermore, these data sets provided no longitudinal data for 36 months or later and no information on criminal activities.[9] The PSID also included a sample of fewer than 1,000 eligible youths and provided only limited information on those who were not heads of households. Finally, the NLSY provided detailed information on a cohort of 4,000–5,000 youths. Unfortunately, these youths were aged 14 to 21 in 1978 and would have been a promising comparison group for Job Corps applicants in 1981: by 1995, however, they were aged 31 to 38.

Other difficulties with using existing surveys for comparison groups are worthy of mention. These difficulties relate to demonstration treatment samples that are not nationally representative or that measure outcomes in idiosyncratic ways (which may reflect limitations of the demonstration rather than the potential comparison sample):

- In recent years, the number of state-based policy evaluations, particularly in the area of welfare reform, has been increasing; similarly, a demonstration of a school reform initiative might be concentrated in one or a small number of states. Existing national databases may lack a large enough sample in the states of interest; furthermore, some data sets may not provide state identifiers in their public use data files.[10]

- Many demonstrations take place in a judgmental (that is, not randomly selected) sample of sites that may not be representative of the national target population for a policy initiative. These demonstrations must confront a methodological tension between identifying a comparison group that is as similar as possible to the treatment population (to promote internal validity of the results) and extrapolating findings to a national target population. To the extent that both variants of a comparison population can be identified in a national data set and their differences measured, use of a national database as a comparison reference, rather than an independent but nonrepresentative comparison

group, could enhance our ability to draw policy implications from demonstrations not conducted with a representative sample.

- Designers of demonstration evaluations often complain that national data sets do not measure potential outcome variables in a manner appropriate for assessing policy impacts of interest. This criticism cuts both ways, however. To the extent possible, program evaluators should attempt to express their findings in terms of broadly available outcome measures in order to promote the interpretability of their results. On the other hand, although certain types of information involving such subjects as criminal activity, drug use, or sexual activity may be inappropriate for broad-based longitudinal data sets, designers of survey instruments for future longitudinal data sets should attempt to incorporate the information required to construct variables that are widely usable as outcome measures for policy evaluations.

The discussion here about the deficiencies of specific data sets relative to an independently tailored comparison group may be moot in an evaluation that rejects a "well-constructed" comparison group in favor of a randomized control group. Most of these issues will remain relevant, however, when discussing the possibility of defining future longitudinal data sets that incorporate or can accommodate a formal experimental design.

Deficiencies of Comparison Groups Relative to Randomly Selected Control Groups[11]

The classical statistical methodology underlying randomized experiments requires that we compare two independent random samples—one that receives the intervention of interest—drawn from the same population. When this condition is met, simple statistical tests reveal the likelihood that any observed differences could be due to chance rather than to systematic differences created by the intervention.

Random assignment fulfills this condition proactively, if neither the sample selection and randomization process nor the method of introducing the intervention creates contaminating effects that could be confused with the intervention's impact. Comparison group methods, on the other hand, use assumptions, measurement of other sources of differences, and statistical models to eliminate differences that could derive from reasons other than the intervention. If these efforts are successful, a residual difference can be identified as resulting from the intervention, perhaps with some measure of statistical confidence.

Continuing debate about whether nonexperimental comparison groups can be used to provide convincing measures of program impacts has been fueled by a number of studies comparing impacts estimates based on control and comparison groups.[12] The debate has also been advanced by an increasingly rich econometric literature about methods to deal with the problem of "selection bias," which results from sources of unmeasured or unmeasurable differences between treatment and comparison groups.[13]

Successful use of nonrandomized comparison groups requires that we be able to measure and control for all systematic differences (other than the intervention) between the samples. Even if all differences can be measured and controlled for, we must keep in mind that the correction process

"uses up" statistical power that is no longer available for testing the intervention's primary impact. Time and time again, statistical tests appropriate for randomized experiments are misapplied to nonrandomized comparison groups, with a resulting vast overstatement of the strength of the results.

Similar problems exist with the statistical methods available to test for the presence of and correct for selection bias. Tests for selection bias produce three possible outcomes: 1) bias is present, but we lack an acceptable method to correct for it or perhaps even to detect it; 2) bias is present, and available methods permit us to correct for it; and 3) no systematic bias appears to exist. Each of these outcomes poses problems:

- In the first case, internally valid estimates of impacts cannot be obtained, and the researcher must seek alternative data sets. This is a useful result for researchers evaluating alternative secondary data sets, but scarce comfort for those who have just completed a demonstration with a primary data collection effort.

- In the second case, increasingly sophisticated statistical methods have been developed to correct for the source of bias. However, they typically require the availability of measures for both the treatment and comparison groups that are correlated with program participation but not with program impacts, and tend to produce unstable, nondefinitive results. Even when successful, they absorb statistical power in the correction process and often produce standard errors of impact estimates that are approximately *three times* those produced with demonstrations using control groups. When this happens, sample sizes for a comparison group design have to be as much as *nine times* larger to measure program impacts with the same statistical precision as with a properly designed randomized experiment.

- Only in the last case can we proceed with no statistical correction for bias. Again, however, using the full sample as if random assignment had occurred implies not only that "we have failed to detect evidence of selection bias," but also that "we know with certainty that it is absent."

In any event, we would not know which case applies until a demonstration has been completed and the data have been collected.

The current array of methods available to measure program impacts with nonexperimental data are extremely valuable when time, resources, or other circumstances prevent the design and execution of a randomized experimental design for testing a new policy intervention or an existing program. They are also important for helping to counteract the inevitable imperfections in formal experiments implemented in actual demonstration or program environments.[14] Yet, nonrandom comparison groups—whether "made to order" or drawn from currently available or future longitudinal data sets—are unlikely to return as the methodology of choice for major impact evaluations that place priority on obtaining convincing results. Thus, future longitudinal data sets are unlikely to play a prominent role in impact evaluations unless they can be adapted to accommodate a formal experimental structure for program evaluation purposes. The next section looks at this topic.

Adapting National Data Sets to Accommodate Formal Policy Experiments

If national data sets can be adapted to accommodate formal policy experiments, they could contribute a vital element commonly absent from such experiments: a nationally representative context in which to test a policy.

Internal validity and external validity are two concepts central to sound evaluation design. Internal validity addresses whether what we observe is in fact caused by an intervention. External validity involves whether observed demonstration impacts would be replicated if implemented in broader settings and/or on a larger scale. Although both concepts are crucial for policymakers, it is in the realm of internal validity where well-designed randomized experiments have established their clear superiority over comparison group methodologies. Experiments as typically implemented fall short of standards of external validity, leaving the analyst to engage in nonexperimental, often judgmental methods to establish policy relevance.

An implicit but major controversy in the evaluation community exists between those who focus on establishing an internally valid experimental setting—often by creating an artificial program in an analytically precise environment in one or a small number of nonrepresentative sites—and those who are willing to sacrifice "design rigor" for evaluating a program in a more representative setting. Frequently, researchers face the tension between asking the right question with a weak methodology and asking the wrong question with a sound methodology.

Only recently have there been any significant attempts to place randomized designs in a nationally representative operational setting. The Upward Bound and Job Corps evaluations are prominent examples of these efforts. By providing a national context—or a well-defined target group, such as inner-city students or rural schools—future national databases may provide a vehicle for implementing policy experiments with a presumptive claim of external as well as internal validity for evaluation results.

In the introductory section, I suggested ways in which experimental designs might be integrated into ongoing longitudinal databases: 1) implementing a specific experiment with the initiation of a longitudinal survey; 2) designing a longitudinal survey to accommodate one or more as-yet-unspecified future experiments; 3) augmenting a survey with supplemental sampling units that will receive an experimental intervention, or expanding a longitudinal sample to incorporate separately defined demonstration treatment and control groups for common data collection efforts; and 4) providing a sample frame for the random selection of schools for testing school-based innovations. This section provides examples to illustrate the potential and the drawbacks of each of these approaches.

Implementing a Specific Experiment With the Initiation of a Longitudinal Survey

Suppose we wish to test a new approach to enhance reading skills, beginning in the 8th grade for students in inner city schools, and that we are prepared to implement a test of this method that coincides with the initiation of a new NELS-type survey—that is, a longitudinal survey of a random sample of 8th-grade students drawn from a first-stage representative sample of perhaps 1,000 schools. Combining these two initiatives could provide four distinct advantages to the evaluation:

1) With a random sample of students from inner city schools from the NELS frame, the evaluation results could be interpreted directly in terms of the target population (external validity);

2) With common data collected from both students in the demonstration schools and the full sample, the performance of targeted students could be compared with that of their designated control group and that of all students nationwide;

3) With continued tracking of the sample on a longitudinal basis, long-term impacts of the demonstration could be measured beyond the initial evaluation effort; and

4) The incremental cost of the demonstration is likely to be lower than that of a stand-alone study.

The experiment could take one of two general forms—*within-school* versus *across-school* random assignment—each with distinct methodological and operational implications. A demonstration with within-school random assignment of students to treatment and control groups (or more broadly, within-site randomized demonstrations) is the most common design for a policy experiment. A less frequently observed design—but very promising in many contexts in my judgment—involves the random selection of treatment and control *schools*, with all eligible students in the respective groups of schools constituting the treatment and control samples of students.

Demonstrations using within-school random assignment require that the scale of the program intervention in each site be smaller than the potentially eligible population. They also require that the nature of the intervention be such that none of its benefits "spills over" onto the control group, such as when instructional methods for the control group are affected by what teachers learn from the demonstration, or when innovations or reforms are schoolwide in their potential impact.

Within-school designs also require overcoming school resistance to denying program services to some eligible students on a random basis. This resistance increases if there is a risk that some program slots may remain vacant because some applicants are diverted to a control group. The Upward Bound demonstration dealt with this problem by assigning some of the control group (on a random basis) to a waiting list, from which students could be selected to fill vacant slots.

An advantage of using across-school random assignment is that treatment-group schools would not have to deal with the mechanics of random assignment. Control schools would be treated like all others in the longitudinal sample, except to the extent that specialized data collection or an increased sample of students is required.[15]

Innovations tested with this approach must be applied either to the *entire* eligible student population, however, or to subsets of the population identified by student characteristics that can be readily measured in data collected for students in the control schools. Interventions targeted at a small number of volunteer applicants from a larger, nominally eligible group—such as Upward Bound—would not be well suited for this approach, because attempts to identify the comparable

group of students in the control schools would suffer from the same selection bias problems that plague nonrandomized comparison groups.[16] Furthermore, the across-school approach could not be used for evaluating existing programs, which are likely to be present already in the control schools.

Finally, in situations for which either design would be methodologically appropriate, across-school designs would typically require larger sample sizes than within-school designs for equal statistical precision, because both individual and school characteristics would vary randomly between the treatment and control groups. Within-school random assignment, on the other hand, eliminates variations in school characteristics between the treatment and control group.[17]

The issues discussed here involving the choice between within-school and across-school randomized designs are relevant whether or not a demonstration is integrated into a longitudinal survey. Special problems to be considered when integrating either approach into a longitudinal survey include the following:

- We must have identified the experiment of interest in time for implementation at the beginning of the longitudinal survey. *More importantly, the target age cohort for the experiment must coincide with a cohort included in the survey*. If the survey is tracking a cohort of 8th graders, for example, a demonstration targeting that group could be included, but not one focusing on 10th graders.

- Planners of demonstrations typically solicit applications from schools or sites willing to participate. The strategy discussed here requires approaching a random sample of schools and inviting them to participate. This approach is feasible only if the offer of participation is sufficiently attractive to achieve high participation rates.

- The number of students per school in an NELS-type survey is unlikely to be large enough to support the sample-size requirements of a demonstration. Thus, the sample of students would have to be augmented in the treatment schools and probably in the control schools (in the across-school design) as well.

- The content of the longitudinal survey may have to be modified to ensure that it includes appropriate outcome measures for evaluating the long-term impact of the intervention. In addition, supplemental data collection may be required for the demonstration (for example, if achievement test scores are desired).

- Students in the treatment sample, *by virtue of their receipt of the program innovation*, would no longer be representative of their cohort. Thus, the size of the longitudinal sample, excluding the demonstration sample, would have to be large enough to serve the general purposes of the longitudinal survey.

Designing a Longitudinal Survey to Accommodate Future Experiments

Simultaneous initiation of a randomized demonstration and a longitudinal survey requires that an uncomfortably large number of planets be in proper alignment. The increased flexibility of a longitudinal survey that could accommodate one or more experiments *after* its initiation would be desirable. For example, we may want to test an initiative targeted at 10th-grade students 2 years after

the beginning of the longitudinal survey, or we may not yet have settled on a policy initiative worthy of experimentation.

What characteristics must the survey sample have to offer this flexibility, and what special problems would have to be resolved in designing subsequent evaluations? For the purposes of this discussion, assume that the longitudinal survey would track a cohort of 8th-grade students, with follow-up interviews scheduled every 5 years.

Several general issues would have to receive special attention in the design of the longitudinal sample to accommodate future experiments, some of them already identified. First, the questionnaire content might have to be examined in terms of its measurement of student characteristics and outcome variables likely to be important for evaluating future policy demonstrations. If the demonstrations require supplemental data collection, especially on a continuing longitudinal basis, much of the advantage of attempting to integrate the demonstration with the survey would be vitiated.

Second, we would have to consider the available sample sizes for all potential evaluation target groups, both for potential demonstrations and for the remaining sample available for general users of the longitudinal database. Realistically, most strategies for appending a demonstration would involve adding supplemental samples (both school sampling units and students within schools) to the survey at the time the demonstration is implemented.

Third, a survey like NELS, restricted to a single-grade cohort, would be particularly restrictive in terms of the future timing and range of potential demonstrations. For example, a longitudinal sample of 8th graders could be integrated with a policy initiative directed at high school sophomores after 2 years, but not at any other time. A survey with more than one cohort would be more flexible in terms of its potential accommodation of future demonstrations.

Finally, if inclusion in the longitudinal sample places schools or students “at risk” of inclusion in a future demonstration, there may be issues of informed consent to consider. (This is more likely to be a problem for demonstrations calling for within-school random assignment than for the across-school approach, if responding to an interview increases an individual's exposure to future selection for participation in an experiment.) Such consent, if required, could lower response rates in the longitudinal survey, a problem that compounds in subsequent waves of the survey. Again, this problem is mitigated if we think in terms of supplemental samples for demonstration implementation.

Returning to the example of testing a policy initiative targeted at high school sophomores 2 years after a longitudinal survey of 8th graders has been initiated, the designer of the demonstration would face several obstacles:

- Timing is everything, as already suggested. Two years after the initiation of a longitudinal survey focused exclusively on an 8th-grade cohort is the *only* time a demonstration targeted at 10th graders could be implemented.

- In this example, baseline data are 2 years old and would not exist for any augmented sample required for the demonstration. If baseline data in addition to student records are required, a supplemental baseline survey would have to be implemented.[18]

- During a 2–year period, students in the longitudinal sample may have dispersed to different high schools in their districts, moved out of the area, dropped out of school, or otherwise disappeared from the sample. Sample students remaining in the same school districts would not be representative of all students in those districts, because students who changed districts in the past 2 years would be excluded from the sample. These factors would severely complicate attempts to implement a demonstration using students already included in the sample, even if sample sizes available for the demonstration were adequate.

- If a supplemental sample is drawn for a demonstration treatment group, the above complications could compromise the suitability of using the regular longitudinal sample in a selected set of schools as a control group. A potential solution to this problem might include adding a supplemental sample of control students who arrived in the sampled schools since the definition of the longitudinal frame.

- Demonstrations that combine the "new" and "original" sample would have to deal with potential differential sample attrition over time, resulting from the different "longitudinal ages" of the two portions of the sample.

These considerations are likely to make separately drawn treatment and control samples more attractive to program evaluators than designs relying heavily on the "original" sample from a previously initiated longitudinal sample. The question is whether these samples should retain a structural link to the longitudinal survey, or whether the current practice of implementing randomized demonstrations independently of national longitudinal samples should continue.

Augmenting a Survey With Supplemental Sampling Units to Receive an Experimental Intervention, or Expanding a Longitudinal Sample to Incorporate Separately Defined Demonstration Treatment and Control Groups for Common Data Collection Efforts

The discussion here has implicitly moved us in the direction of a more limited integration of randomized demonstrations with longitudinal data sets. Three possibilities come to mind:

1) Augmenting the longitudinal sample with supplemental sampling units—*selected in the same manner as schools forming the basis of the longitudinal survey*—to receive the program intervention being evaluated, and using the longitudinal sample as a control group;

2) Choosing supplemental sampling units (in the same manner) for *both* the treatment and control groups, but integrating the demonstration into the longitudinal data collection sample; and

3) Defining a demonstration sample by procedures not related to the longitudinal frame, as is currently done, and limiting the link to common longitudinal data collection.

The first approach forces the treatment sample to be nationally representative of the target group in question, a major advantage over most contemporaneous randomized demonstrations. This approach would also be viable as a variant of the first scenario described in this section, in which a demonstration is implemented at the same time the longitudinal survey is initiated. When the demonstration is initiated *after* the longitudinal baseline, however, a number of issues related to the comparability of the treatment and control groups (discussed earlier) could compromise the experiment's validity.

The second approach would utilize supplemental, representative sampling units for both the treatment and control groups. Although not using the longitudinal sample in a literal sense, this approach combines the advantages of providing a nationally representative test of the program intervention on a sample defined in the same manner as that used to track students nationally, with the economic advantages and the interpretative consistency of commonly collected data for demonstration participants and the general student population. This approach could strengthen program evaluation methodology significantly and is an option worth pursuing where feasible.

The third option leaves demonstration designers free to define independent treatment and control groups, while retaining the advantages of common data collection efforts. This approach may be an improvement over current practice, but I would find it to have a rather disappointing outcome: guiding randomized demonstrations in the direction of nationally representative rather than pragmatic implementation venues, which would be achieved by the previous option, is an important priority for the evolution of program evaluations.

Providing a Sample Frame for the Random Selection of Schools for Testing School-Based Innovations

All the design options discussed here have been "school-based" but focused on measuring outcomes through longitudinal data collection efforts patterned after NELS, with schools serving as the primary sampling unit for selecting students and as the venue for implementing the demonstrations. The tested policies were viewed as affecting specific students enrolled in the demonstrations, rather than as broader school reforms that might have schoolwide impacts.

Here, the discussion expands to include experiments in which the school is the target of the innovation, and the design is clearly across-school in character. Measured outcomes might take the form of longitudinal observations of students, as before, or repeated outcome measures for successive cohorts of students in a longitudinal sample of schools. In the latter case, measured outcomes could be based on administrative records, test scores, or aggregate measures for each school, as well as student interviews.

The design objective here is to use existing survey sample frames to select random samples of schools for testing a reform or innovation in a formal experiment, rather than to follow the more traditional approach of comparing judgmental treatment and comparison samples of schools.

For example, suppose we wish to test the effect on mathematics achievement or other outcomes of making personal computers readily available to students in rural schools.[19] In order to implement such a demonstration, we would select a random sample of rural schools from the NCES Schools and Staffing Survey (SASS) frame (or augment the sample if there are not enough schools) and invite these schools to participate in the demonstration.[20] Rural schools not selected for the offer of participation would constitute the control group and could be augmented by an additional sample of schools, if necessary. Outcomes could be measured with supplemental data collection efforts in conjunction with future waves of SASS.

In order for experiments of this sort to be effective, certain conditions would have to be met:

- The SASS design would have to be modified to be more longitudinal in character. (It is my understanding that such a modification is under consideration.) Furthermore, it would be desirable to investigate the possibility of adding summary outcome measures relevant for a range of potential school innovations and reforms to limit the extent of supplemental data collection efforts.

- As noted, the tested initiatives would have to be attractive enough that a large proportion of the selected schools would agree to participate, because the treatment group would be properly defined as all who are offered participation (not just participants). Furthermore, nonparticipants would dilute the power of the experiment.

- The potential impacts of the intervention would have to be schoolwide or serve a high fraction of identifiably eligible students. The impacts would also have to be measurable in tangible terms that could be measured consistently across schools.

- If all relevant output measures could be obtained from standard survey data, there would be no need to obtain any special consent from the control schools. Agreement of control schools to participate in supplemental data collection efforts would have to be solicited, but they would not have to be involved in the demonstration in any other material way.

- If the initiative is widely publicized, inexpensive, and easy to implement, there is the risk that control schools will implement a similar program on their own. If this happens too quickly, the outside world will "catch up" to the innovation before its impacts can be measured. The demonstration is more likely to be successful in measuring impacts if the innovation requires significant resources and/or technical assistance to implement, and if premature publicity surrounding the demonstration is kept to a minimum.

SASS may be less promising (in terms of its structure and traditional content) than longitudinal student samples as a vehicle for collecting required outcome measures, increasing the likelihood that specialized data collection efforts would have to be implemented in conjunction with randomized demonstrations. Even if supplemental data collection is required, however, I place high priority on the possibility of executing randomized tests of school-based interventions within the standardized framework that a nationally representative database can provide.

CONCLUSIONS: A REALITY ASSESSMENT

In the discussion here, I reviewed a range of reasons why demonstrations and policy evaluations have not made significant use of existing national data sets, and considered a number of ways in which these data sets might be adapted to alter the conduct of future evaluations. My conclusions are as follows:

- The descriptive value of well-structured national data sets for framing policy issues ought not to be minimized, and precautions should be taken so that efforts to accommodate policy experiments do not dilute this value.

- Attempting to improve the attractiveness of national data sets as general-purpose (nonrandomized) comparison groups would not be, in my opinion, a noble objective. The intrinsic weakness of comparison groups relative to randomized control groups makes this effort an unpromising investment.

- Experiments with within-site treatment and control groups are difficult to incorporate into a national longitudinal sample, unless timing of a demonstration implementation converges fortuitously with initiating a longitudinal survey that has a compatible age cohort. There is potential for improving the efficiency and comparability of longitudinal data collection, however, and for moving such experiments in the direction of using representative sample frames compatible with national data frames.

- Attempts to append an experiment to a longitudinal survey after the survey's initiation point would be fraught with difficulties, unless supplemental samples are drawn for the demonstration treatment and control groups.

- The potential for implementing demonstrations with across-school random assignment appears to have been severely underestimated, both for student- and school-targeted initiatives. Coordinating the design of such evaluations with the representative frames of national surveys and engaging in integrated data collection activities, where possible, could produce significant improvements in both the methodology and the efficiency of future policy experiments.

NOTES

1. The evaluation strategies for both Upward Bound and Job Corps depended on pools of potential eligibles that exceeded the available number of program slots.

2. SIPP utilizes an overlapping panel design rather than a strict longitudinal design; the Current Population Survey, another widely used continuing data set, utilizes overlapping panels of household locations. The NCES Schools and Staffing Survey, which will be considered later in this paper along with the NCES longitudinal studies for adaptation to experimental evaluations, utilizes repeated cross-sections with an approximate 30 percent overlap of schools between successive interview waves.

3. This paper defines a "control group" as a sample selected through random assignment between treatment and control students or schools, and a "comparison group" as one chosen to be as similar as possible to a treatment group, but without random assignment.

4. An additional potential focus for incorporating experiments into NCES national data collection efforts—not the subject of this paper—would involve testing alternative data collection methodologies. Designing these experiments would involve substantive issues related to the data collection methodologies being tested, but the sampling and experimental design issues would be relatively straightforward.

5. For example, the initial design for the ED's evaluation (conducted by Mathematica Policy Research) of the tech-prep educational program called for using data from the National Education Longitudinal Study as a comparison group, but this approach was abandoned after critical examination by both Office of Management and Budget and project staff. NELS was one of several longitudinal data sets considered for creating a comparison group for DOL's Job Corps evaluation before a randomized design was chosen as a superior approach.

6. School-based interventions would require longitudinal samples of *schools*, but associated samples of students might appropriately be repeated cross-sections, depending on the evaluation.

7. The upper age limit was increased from 21 to 24 in 1993.

8. A previous evaluation of Job Corps completed in 1982 (Thornton et al. 1982) identified a reduction in criminal activities as a prominent benefit of the program.

9. The rotation pattern of the SIPP panels provided longitudinal data for 30 months or less; the CPS utilizes rotating panels based on household location and provides no actual longitudinal data.

10. NCES usually maintains both public use and restricted use data files. The restricted use files may permit identification of states and other locations.

11. Portions of the following discussion are adapted from Metcalf and Thornton (1991).

12. See Ashenfelter and Card (1985), Lalonde (1986), Lalonde and Maynard (1987), and Fraker and Maynard (1987).

13. For example, see Maddala and Lee (1976), Heckman (1979), and Heckman and Hotz (1989). For discussions of effective use of tests and corrections for selection bias in nonexperimental data, see Heckman and Robb (1985) and Heckman et al. (1987).

14. For example, the potential presence of selection bias must be dealt with when 1) fewer than 100 percent of the individuals selected for a treatment group choose to participate in a program; 2) separate impact estimates are desired for different program elements provided to nonrandom subsets of the treatment population; and 3) longitudinal data collection efforts produce differential attrition rates for the treatment and control groups.

15. In the limiting case, all schools in the longitudinal sample with the characteristics of the treatment schools—in this example all inner-city schools—would be part of the control group by virtue of their inclusion in the longitudinal sample. The control group schools have no special knowledge about the existence of the demonstration in the treatment schools.

16. Interventions that require voluntary enrollment could be tested if the participation rate is high (for example, 70 percent or greater). The defined treatment and control groups, however, would include all eligibles, inclusive of nonparticipants. The presence of nonparticipants would dilute the precision of measured impacts by a factor proportional to $(1/P^2)$, where P is the participation rate.

17. By eliminating this major component of variance, within-school random assignment improves the statistical precision of internally valid estimated impacts for the demonstration schools. Extrapolations to national estimates would still have to account for design effects due to the clustering of the student sample into a small number of schools, but a within-school design would retain its statistical advantage for extrapolations as well.

18. In principle, baseline data are not required for comparing treatment and control outcome data in a properly constructed experiment. Baseline data, however, can be used to reduce the variance of impact measures by controlling for student characteristics, and can be invaluable for interpreting future problems of sample attrition. Student records might serve some of this purpose, if informed consent issues can be resolved.

19. Alan Hershey of Mathematica Policy Research suggested this example. Recently it has come to my attention that a similar program already exists.

20. Alternatively, one might choose schools serving as primary sampling units in a longitudinal student survey, but this approach would provide relevant student data only if the demonstration were implemented at the initiation of the longitudinal survey, and only if the survey included a broad enough age span of students to encompass the intended target of the reform.

REFERENCES

Ashenfelter, O. and Card, D. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67 (4).

Fraker, T. and Maynard, R. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources* 22 (2).

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47.

Heckman, J. J. and Robb, R. 1985. "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data.* Eds. James J. Heckman and Burton Singer. Cambridge, MA: Cambridge University Press. 1985.

Heckman, J. J., Hotz, V. J., and Dabos, M. 1987. "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?" *Evaluation Review* 11.

Heckman, J. J. and Hotz, V. J. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84 (408).

Lalonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76.

Lalonde, R. J., and Maynard, R. 1987. "How Precise are Evaluations of Employment and Training Programs: Evidence from a Field Experiment." *Evaluation Review* 11 (4).

Maddala, G. S. and Lee, L. 1976. "Recursive Models with Qualitative Endogenous Variables." *Annals of Economic and Social Measurement* 5.

Metcalf, C. E. and Thornton, C. 1992. "Random Assignment." *Children and Youth Services Review* 14(½): 145–156.

Thornton, C., Long, D., and Mallar, C. October 1982. "A Comparative Evaluation of the Benefits and Costs of Job Corps After Forty-Eight Months of Post Program Observation." Princeton, NJ: Mathematica Policy Research, Inc.

U.S. Department of Education, Office of Educational Research and Improvement. January 1995. *Programs and Plans of the National Center for Education Statistics, 1995 Edition.* Washington, D.C.: U.S. Department of Education.

Discussant Comments

DONALD B. RUBIN

I congratulate Chuck Metcalf for writing a clear and direct article advocating the increased use of randomized experiments in educational research, a point with which I fully agree. He is also to be congratulated for providing a list of good recommendations on how to conduct such studies (e.g., by imbedding them in longitudinal studies and doing treatment assignment at an appropriate level to avoid issues of interfering units). It is especially rewarding to see a distinguished, practically experienced economist strongly eschew the naive application of simple OLS models, structural equations methods, and instrumental variables techniques that have been advocated by many in economics (e.g., Heckman 1979, and other more recent references cited by Metcalf).

I am particularly interested in his citation of LaLonde (1986) to support his advocacy of randomized experiments because that article has become a focal point in a course on "Causal Inference," which I have been teaching with Professor Guido Imbens in the Department of Economics at Harvard University. Specifically, the LaLonde article shows that the standard techniques typically used by statisticians and economists with nonrandomized data cannot be trusted to provide the "correct" answer, where correct is defined by the answer provided in a randomized experiment. In this study, the treated group, consisting of about 200 from a randomized experiment concerning a job training program, was considered as the treated group in an observational study, whereas the comparison group was to be derived, as typical in such observational studies, from a large-scale database (e.g., either the CPS or the PSID). Estimates of the treatment effect were then obtained using the standard array of econometric/statistical modeling tools on the actual treated units and the observational comparison units. These tools provided answers that were typically wild, and often absurd, when compared to the benchwork estimate available from the randomized experiment. The conclusion, which is I believe consistent with Metcalf's position, is that this documents the fact that such observational studies cannot be trusted to produce honest policy-relevant estimates of treatments.

When Imbens and I presented this example in class, it was in the context of already having warned the students of the extreme extrapolation often implicit in estimates based on such methods, and of already having exposed them to propensity score methods (Rosenbaum and Rubin 1983, 1984, and 1985; Rubin and Thomas 1992a, 1992b, 1996), which avoid such extrapolation. Propensity score methods can also directly lead to the conclusion that, despite the apparent wealth of comparison information available in large databases such as the CPS and PSID, the treated and comparison groups may be so far apart that there are no trustworthy conclusions possible. Two economics students, Sadek Wahba and Rajeev Dehejia, pointed out that the conclusion from LaLonde, to the effect that such studies are hopelessly unreliable, should be decomposed into two

crisper issues. First, are the *data* from studies such as LaLonde's hopelessly unreliable? Second, are the *standard methods* used to analyze such data hopelessly unreliable? We all seem to agree that the latter is true, but what would happen if LaLonde's data were reanalyzed using the far more appropriate propensity score methodology, now very popular in much of social science and medical research (e.g., U.S. GAO Report, "Breast Conservation Versus Mastectomy," 1994).

LaLonde graciously supplied the data, and Sadek and Rajeev went to work. Despite the thousands of potential control units in these large-scale data sets, only about 200 or fewer were similar enough to the treated groups, with respect to their propensity scores, to be considered to constitute a reasonable comparison group for the treated. Adjustment then took place using special versions of either subclassification (Rosenbaum and Rubin 1983) or matching (Rosenbaum and Rubin 1985) on the propensity scores, with possibly some simple OLS regression for minor adjustments. Of great importance, the results based on the propensity score technology tracked those from the randomized experiment, even with respect to interactions between treatment and some background characteristics. An initial reference for this project is Wahba and Dehejia (1996).

Certainly this work does not show that using propensity score techniques in observational studies *will* always either 1) conclude the treated and comparison groups are too far apart, or 2) provide an estimate like that from a randomized experiment. But Wahba and Dehejia (1996) provide an important "existence theorem," showing that propensity score technology, because it inherently addresses problems of extrapolation, *can* produce acceptably accurate estimates of causal effects from observational data in cases where the standard OLS or selection model methods fail to do so.

My conclusions, therefore, are a tempered version of Metcalf's. That is, we should push for randomized experiments whenever possible, but because observational data will nearly always be cheaper to obtain and more readily available, we should be willing to analyze nonrandomized data, but with great care, using appropriate propensity score methods and avoiding unreliable model-based extrapolations employing standard statistical or econometric models. These models have their place, and the ideas underlying some of them can be extremely useful in some contexts (e.g., see Angrist, Imbens, and Rubin 1996); however, they must be used insightfully and not be used, as often advocated, as off-the-shelf solutions to the problems of possible "selection bias" in observational studies.

References

Angrist, J.D., Imbens, G.W. and Rubin, D.B. Forthcoming 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, Applications Invited Discussion Article.

Heckman, J.J. and Robb, R. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*. Eds. James J. Heckman and Burton Singer. Cambridge, UK: Cambridge University Press.

LaLonde, R.J. 1986. "Evaluating the Econometric Evaluations of Training Programs With Experimental Data." *American Economic Review* 26.

Rosenbaum, P. and Rubin, D.B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41–55.

Rosenbaum, P. and Rubin, D.B. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” *Journal of the American Statistical Association* 79: 516–524.

Rosenbaum, P. and Rubin, D.B. 1985. “Constructing a Control Group Using Multivariate Matched Sampling Incorporating the Propensity Score.” *The American Statistician* 39: 33–38.

Rubin, D.B. and Thomas, N. 1992a. “Affinely Invariant Matching Methods with Ellipsoidal Distributions.” *The Annals of Statistics* 20 (2): 1079–93.

Rubin, D.B. and Thomas, N. 1992b. “Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Covariates.” *Biometrika* 79 (4): 797–809.

Rubin, D.B. and Thomas, N. Forthcoming 1996. “Matching Using Estimated Propensity Scores: Relating Theory to Practice.” Forthcoming *Biometrics*.

U.S. GAO Report to the Chairman, Subcommittee on Human Resources and Intergovernmental Relations, Committee on Government Operations, House of Representatives. 1994. “Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies.”

Wahba, S. and Dehejia, R.H. 1996. “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs.” Cambridge: Department of Economics, Harvard University.

THIS PAGE INTENTIONALLY LEFT BLANK

6 Postsecondary Education

THIS PAGE INTENTIONALLY LEFT BLANK

Tracking the Costs and Benefits of Postsecondary Education: Implications for National Surveys

Michael S. McPherson
Morton O. Schapiro

INTRODUCTION

Our assignment is to advise NCES on ways their data collection activities could help shed more light on understanding the costs and benefits of higher education. This paper begins with a discussion of how educational impacts are identified and valued and then goes on to distinguish between the immediate and long-run consequences of educational investments. This discussion is followed by an explanation of what we mean by "high quality" data in arguing for the importance of certain types of longitudinal data sets. The next section addresses the role of "educational treatments" in identifying how educational efforts and resources translate into impacts on students' learning and concludes with a discussion of the usefulness of simple cost/benefit measures in international comparisons of educational "productivity."

CAUSAL VERSUS EVALUATIVE ISSUES

Appraising the costs and benefits of postsecondary education requires knowledge of the impacts of such education—a problem of identifying causation—and knowledge of the values to be placed on those impacts.

Difficulties in Identifying Causal Impacts

It is often relatively easy to identify *differences* between people who have and who have not attended college, or even among those who have had different types of postsecondary experience. But it is much harder to identify the causes of these observed differences among, for example, college graduates and high school graduates.

Two major statistical problems that make causal analysis in this area difficult are *maturation effects* and *selection effects*. Maturation effects create an important hazard for individuals who try to reflect on how their college experience affects their own lives. Looking back, it may be easy to identify ways in which one was different after college than before attending college. But to some unknown extent, those differences, rather than being caused by the college experience, may have been simply a result of aging. In studying individuals, it is hard to surmount this problem of

distinguishing the effect of the college experience from the simple effect of the passage of time. This is one basic reason for attempting to assess the effects of college by comparing college-goers to non-college-goers, rather than simply looking at changes occurring in the lives of people who did attend college.

But comparing college-goers to non-college-goers raises the problem of selection effects. These arise because the processes that determine who goes to college, as well as who goes to which college, are far from random. A great deal of evidence shows that college-goers differ systematically at time of entrance from non-college-goers. College-goers come from families with higher incomes; they score higher on average on aptitude tests; they are more likely to have parents who attended college, and so on. An important advantage of rich longitudinal databases tracking individual life histories—such as the National Longitudinal Study of 1972 (NLS–72), High School and Beyond (HS&B), and the National Education Longitudinal Study of 1988 (NELS:88)—is that they allow us to observe and statistically control for many of these differences in estimating statistically the impact of college experiences on later life.

Unfortunately, however, no data set is rich enough to enable us to observe all the ways in which college-goers differ from non-college-goers (or the ways in which people with different postsecondary education experiences differ from one another). To the extent that these unobservable differences between college-goers and non-college-goers themselves lead to differences in what we observe about people in their later lives, we are at risk of mistakenly attributing these later differences to the college experience, rather than to the unobservable differences that led one group to attend college while another did not. Econometricians have spent considerable energy and imagination in finding ways to allow statistically for these selection effects, and much progress has been made. Still, selection effects remain a great obstacle to sorting out the causal impacts of college-going.

Even more difficult than measuring the effects of college is understanding why or how those effects occur. Better knowledge of the effects of college on people's later lives might help guide decisions by individuals about whether to attend college or by governments about whether to encourage college attendance. But knowing what *features* of a college experience lead to particular outcomes would be of great help in permitting colleges to improve their operations. Clearly to study such problems requires a much more fine-grained measurement of various dimensions of the college experience than most existing data sets permit.

Difficulties in Evaluating Outcomes

Cost-benefit analysis requires not only the identification of outcomes but also the evaluation of those outcomes, and of the inputs that produce them in systematic and preferably quantitative ways.

Benefits of postsecondary education appear partly in labor markets. It is commonly believed that postsecondary education equips people with skills and knowledge that make them more productive. Such higher productivity may then turn up in higher wages, so that the wage differences between, say, college graduates and high school graduates are an index of the social benefits of higher education. But even when the focus is limited to labor market effects, this analysis is not so simple. First, of course, selection effects like those just noted imply that differences in wages

between high school and college graduates are probably not all due to the effects of college. Indeed, it is possible to develop a coherent economic model of college in which the economic function of higher education is to sort out more and less productive individuals, rather than adding to their individual productivity. But even when wage differences result from changes in individual productivity caused by college, these wage differences may either understate or overstate the impact of college on economic productivity, simply because wages may understate (e.g., school teachers, public defenders) or overstate (e.g., investment bankers, deans) the social contributions of particular jobs.

Quite apart from labor market effects, higher education may make people more valuable in other ways, as by making them more politically active or more community minded. These effects are hard to measure in a causal sense, and even harder to measure in a cost-benefit sense, since putting dollar valuations on such effects is difficult.

It is also important to note in passing that higher education makes major contributions to social productivity through its contributions to knowledge accumulation and basic research. Examining ways to improve measurement of these benefits and of the cost of producing them is beyond our scope here.

Unlike the measurement of benefits, the measurement of costs may appear straightforward. But it is actually more difficult than it may appear. One difficulty is that of attributing costs to particular activities and hence to particular outcomes. It is, for example, quite difficult to separate the costs of graduate and undergraduate education in most existing data sets. (There are, of course, conceptual problems in trying to allocate those university costs that contribute jointly to undergraduate and graduate education, but even simple measures, like the number of graduate and undergraduate courses taught by faculty members, are quite hard to come by.)

Another difficulty is sorting out private and social costs and being clear about who pays. The price charged at most colleges, and especially at public colleges, is well below the cost of production. At private colleges, gifts from alumni (often accumulated in endowments) and at public colleges, appropriations from state governments, keep the price to families down. Thus, the calculus of whether college pays in a cost-benefit sense for the family is quite different from the question of whether it pays for society.

It is also important in measuring private costs to gain clarity about what the student and family actually pay. Because of the importance of financial aid, both grants and loans, the actual costs of attending a particular school may be quite different for different individuals. Further, the most important cost of college for most people is not the out-of-pocket price paid to the school, but rather the opportunity cost of student time—the earnings forgone by reducing or eliminating work hours to attend school.

What is the role of NCES in contributing to these evaluative questions? We would underline the importance of NCES recognizing the limitations imposed on it by its role as a government statistics-gathering agency. Actually putting dollar values on various benefits (and to a lesser extent on costs) is ultimately a political decision—a public decision about values. The job of NCES is to provide the information to support that public decision. So, for example, it would be a contribution

if NCES studies could shed light on the impact of postsecondary education attendance on the likelihood of one's participation in volunteer public service activities; it would not be smart for NCES to attempt to put a dollar valuation on the worth of such service contributions.

THINKING THROUGH IMPACTS OF POSTSECONDARY EDUCATION

In studying the impacts of higher education on individuals, it is important to distinguish relatively direct and immediate educational impacts from the long-run effects on earnings and quality of life that are the ultimate payoffs of higher education.

Immediate Educational Impacts

Typically studies of educational assessment and educational production functions assume that education aims at certain impacts on knowledge and cognitive capacities that are thought to be directly related to educational inputs. NCES has strengths and weaknesses in developing data for these kinds of studies.

The basic strategy of such studies is to relate variation in educational inputs to available outcome measures. Abstractly, the ideal framework for such a study is an experiment: introducing planned variation in an input of interest, while applying different levels of the input randomly to a set of students. The fact is, however, that experimental studies of this kind are relatively rare. Much more common are "natural experiments," where naturally occurring variations in inputs of interest are related to corresponding variations in outcomes.

The principal strength of NCES for such studies lies in its ability to develop reliable comparative data for different institutions. Having comparative data across institutions is valuable because it allows for more variation in both inputs and outputs than one is likely to observe within a single institution or a narrowly confined set of institutions. NCES longitudinal surveys like High School and Beyond have enough reach to incorporate institutions with widely varying input levels—large versus small average class size, rich versus meager library resources, and so on.

The principal weakness of NCES here is the counterpart of its strength: the impracticality of generating in-depth data for individual institutions. Two different students at the same institution may have sharply different educational experiences. Without being able to track such variations internal to institutions, educational production function or outcome studies will inevitably involve *averaging* over both the input levels and the outcomes experienced by different students.

If one considers the existing longitudinal surveys (NLS–72, HS&B, and NELS:88), it is clear that they are richer in their information about individual students than in their descriptions of educational environments and inputs at the postsecondary level. Thus, these surveys report the results of student performance on a battery of tests at high school completion and in later years and provide some data on the quality of performance in college—GPA and the like. Information on the learning environment—class size, pedagogical techniques employed, characteristics of faculty—are not a

focus of study, and can be inferred, if at all, mostly through linking the survey data to information from the IPEDS financial statistics survey, which itself tracks only very general institutional characteristics, such as spending, in broad categories. And as noted, these surveys do not permit any tracking of differences among students within a school on the educational inputs directed toward individual students.

These limitations are not surprising, and imply no criticism of the existing surveys. The basic fact is that the longitudinal surveys have not been designed principally with the goal of studying direct educational impacts at the postsecondary level. It is important to appreciate that, as a result, they are poor instruments for this purpose. And because this imposes a real gap in our knowledge of the causal consequences of postsecondary education, it also limits the value of these surveys in studying the costs and benefits of higher education. The discussion below will focus on what kinds of efforts NCES could make to address these limitations.

Long-Run Consequences of Educational Investments

We have just been noting limitations on the ability of existing surveys to shed light on what actually happens to students as a direct consequence of their educational experiences. Fortunately or unfortunately, much economic analysis of the long-run effects of college attempts to measure these effects while sidestepping completely the question of how those effects are produced. This is of course the model of the classic "human capital" study, which attempts to measure the private and social returns to education while treating the educational treatment itself as a "black box."

Many studies of the returns to education have treated the basic unit of education as the "year," and have viewed the returns to an added year of education simply in dollar terms, comparing the earnings of those with more and less schooling. This formulation makes the educational input a homogeneous commodity, the "year," and makes the educational output another homogeneous commodity, "the dollar earned." Much has been learned from models that employ such radical simplifications, but plainly much is also omitted. In particular, such studies are worthless from the standpoint of asking how to improve education—whether one type of education or one way of "doing education" is more valuable than another.

More recent studies have added complexity to this simple model of educational effects. On the input side, there are attempts to recognize that educational inputs differ in their intensity (measured, for example, by dollars spent per student on instruction), as well as their duration (measured by years of school). Researchers have attempted also to measure the returns to different types of schooling—public versus private, community college versus proprietary vocational school. Studies of this kind are potentially of great importance in guiding decisions about public investment, which of course raises the stakes in ensuring that such studies can be conducted reliably. On the output side, there are efforts to recognize that the impact of college experience may show up in places other than the paycheck—in choice of vocation and in the various non-pecuniary dimensions discussed above.

Plainly attempts to move in these directions, recognizing the multidimensionality of both educational inputs and outputs, raise data demands rapidly. On the input side, one runs into the

problem discussed in the previous section that existing longitudinal surveys have only quite gross measures of educational characteristics of institutions and provide virtually no information on differences in the educational inputs applied to different students. On the output side, the surveys are richer, since they include significant attention to attitudinal questions and to activities outside the workplace. It is our sense that these dimensions of the data in the existing national surveys may have been underexploited. (We think this is clearly true of work by economists, but are less well informed about work done with these data in other academic specialties.)

A major question here is that of the *pathways* through which college experiences influence activities in later life. Consider, for example, evidence that college graduates are more likely to participate in community service activities. This could just be a selection effect—that people who are more likely to engage in public service are also more likely to attend college. But even if the result is not spurious in this way, interpreting it remains a complex matter. Is this because college has changed their attitudes—increasing the value they attach to public service; changed their *capacities*—so that they are asked to do more because they do it more effectively; or changed their opportunities—so that they are offered more interesting or rewarding service opportunities owing to their higher status as college graduates? It is far from clear that survey data can help much in answering these questions, but they are certainly important ones to keep in mind in evaluating research findings.

THE NEED FOR GOOD LONGITUDINAL DATA

As noted earlier, the ideal way to study both short- and long-term effects of college experiences would be through conducting experiments involving random assignment of subjects. Without discounting the possibility of doing this in some settings, it is clear that most knowledge about college effects will not come from this source. Rather, we are thrown back on the "natural experiments" generated by the educational system.

We can make no more important point than that high-quality longitudinal data is an essential component of reliable studies of college effects in non-experimental settings.

This point applies to studies of immediate effects of college experience on student attributes as well as on long-run studies of educational impacts on life outcomes. For the study of direct effects of college, longitudinal data provide benchmark information on student attributes before or at the time of college entry in order to examine how college *changes* these attributes. It is further necessary to make comparisons between the changes experienced by those with and without college experience to distinguish college effects from maturation effects.

More subtle distortions can also be corrected with adequate longitudinal data. Suppose, for example, that a group of students enter college with scores on cognitive tests equal to those of a group of non-college-bound students, and 4 years later the college students have improved their scores more than the non-college-bound. Since we do have pre-treatment data, can we safely attribute the difference to the college experience? Not necessarily, for the non-college-bound, even with the same test scores, may differ from the college-bound in ways that are not picked up in the test score data. They may, for example, attend college because they are more motivated, or because they have reason to believe they will learn faster, and so on. A rich longitudinal data set that tracks pre-college

differences among youth may provide measures of variables that correlate with unobserved differences like those just noted, allowing statistical control for these differences that will otherwise confound results. Ultimately, there is no sure cure for such unmeasured effects except random assignment, but good longitudinal data are helpful.

In studies of long-run effects of college, pre-college data are needed for all the same reasons. Post-college data are quite valuable as well. Obviously, one must have data for that point at which the long-run effects of college are measured. And indeed, one can do good work limiting one's data to such "end-point" information. In such work, one is lumping together all the very different kinds of effects college may have on life outcomes, and all the different pathways through which these effects may operate. This sort of "reduced form" or "black box" approach is legitimate, but limited. With good data, much can be added by a more "sequential" approach. Thus, for example, a particular type of college experience may increase a person's likelihood of attending a professional school after college, thereby influencing future vocational choice and career opportunities and ultimately earnings. Sorting out such causal pathways can be instructive in ways that simple bottom-line assessments of the impact of college are not.

What is "Good" Longitudinal Data?

Several times we have referred to the value of "good" longitudinal data. This section concludes by being more specific about what we mean by good data in this context.

Obvious statistical requirements include sample sizes that are adequate to the levels of detail in the analyses that are contemplated and high, preferably uniformly high, response rates among population subgroups. We would also stress three desiderata that are more specialized to longitudinal surveys.

First, it is very helpful to minimize reliance on recall. For pre-college information, this points to the advantage of beginning surveys when subjects are young, so that contemporaneous information on their background, environment, and characteristics can be collected. For post-college information, this points to the advantage of reasonably frequent resurveys in order to minimize reliance on recall to fill in the gaps. The reason for avoiding reliance on recall is obvious: recall is not just imperfect but frequently biased, as the hundreds of thousands of people who claim to have seen Don Larsen's World Series perfect game in person would, unfortunately, not attest.

Second, it is important where possible to cross-check individually reported data against administrative records. One striking illustration of this point is provided by the Postsecondary Educational Transcript Study, which recovered data from colleges on the educational records of students in the NLS–72 study. The re-study turned up large inaccuracies in student reports of their transcripts. The NPSAS studies of student aid do an excellent job of this kind, by corroborating student and parent reports of college financing arrangements against college student aid records.

Our third desideratum for a longitudinal survey is a long span of years, both pre- and post-college. Owing to the problem of selection effects, good data on pre-college background and

experience, extending even as far as early childhood, can be of enormous value. Following a cohort well into the post-college years is also of great value, since it is very plausible that the effects of college are long-lasting, and some may take a long time to manifest themselves. People often do not attend college immediately after high school, even those who graduate from college typically now take more than 4 years to complete, and many people obtain post-collegiate education. After education is complete, there is often a period of job experimentation and search that may last 3 to 5 years or more. For many people a long-term career profile does not begin to jell until they are in their early 30s.

These long time spans are obviously quite frustrating for two reasons. One is the perennial desire for prompt answers to urgent questions. The other is the worry that the world changes so fast that data obtained about the college experiences of people now in their 40s may be irrelevant to the educational experiences of those in school now. In practical terms, and given limited budgets, at any particular time this question comes down to two more focused choices: should we do another round of an "old" longitudinal survey, or should we use those resources to start a new one? And, should a new longitudinal survey start with early childhood, or pick up people at a point closer to maturity? Although the answer is always a matter of judgment, we would express a preference for the long view, based on the suspicion that the system will always tend to be biased in favor of a short-run view. Our reasoning is this: we really should view work in this area as "basic research." The social science community is far away from having reliable knowledge about the effects of college or of how those effects are brought about. One of the advantages of a deeper understanding would be an ability to explain how differences in the educational system influence educational outcomes from one decade to the next. Investing now in the data collection efforts that will eventually bear this fruit, and summoning the patience to await the maturing of those data sets, seems to us the more sensible course.

THE NEED FOR BETTER DATA ON EDUCATIONAL TREATMENTS

One of the major lessons of this review is the high potential payoff from data that get closer to the actual educational "treatment" than existing national data sets do. Unfortunately, another lesson is the great difficulty of getting such data in a reliable form, and at reasonable cost. The appeal of such data, as should be clear from our earlier discussion, is the opportunity they would provide to get a clearer fix on how educational efforts and resources get translated into impacts on students' learning and hence on their later lives. Such data would help overcome two major limitations on existing work on higher education based on national data sets. First, available data will often fail to detect what may be large differences in the educational treatments received by students. For example, two institutions may have identical levels of instructional spending per student, or of numbers of library volumes, while offering very different instructional or library experiences to their students. Although this fact will not introduce any econometric bias into studies that ignore these differences, it will reduce, perhaps substantially, the precision of any findings. Second, existing studies average over educational experiences that, quite likely, vary substantially across students in the same institution. Ignoring this variation will also reduce the precision of estimates. But, more significantly, unmeasured variation in the educational treatments applied to different students may be a source of bias. If differences in such student characteristics as social background, academic ability, or motivation influence the educational treatments those students receive, there will be a tendency to

overestimate the impact of these background variables on educational results, and to underestimate the impact of educational resources.

Better measurement of educational "treatments" would be of great value in estimating educational production functions, in assessing the returns to different types of education, and in studying the cost-effectiveness of different educational strategies.

As mentioned above, in existing longitudinal studies information about the educational environment of the colleges attended by a student in the sample is provided principally by linking the survey data to Integrated Postsecondary Education Data System (IPEDS) data on the institution. IPEDS, an institutionwide survey, provides no information on differences in educational resources provided to different individuals in the same school, or even to different groups of students (such as graduate and undergraduate students) within the school. Moreover, even the information on the resources applied to the average student are limited. IPEDS, a finance and enrollment survey, does not describe physical inputs, but rather only the dollar amounts spent in broad categories. It is also difficult in some cases to distinguish dollars spent on educational purposes from dollars spent for other institutional purposes in the IPEDS data. Finally, one important educational input—the quality of other students—is not measured at all in the IPEDS data.

Improving this situation would be a great help in improving understanding of the costs and benefits of higher education, and especially in helping learn about the relative costs and benefits of different types of or approaches to higher education. Two rather different kinds of improvements in data on higher education inputs should be distinguished. First is better measurement of actual inputs, rather than simply dollars. Thus, data on class sizes, on instructional methods employed, on the role of graduate assistants versus faculty in teaching, and so on, could be enormously helpful. Ideally, one would have data individualized to students (such as the sizes of classes experienced by a given student in a longitudinal sample). More realistically, one might hope for such data by classes of students (freshmen, sophomores, and so on) But even to have such data for the average student in a school would be a real improvement.

The second type of data improvement would be more refined measures of costs. Thus, for example, it would be very helpful to be able to distinguish costs of graduate and undergraduate education in the IPEDS data. Refinements of some expenditure categories in the IPEDS survey would also be welcome—a favorite example is including the costs of the admissions office in student services.

Conceivably, some refinements of the latter sort might be introduced in future generations of the IPEDS survey. However, as a survey intended to be a census of all postsecondary institutions, it would be unreasonable to expect IPEDS to be a vehicle for collecting detailed data on educational treatments. Several strategies are offered here that may be worth considering to enable NCES to make progress on this front.

First, NCES might consider doing a "long-form" IPEDS for a sample of institutions, analogous to the Population Census long forms. For example, if 5 percent of postsecondary institutions were selected for more intensive treatment, that would amount to about 160 public and

private not-for-profit institutions, and a somewhat larger number of proprietary institutions. Ideally such a long form should be administered at random (as the Population Census does), but even if the institutions had to be selected on a voluntary basis, the effort might be worthwhile. It would also be reasonable for NCES to reimburse institutions for the expense of undertaking a more thorough study.

A variation on this idea would be to link an intensive effort to measure institutions' educational practices to the participation of those institutions in a longitudinal survey. It would be reasonable in such a framework to include fewer institutions in the study, with more students from each institution. The trade-off is that one would have less variation among institutions but more information about each one. If it were possible in the context of such a study to measure actual variation in the educational resources provided to different students, being able to include this kind of variation would probably more than make up for having fewer institutions in the sample.

Finally, NCES might consider ways of approaching getting these data through cooperation with institutions that are interested in doing such studies internally. Some institutions are interested in gathering detailed data on their internal educational practices, and in using those data to improve their practices. For the individual institution, the inability to make comparisons to other institutions is a real drawback. NCES might have some opportunity to help to support individual institutions in making such efforts, and might be able to help standardize the efforts of different institutions in order to facilitate comparisons. The loss of randomization implied by this strategy is a significant drawback, but the advantages of having institutions as enthusiastic partners in the effort would be considerable.

We offer all these suggestions tentatively. We recognize that any of these efforts would be expensive and would challenge a general reluctance of colleges and universities to make detailed information about their internal practices known. Yet the potential gains in understanding are considerable.

MAKING INTERNATIONAL COMPARISONS

More and more countries have been attempting to measure the performance and effectiveness of their higher education industries, giving rise to the obvious question of whether there are particular indicators that would be of use in making international comparisons. If so, it would be important to make sure that NCES data sets include such information.

A recent monograph (Gaither, Nedwek, and Neal 1994) reviews some of this literature, dividing performance indicators into three types: input measures (test scores and secondary school performance of entering students, prestige of programs from which faculty received Ph.D. degrees, and so on); process indicators (library use, meetings with faculty advisors, and so on); and output measures (number of degrees awarded, graduation rates, faculty publications, percentage of students going to graduate school, and so on).

In their discussion of performance indicators in Britain, they point out that most of the indicators are input rather than process or output measures. Key measures include admittance rates and entry scores for undergraduates, their subsequent graduation rates and postgraduate employment

experiences, and for faculty, research grants and publications. In terms of cost measures, staff/student ratios, unit costs, and institutional revenue and expenditures data are all used, although the authors report that difficulties in cost allocation procedures have made it very hard to evaluate efficiency.

Indicator systems in Canada also center on simple input and outcome measures, with relatively little on the process side. Popular indicators include time to degree, degrees granted, and various expenditure types. Typical indicators used in Australia are graduation rates, class size, and a series of "destination outcomes," including post-graduation employment, study, and salary. Again, process measures are neglected. The Netherlands concentrates on such teaching indicators as the number of students and their length of study, while Finland relies on similar aggregate measures. Sweden also concentrates on basic student enrollment and graduation indicators. Denmark supplements this sort of data with "customer satisfaction" information gathered from interviews with students, graduates, and employers.

In summarizing the indicators used in the seven countries they examine, Gaither, Nedwek, and Neal conclude that certain simple input and output measures—with some variation—are commonly used by educators and government officials in a variety of contexts. This raises two questions: do existing data sets in the United States allow us to compute these measures; and are these measures really of use in comparing the costs and benefits of higher education across the world?

The answer to the first question is "yes." It is not very difficult to collect information on the number of degrees awarded or total educational expenditures. In fact, variables of this type were mentioned in our earlier discussion. But, for example, in an analysis of the cost effectiveness of public higher education expenditures in the United States, we would hesitate to simply divide the number of degrees by state spending and compare that "productivity" measure across states. There is no reason to expect that the educational quality is similar enough to give any real meaning to this ratio, and the same can certainly be said for comparisons across countries.

We are therefore rather skeptical that data could be developed that would permit meaningful international comparisons of the relative costs and benefits of America's postsecondary education enterprise. However, the recommendations we have made here concerning data collection and analysis could help us increase our understanding of the costs and benefits of higher education *within* our country. The payoff would be considerable, both from the standpoint of individual students and colleges and from the nation as a whole.

REFERENCES

Gaither, G., Nedwek, B., and Neal, J. 1994. *Measuring Up: The Promises and Pitfalls of Performance Indicators in Higher Education* (ASHE-ERIC Higher Education Report No. 5). Washington, D.C.: The George Washington University, Graduate School of Education and Human Development.

Special Issues in Postsecondary Education and Lifelong Learning

David W. Breneman
Fred J. Galloway

ABSTRACT

In an effort to improve the data collection abilities of NCES, this paper identifies six emerging research areas in postsecondary education and lifelong learning: institutional finance, postsecondary assessment, loans and student indebtedness, the school-to-work transition, technological change and distance learning, and the proprietary sector. For each of these emerging issues, we provide both a contextual discussion and a review of the extent to which existing NCES databases can respond to these emerging issues.

Recommendations are provided for both data collection and data dissemination activities and are ordered by our perception of where "the biggest bang for the NCES buck" might be. These include increases in the proposed coverage and sampling frame of several of the data sets; the establishment of agreements with outside agencies to provide information that was previously self-reported; the creation of a new database that surveys high school graduates each year; an increase in the frequency with which the National Postsecondary Student Aid Study (NPSAS) is administered; an increase in the number of analysis reports issued each year; and an increase in both the coverage and availability of the public access versions of several of the data sets.

INTRODUCTION

Few areas of social policy are devoid of turmoil and disagreement regarding future directions, and the world of postsecondary education is no exception. Indeed, during the first years of the 1990s, higher education funding from state governments was the one area of broad state responsibility that saw a percentage decline in support. As a consequence, tuition levels increased sharply, access for low-income students was reduced, and worries about college affordability increased for middle and even upper income families. Adding to the dilemma of families and students is the pivotal role of postsecondary education as the gateway to challenging and remunerative employment, coupled, however, with a growing dispersion of opportunities and earnings, even among the college-educated. As higher education becomes essential, its economic payoff appears more like a lottery, with big winners and losers. The recent explosion of debt financing adds further tension to this relationship between investment in college and economic return.

Broad social and economic developments such as the above are beyond the power of any data collection exercise to anticipate or influence; nonetheless, as the scale of costs and benefits to students and to society expands, it is incumbent on the National Center for Education Statistics (NCES) to monitor and help policymakers and others interpret trends in the industry. Many of the data sets currently collected on postsecondary education and its students perform that function effectively, but we believe that cost-effective improvements are possible. Our discussion is organized around six emerging research issues, described in the next section. Following a brief discussion of each issue, the paper examines the current state of data collection in each area. The paper then concludes with recommendations for modifications and enhancements of NCES data collection practices.

EMERGING RESEARCH ISSUES

Institutional Finance

As one considers the last four decades, the overriding picture of postsecondary education is of an expanding, growing industry, with increasing enrollments, growth in the number and size of institutions, employment, and resources. Only recently have these patterns begun to shift toward stasis, with a focus on retrenchment, doing more with less, and growth by substitution. No one knows for certain whether this recent trend will continue, but no key revenue source seems poised for sharp increase. As noted in the Introduction, state support has slowed, and federal dollars for student financial aid and for research are under similar budgetary stress. Tuition increases of recent years have slowed, as private colleges and universities fear that they are pricing themselves out of reach, while political reaction to public tuition increases has forced a slow down. Philanthropy appears to be the source most emphasized, as both private and public institutions step up their fund-raising efforts. Suffice it to say, however, that the resource outlook for most colleges and universities is as cloudy today as it ever has been.

Among the responses of colleges and universities, two will serve as examples of the sharp changes under way. In the private, nonprofit sector, institutions are engaging in calculated price discrimination in the form of student aid discounting to fill their classes and to attract students of particular quality. This tendency toward discounting has accelerated in recent years, as colleges literally fight, in some cases, for survival. The economics of discounting in this sector is only beginning to be understood, and the Integrated Postsecondary Education Data System (IPEDS) database is only partially adequate for analysis tasks. In particular, that database does not differentiate between types of discounting, and is not sufficient for monitoring institutional financial stress in a rapidly changing environment.

In the public sector, talk is increasing of privatization in some form, with state universities becoming state-assisted institutions, relying more on tuition and private fund raising than on state support. The extent to which this trend is occurring is a matter of conjecture, because databases are not clearly focused on such issues. In both this and the prior example, higher education could be better served by improvements in financial data collecting, and our suggestions for change are noted in the next section.

Assessment

As the private and social costs of postsecondary education have grown, it is not surprising that both families and policymakers have sought more information about the benefits of higher learning. We think of the "assessment movement" as a rational response to the need for better measures of educational outcomes, thereby permitting individuals and society to calculate rough cost-benefit ratios. As postsecondary education has evolved from an elite to a mass phenomenon, more sophisticated measures of educational results have become necessary, responding to the diversity of students and reasons for enrollment.

Economic rate of return calculations, first developed in the 1960s, helped to fuel the growth of college enrollment, as the basic message was that college was a good investment. Today, however, many question whether the country has moved too far, with growing numbers of students enrolled in remedial courses and not completing their degrees. Some argue that we have too limited a range of postsecondary learning options, pointing to German apprenticeships as a better model. And how does one evaluate the many students who begin but do not finish programs? Should such students be viewed as "wastage," with the focus turned to retention, or have they gained something of value, and is concern misplaced? These are among the important policy questions that NCES longitudinal data sets can help to answer, provided certain changes are made.

Loans and Student Indebtedness

One of the most dramatic shifts in college finance in recent years has been the growing share of economic costs borne by students, financed primarily through increased student loans. While much of the policy focus has been on aspects of loan repayment—default rates, income-contingent options, and so forth—more fundamental, long-term issues of human behavior are involved. High levels of student debt may affect career choice, marriage, and child-bearing decisions, as well as patterns of saving and consumption over the life cycle. While many have speculated about such issues, very little empirical information has been available to analysts seeking to understand these relationships more clearly. The growth of student debt appears unstoppable; thus, it behooves us to begin to collect data that can help us understand the long-term implications of this social choice.

School-to-Work Transition

As the labor market grows ever more complex, the old verities about high school transitions to work, or high school transitions to college and then to work, are increasingly inaccurate. High school graduates today face a limited, and for the most part, unappealing set of choices—dead end jobs, the military, unemployment, crime, or college. Not surprisingly, college appears to be the best choice, but then there is the issue of which college, which major, and at what price. After college graduation, the choice of work or graduate school comes up, and throughout one's career, the decision to return full or part time for further education is a continuing dilemma. In short, the worlds of formal education and of work are no longer clearly segmented by age or employment situation. This blurring of circumstance yields a need to know more about the choices facing people at several stages in life, and the realistic options open at each stage. Modifications of the several longitudinal files maintained by NCES is the obvious way to enhance our knowledge in this area.

Technological Change and Distance Learning

Perhaps the greatest imponderable in our current situation is the prospect of technology for transforming the way we deliver education. One can hear the voices of prophets proclaiming a new millennium, in which education as we have known it will diminish, even vanish, from the scene. No longer will physical places called "colleges" or "universities" be necessary because anyone will be able to tap the resources of electronic information systems and video texts. New suppliers are expected to enter the market, providing education at much lower cost because they are not freighted down with either the physical plant or the outmoded traditions of academia. In this view, all that saves colleges and universities is the near monopoly on credentials, a too fragile reed to survive the onslaught of technological advance.

Others see the new technology as the salvation of college and university education, because at last a way may be found to escape the "cost disease," the tendency for unit costs to rise annually by about 3 percent above inflation. And still others dismiss the talk about technological revolution as yet the latest over-promoted fad, analogous to the promises made in an earlier generation for educational television. Much rides on which vision is the accurate one, and information about trends is clearly crucial to the evaluation of claims and promises. NCES can play a key role in helping to shed light on this important, but vexing, topic.

Proprietary Institutions

Much of the terrain of traditional, non-profit higher education is well-mapped by the various NCES databases, but the burgeoning universe of profit-making schools and colleges is only lightly covered by existing surveys. These schools are the largely hidden world of postsecondary education, having grown dramatically in response to eligibility for federal grants and loans by their students. Claims and counterclaims about their effectiveness are lobbed back and forth by educators and policy analysts, reflecting largely newspaper accounts and anecdotal information rather than hard data. The simple fact is that we do not know how many are doing a good job and how many are simply exploiting our most vulnerable young people. NCES would do society a great service by focusing on this group of schools and developing a systematic data collection effort that could be effectively implemented.

CURRENT STATE OF DATA COLLECTION

Using the above six categories of emerging issues, this section of the paper discusses the extent to which existing NCES and other databases are able to respond to the questions raised in each area. This discussion leads, in turn, both to modest suggestions for incremental change and to a small number of recommendations for substantial new surveys.

Institutional Finance

In an effort to maintain enrollment levels and ensure a diverse student population, many institutions have dramatically increased their contribution to the student's financial aid package. Although this growth has occurred among all types of institutions, the biggest increase has been at private, 4-year colleges and universities that either use institutional aid to meet enrollment targets, or as a form of merit aid to attract academic stars. This leads to two sets of research questions: first, what effect does this increasing reliance on institutional aid have on institutional health; and second, how does the growth in institutional aid affect such student outcomes as enrollment and persistence?

Unfortunately, current NCES databases provide little if any help in addressing these research questions. For example, one important bit of information needed to address both research questions is the ability to distinguish among institutions that use institutional aid to meet enrollment targets versus those schools that use it solely to attract stars. Without the ability to discern institutional motive in the awarding of aid, it becomes increasingly difficult to apply the appropriate standard of institutional health. For example, for institutions seeking to diversify their student population, net tuition revenue (gross tuition revenue minus institutionally provided aid) seems an inappropriate metric, yet for those institutions trying to meet enrollment targets, it may well be the appropriate measure of institutional health.

Currently, IPEDS collects information on institutional characteristics, including enrollment and financial statistics. However, the categories used to collect the information provide only gross measures, nothing approaching the context required to differentiate institutional motive in the awarding of this particular kind of aid.[1] Even those measures that might provide some hint of administrative context get “scrubbed” by the state higher education associations before arriving at NCES, further reducing potentially interesting variation across institutions. When combined with the other well-known limitations of IPEDS, one wonders if this data set could be successfully reconfigured to address these issues, or if some “student aid management” survey needs to be created.

The limitations embedded in the IPEDS system also extend to the student-based research questions involving enrollment and persistence. Even if IPEDS allowed us to differentiate among institutional motive in the awarding of this type of aid, the measurement of student-based outcomes would most likely occur through NPSAS, where enough financial aid information is collected on individual students to effectively address the second research question. To do this, however, would require that IPEDS and NPSAS be linked, so that the characteristics of IPEDS institutions would be matched with individual students in NPSAS. Even with this linkage, however, the NPSAS sampling frame would probably need to be increased so that there would be enough students in the sample to provide an adequate statistical test for the various student-based propositions concerning institutional aid. And if the information is to be used for policy decisions, then it needs to be available in a timely manner, something that is currently unavailable within the 3-year cycle under which NPSAS operates.

Assessment

As more students return to school for just a few skill-specific courses, the quality of instruction and amount of learning that takes place in postsecondary classrooms becomes increasingly important. To address effectively the growing importance of assessment in postsecondary education, several important modifications must be made both in terms of the information collected and the way in which it is collected.

To understand the importance of the particular information that needs to be collected, it may be helpful to classify students into one of three groups: degree earners, those with some college experience, and those who just enroll in a few specific courses. Although the later group may be growing the fastest, each group faces its own unique assessment needs. For those with college degrees, typical measures of societal assessment include the degree itself, the school attended, annual earnings, cumulative grade point average (GPA), as well as scores on such standardized tests as the Scholastic Aptitude Test (SAT), the American College Test (ACT), and the Graduate Record Exam (GRE). Currently, most of this information is collected by NCES in NPSAS, Beginning Postsecondary Students (BPS), and Baccalaureate and Beyond (B&B), so little additional information is needed for this group of students.

For those individuals with some college experience, the most important assessment measure may be the number of credits earned, the school attended, and annual earnings. Fortunately, most of this information is also collected by NCES. However, for those students who just enroll in a few postsecondary courses, two important pieces of information need to be collected. The first, and perhaps most important, concerns the motivation of the returning student. It seems that if the individual is taking the course for entertainment or personal enrichment, it makes little sense to apply any tools of assessment to the student's performance in the class. However, if the returning student is taking the course for a job-related reason, then some sort of value-added assessment measure is appropriate. The selection of appropriate assessment measures for these students, however, is quite controversial. Short of requiring both pre- and post-tests for this group, traditional measures such as the grade in the course, the quality of the instructor, or any increase in earnings may have to suffice. To the extent that this information is not currently collected, it should be added to the student-based NCES data sets.

Perhaps even more important, however, is the reliability of the information currently being collected. Although much of the relevant assessment information collected in NPSAS is done through transcripts, some of the information is self-reported. As demonstrated by numerous researchers, such self-reported data as annual earnings, test scores, GPAs, and years of education tend to be significantly overstated, introducing enough measurement error into the variables to make them virtually unusable in any statistical analysis. Since this information is already being collected, it makes sense for NCES to go straight to the source wherever possible. For example, test score information could be gathered from the College Board, and income and earnings information from the Internal Revenue Service, rather than relying on any self-reporting. In fact, if such a match could be accomplished, fewer questions would have to be asked in the surveys, thereby freeing up additional resources either to expand the sampling frame or to solicit additional pertinent information.

Loans and Student Indebtedness

In the last few years, students have become increasingly reliant on loans to help finance their postsecondary education. Since increases in student indebtedness have implications for future patterns of domestic consumption, an emerging issue concerns both the extent of the problem (exactly what is the combined debt load of postsecondary graduates) and how these higher debt levels influence acquiring such major items as automobiles and new homes.

Although the lack of up-to-date information on indebtedness has been a major problem in fighting to save subsidized student loans during the recent federal budgetary debate, the timing of this information is more of an issue than its ultimate acquisition. As currently configured, information on student indebtedness is available for both undergraduates and graduate/professional students through NPSAS, and will be available in the future through BPS and B&B. However, the 3-year cycle that drives NPSAS means that to get information from the 1992–93 academic year, one needs to wait roughly 3 years. Given the rapidly changing nature of financial aid programs in this country, the 3-year cycle means that analysts are always behind the curve, forced to speculate on emerging patterns or to conduct their own surveys. Moving NPSAS to a 2-year cycle would help ameliorate this problem.

To a large extent, the same timing issues are relevant in the discussion of BPS and B&B. Originally designed to alternate with each other as a companion to NPSAS, these data sets contain important information on overall student debt levels, but suffer from the same long-cycle problems as NPSAS. Even more important, however, is the need for these surveys to follow students well into their careers, so that the full effects of their postsecondary financing decisions can be documented. Given the standard 10-year repayment period for most student loans, it would seem that individuals would need to be tracked for at least 10 years, and probably more, to capture the behavioral changes that occur as their student debt is finally retired. As such, we strongly advocate both shorter cycles and more follow-ups for the surveys to become truly effective tools for both researchers and policymakers.

School-to-Work Transition

Given the rapidly changing nature of work in this country, today's high school and college graduates face an uncertain future in terms of job availability and rapidly changing skill requirements. To help them plan for this transition, more information is needed on the career paths of recent graduates, as well as intertemporal changes in the distribution of job offers for recent graduates.

To address these issues, contemporary information is needed on two sets of graduates, high school and college. For college graduates (both undergraduate and graduate), the amount of information currently collected by NCES may have to be expanded to include more information on job offers, search strategies, and starting salaries, but the larger issue is the frequency with which the data are collected. Since most of the pertinent information is contained in B&B (which alternates with BPS in NPSAS), the resulting 6-year cycle provides information that is of little practical value to the ultimate consumers of such information—researchers and recent graduates.

Additional support for this proposition comes from both the National Academy of Sciences (NAS) and Tom Mortenson's "Postsecondary Education Opportunity Research Letter." As described in the NAS publication *Reshaping the Graduate Education of Scientists and Engineers*, members of the Committee on Science, Engineering, and Public Policy argue:

> Graduate scientists and engineers and their advisors should receive more up-to-date, accurate, and accessible information to make informed decisions about professional careers. We recommend that a national database on employment options and trends be established (National Academy of Sciences 1995).

Ironically, their recommendation comes at a time when the most reliable source of undergraduate starting salary information has recently been discontinued. In writing about the termination of the Endicott survey on starting salaries of college graduates, Tom Mortenson writes:

> Currently, several of the data sources that reveal the condition of educational opportunity in the United States are under assault. The Endicott survey data on starting salaries of college graduates . . . that was collected and reported by Northwestern University since 1947 was ended in 1994. A 48-year time series of data used in numerous econometric studies of student demand for education has been terminated (Mortenson 1995).

To provide this information in a more timely manner, we recommend that B&B be either included in every NPSAS survey, or that B&B continue to alternate with BPS, but that NPSAS be moved to a 2-year cycle. In this manner, the requisite information would be available every 3 years under our first option, or every 4 years if the second option were adopted.

While our recommendations for improving the timeliness of information on recent college graduates may be resolved by simply changing the cycle on which several databases operate, a more serious structural problem exists for high school graduates, the most overlooked group of individuals in the NCES sampling universe. Although information is collected every 3 years (through NPSAS) for those high school graduates who enroll in college, no information is collected on those who directly enter the work force. For these individuals, a national "black hole" currently exists in terms of up-to-date information on starting salaries and potential career paths. To generate this information, we recommend a short longitudinal study, conducted every year, of our nation's graduating seniors, with at least one 2-year follow-up survey. In this manner, contemporary salary and career information could be gathered and made available in a timely manner for this long-neglected group of individuals.

Creating such a national database would also provide a wealth of information on access and choice for those graduating seniors who elect to continue on to postsecondary education. Surprisingly, this information has been collected only four times in the last 35 years, in 1972, 1980, 1982, and 1992, through the National Longitudinal Study (NLS), High School and Beyond (HS&B), and the National Education Longitudinal Study (NELS) databases. Given that access and choice are two of the main reasons for the very existence of financial aid, it is truly shocking that this information is not collected regularly by NCES. In fact, if such a database were created, it could be linked up with the Common Core of Data (CCD), Schools and Staffing Survey (SASS), and IPEDS

data sets, so that from an information perspective, the entire transition from high school to college (including the characteristics of the high school, the college application process, and the college eventually selected) would be seamless.

Technological Change and Distance Learning

As information technologies continue to revolutionize the way individuals both work and learn, an increasing number of students will spend time in "nontraditional" classrooms. To evaluate the effectiveness of these new modes of teaching and learning, information must be gathered not only on the methods of delivery but also on a variety of student-based outcome measures.

In collecting this sort of information, there are several issues that NCES needs to address. The first involves from which end of the delivery system the data should be collected—the user or the institution. To provide overlapping coverage, we recommend that the data be collected at both ends. In this manner, questions could be added to NPSAS and BPS that measure the availability and frequency of this type of learning at the student level, and similar questions could be added to IPEDS and the National Survey of Postsecondary Faculty (NSOPF) at the faculty and institutional level. In this manner, emerging trends could be identified at both the provider and consumer level, instead of lumping them together into the less interesting "user" level.

Another important issue is the timing of the data collection. Although the NSOPF appears to be on at least a 5-year cycle, the annual nature of IPEDS makes it a useful vehicle for collecting and reporting this sort of information. At the student-level, however, the timing problems previously discussed with NPSAS and BPS are again relevant. To remedy these problems, we encourage NCES to move NPSAS to a 2-year cycle, thereby providing consumer-based information on distance learning in a timely manner.

Perhaps the most important issue, however, is the ability of NCES to go "where the action is"—in this case, the proprietary sector. Although distance learning is occurring across all institutional types, NCES must be able to gather information from this sector or risk misstating the extent of this emerging technological innovation. Unfortunately, the ability of NCES to adequately measure anything in this sector is relatively weak—due largely to the refusal of many schools in this sector to share any information for fear of increased federal regulation. Although no simple solution seems apparent, NCES must increase their coverage of this sector, or risk relying on student-based information to capture this emerging and important trend.

Proprietary Institutions

As described in the last section, the coverage of the proprietary sector by NCES must be increased if the effectiveness of these for-profit institutions is to be debated publicly. Although both IPEDS and NPSAS provide some sectoral coverage, the lower response rates typical of schools in this sector make statistical inference an increasingly difficult task. When combined with the large numbers of schools regularly entering and exiting, even the notion of a "steady state" in this sector becomes somewhat meaningless.

To address these issues, NCES needs to find a way to increase institutional participation among for-profit institutions. If such a method were devised, the institutional sampling frame in the NPSAS and IPEDS databases could be increased, and inferences regarding this sector made more robust. Furthermore, by matching these institutions with their Internal Revenue Service records, financial information could be taken directly from their tax records, effectively solving the "self-reporting" problem. In this manner, both the quality and quantity of data from the proprietary sector would be significantly improved.

RECOMMENDATIONS

In this section of the paper, our recommendations for NCES will be presented. They flow logically from the previous discussion, and are divided into two groups: those dealing with the data collection process itself, and those dealing with the dissemination of information derived from this process. Within each group, the recommendations are ordered by our perception of where the "biggest bang for the NCES buck" might be.

Data Collection

Recommendation #1: Add the following information to the IPEDS, NPSAS, NSOPF, and BPS databases:

Although many of the changes recommended here represent only marginal additions to existing NCES data sets, we believe that their value added greatly exceeds the costs of implementation. For example, the IPEDS database could be made more useful in at least three ways: by adding a set of contextual questions designed to determine institutional motive in the awarding of various types of aid; by including a set of questions designed to solicit information on technological change and distance learning; and by expanding the sampling frame to include more proprietary institutions. In a similar manner, the NPSAS database could be improved by also expanding its sampling frame (which would make a potential linkage between IPEDS and NPSAS even easier), and by including questions on technological change and distance learning. Finally, both the NSOPF and BPS data sets could also be expanded to include questions on technological change and distance learning.

Recommendation #2: Enter into an agreement with the Internal Revenue Service, the College Board, and Educational Testing Service to provide some of the information currently collected through NCES surveys.

Although establishing such a linkage might require a substantial expenditure of political capital, the rewards would be enormous. For starters, such previously self-reported information as income, earnings, and some scores on standardized tests would be made substantially more reliable. In addition to the obvious benefits for both the consumers and practitioners of educational research, this would also mean that fewer questions would be asked in NPSAS, BPS, and B&B, thereby freeing resources either to increase the sampling frame in these databases or to ask other policy-relevant questions. By any measure, such a linkage would increase both the reliability of the data and

subsequent analyses, in addition to either cutting programmatic costs or increasing the scope of the overall coverage.

Recommendation #3: Create a new database that surveys high school graduates every year, with at least one 2-year follow-up survey.

Creating such a database would allow researchers to address intertemporal issues of access and persistence among those high school graduates applying for college, as well as provide salary and career information on those students who enter the work force directly. Since this information has been collected for only 4 years out of the last 35, it would help researchers identify emerging trends among high school graduates, and could help current high school students decide on an appropriate career path. The data set itself could be linked with the CCD, SASS, and IPEDS databases to provide maximum information for the educational researcher and could be relatively "short and sweet," limited to perhaps as few as 8,000 high school graduates annually.

Recommendation #4: Move NPSAS from its current 3-year cycle to a 2-year one.

If NPSAS were administered every 2 years instead of 3, the timeliness of the resulting information and analyses would be greatly improved. Given the dynamics of postsecondary finance, this information needs to be collected at least every 2 years if researchers and policymakers are to stay reasonably ahead of the curve. Furthermore, since BPS and B&B alternate with each administration of NPSAS, the timeliness of their information would also be improved.

Data Dissemination

Recommendation #1: Produce more Postsecondary Education Descriptive Analysis Reports (PEDAR) reports.

Although many NCES users have restricted-access versions of the NCES data sets and many more use the DAS table-generating software, the PEDAR reports have perhaps the widest usage among consumers of educational research. Currently, five of these reports are produced each year, with topics ranging from the packaging of institutional aid to minority representation in higher education. Since the selection process involves a dozen or so proposed topics, it makes sense to produce at least a couple more reports a year, given the dependence of the research community on the reports. Furthermore, if the scope and coverage of some of the NCES databases are increased, then this should naturally be accompanied by the increased dissemination of analyses.

Recommendation #2: More public access versions of NCES databases.

As described in the above recommendation, those individuals without a restricted-access version of a particular NCES database are forced to rely on the PEDAR reports or to use the DAS software. Since this software limits the user to simple crosstabs and correlation coefficients on a subset of the variables, the question arises as to how much information the public should be able to access. At the least, we think that there should be public access versions of all the main NCES data sets, and if time and money permit, these public access versions should allow analysis on as many variables as possible. In this manner, the data that NCES worked long and hard to gather and clean would be made available to as many researchers as possible.

APPENDIX

To help identify the databases referenced in this paper, the following descriptions are provided by the National Data Resource Center:

Schools and Staffing Survey (SASS)

The SASS is an integrated sample survey of public and private schools; school districts; and principals and teachers. SASS was first administered during the 1987–88 school year, and again in 1990–91 and 1993–94. It will be conducted again in 1997–98. SASS consists of eight questionnaires: Public and Private School Administrator; Public School; Public School Teacher; Public School District Teacher Demand/Shortage; and Teacher Follow-up Survey. The following questionnaires were added for the 1993–94 school year: Public and Private School Library Media Center; Public and Private School Library Media Specialist/Librarian; and Public and Private School Student.

National Survey of Postsecondary Faculty (NSOPF)

The NSOPF is a survey of faculty in postsecondary institutions. The survey was initially conducted during the 1987–88 school year and was repeated during the 1992–93 school year. It consists of the following surveys: Institutional, Faculty, and Department Chair.

Common Core of Data (CCD)

The CCD is a set of five surveys sent to state education departments to collect data about all U.S. public elementary and secondary schools, local education agencies, and state education agencies. CCD contains three categories of information: general descriptive information on schools and school districts; data on students and staff; and fiscal data. The descriptive information includes name, address, phone number, and type of locale; the data on students and staff include demographic characteristics; and the fiscal data cover revenues and current expenditures.

High School and Beyond (HS&B)

The HS&B describes the activities of seniors and sophomores as they progressed through high school, postsecondary education, and into the workplace. The data span from the years 1980 through 1992 and include parent, teacher, high school transcript, student financial aid records, and college transcripts, as well as student questionnaires.

National Postsecondary Student Aid Study (NPSAS)

The NPSAS describes all types of postsecondary enrollees, ranging from full- and part-time students who attend private, for-profit (proprietary) institutions to those in prestigious public

universities. Administrative records, with exceptional detail concerning student financial aid, are coupled with student interviews and data from a subsample of parents. Data are available from academic years 1986–87, 1989–90, and 1992–93.

National Longitudinal Study (NLS)

The NLS describes the transition of young adults from high school through postsecondary education and the workplace. The data span from the years 1972 through 1986 and include college transcripts.

National Education Longitudinal Study (NELS)

Beginning with an 8th-grade cohort in 1988, NELS provides trend data about critical transitions young people experience as they develop, attend school, and embark on their careers. Data were collected from students and their parents, teachers, and high school principals and from existing school records such as high school transcripts. Cognitive tests (math, science, reading, and history) were administered during the base year (1988), first follow-up (1990), second follow-up (1992), and third follow-up (1994). All dropouts were retained in the study.

Integrated Postsecondary Education Data System (IPEDS)

The IPEDS surveys most postsecondary institutions, including universities and colleges, as well as institutions offering technical and vocational education beyond the high school level. IPEDS began in 1986, replacing the Higher Education General Education Information Survey (HEGIS), which began in 1966. The components of IPEDS include Institutional Characteristics; Fall Enrollment; Salaries; Tenure and Fringe Benefits of Full-Time Faculty; Financial Statistics; Staff; and Academic Libraries.

Baccalaureate and Beyond (B&B)

Formally known as the Survey of Recent College Graduates (RCG), B&B is designed to analyze the occupational outcomes and educational experiences of bachelor's and master's degree recipients who graduated from colleges and universities in the continental United States. The survey was taken during the 1985–86, 1989–90, and 1993–94 academic years.

Beginning Postsecondary Students (BPS)

The BPS followed first-time beginning students from the 1989–90 NPSAS. NPSAS:90 asked additional questions of students eligible for BPS concerning background and experiences related to completion of postsecondary education. The BPS:90/92 data further describe the experiences during and transitions through postsecondary education and into the labor force, as well as provide information about family formation. Transfers, persisters, stopouts/dropouts, and vocational completers were among those who completed interviews in the first follow-up conducted in 1992. In the second follow-up, conducted in 1994, many will have completed a bachelor's degree as well.

NOTES

1. In response to recommendations of the Financial Accounting Standards Board, a committee made up of members of NCES and NACUBO (the National Association of College and University Business Officers) is working on changes to college and university financial statements, which will go a long way toward meeting these objectives.

REFERENCES

Mortenson, T. October 1995. "Starting Salaries of College Graduates 1947 to 1995." *Postsecondary Education Opportunity*.

National Academy of Sciences, Committee on Science, Engineering, and Public Policy. 1995. "Reshaping the Graduate Education of Scientists and Engineers." Washington, D.C.: National Academy Press.

Discussant Comments

JAMIE MERISOTIS

Let me begin by saying that NCES deserves a great deal of credit for what it accomplishes. As an agency that has been plagued by chronic underfunding, which operates with the federal procurement albatross permanently affixed to its neck and has had to fend off occasional attempts to politicize the Center's agenda and data collection vehicles, I have tremendous respect for the content and the quality of the work that NCES does. This conference, with expert guidance from MPR Associates, Inc., is a good example of the foresight and professionalism exhibited by NCES. I am delighted to be here and am honored by the invitation to participate.

The task before us today is to explore issues related to the national collection of data regarding postsecondary education over the next 5 to 10 years. This suggests that we need to have some sense, at least from a national policy perspective, of what the most important issues will be. So in beginning my comments about the two excellent papers from David Breneman and Fred Galloway and from Michael McPherson and Morton Schapiro, I would like to attempt to predict what those key issues will be. Because of the limited time we have, I will focus on just those issues that concern the federal government's interest in and influence on national data collection in postsecondary education.

First, it seems clear to me that the *federal role* in higher education will be a prominent if not dominant topic of discussion in the next decade. Undergirding this discussion of the federal role will be the central question of who pays for and who benefits from investment in postsecondary education. The personal, social, and economic benefits of postsecondary education will need to be clearly delineated and understood in the policy world in order to constructively engage in this conversation. The federal government already has attempted to address this topic at the K–12 level with the National Education Goals effort. In higher education, I think we will be examining how or if the federal government should play a role in setting goals for higher education; how those goals should be measured; and what happens if those goals are not achieved. I also think that the federal role in defining or delineating the distinctions among collegiate education, remedial instruction, and work force training will be important components of this discussion.

Second, the *level of support* that the federal government should be providing to pay for higher education will also be an important topic. What is the appropriate level of investment in postsecondary education from the federal perspective? What should the relationship be between the federal and state investments in higher education? What linkages, if any, should there be between federal support levels and institutional pricing? These are the kinds of questions that are posed in

both papers and will form the core of the debate about federal support levels in the next several years.

Third, the issue of *program integrity* also will be critical. By program integrity I do not mean the current Higher Education Act usage of that phrase, which seems to confuse concerns about fraud and abuse in federal programs with what are the desired educational outcomes of those programs. Instead, I mean that the integrity of what the programs are supposed to do to influence the educational attainment of students—ranging from access to program completion—will be discussed.

Fourth, the appropriate methods for *regulating* or *deregulating* the federal government's interactions with higher education also will be essential. This is the flip side of the program integrity issue, and is related to determining what aspects of federal regulation might be eliminated without negatively affecting the federal government's legitimate interest in stemming fraud and abuse. Because NCES does not play a direct role in program management, this issue will be put aside for the purposes of this discussion.

Thus, in analyzing these two excellent papers in relation to what will be the most prominent issues of policy discussion in the next several years, I think we have two complementary pieces: the McPherson and Schapiro paper provides a road map for tracking the costs and benefits of postsecondary education over the next several years, which is key to determining what the federal role should be; and the Breneman and Galloway paper provides us with the key stops along the road, thereby helping to define what information we will need in setting federal support levels, and how we can track program integrity by deciding which outcomes of postsecondary education should be measured.

With respect to the particulars of the two papers, I am most compelled by McPherson and Schapiro's clear arguments for good longitudinal data. As the paper carefully points out, longitudinal data provide benchmarks on student attributes in order to examine how college changes the attributes. That analysis is critical to the task of determining who benefits from postsecondary education, which in turn will shape how we define the federal role.

Two specific points contained in the McPherson and Schapiro paper deserve careful consideration. First, I very much agree with the idea of creating a “long-form” IPEDS survey to collect detailed data on educational treatments, for the very reasons described in the paper. Participating institutions must be compensated for this extra effort, however. Second, I share the authors' observations regarding the utility of NCES data for making valid comparisons of higher education internationally. In this age of global economic and social systems, the inability to make reasonably precise comparisons represents one of our greatest shortcomings in national data collection.

The Breneman and Galloway paper contains many excellent suggestions regarding information that should be collected but currently is not. At the same time, however, I am wary of adding to the NCES burden in the absence of new resources. Simply put, I don't believe NCES is capable of doing more with less—that is what they have already been doing for more than a decade.

If new resources are available, I believe that the authors' proposal for a longitudinal database of high school graduates is an excellent idea. This should be a priority in any environment where new resources are available, since this database would allow us to conduct the kinds of seamless analyses of access, persistence, and work force performance of college graduates that we have attempted in the past using multiple, often incompatible, data sets. Such a database would take us a long way toward deciding what is the appropriate level of support from the federal government, and in assessing the integrity of programs with respect to influencing the educational attainment of students.

The recommendation by Breneman and Galloway for more published reports is important. If I have a criticism, though, it is that the reports currently produced under NCES supervision are unnecessarily dull. When every report appears to use the same adjectives—taken, no doubt, from an approved list—and when every report is similarly formatted and printed, a kind of mind-numbing effect can sometimes occur. In my office, we argue about which NCES publication contains certain information. These disagreements often end in frustration, since it is virtually impossible to distinguish among them. (“I think it was the blue book” is a sure-fire way to frustrate an opponent in such arguments.)

The only priority that the authors have identified with which I do not agree concerns research on proprietary institutions. Having spent several years during the 1980s conducting such research, I share the authors' frustration about the lack of reliable, consistent data. Unfortunately, I believe that expanding the sampling frame of various NCES surveys would be a day late and a dollar short. Given that the concept of “institution” is about to be radically transformed as a result of technological changes in pedagogy and educational delivery, focusing on 1980s-era concerns about proprietary schools seems misplaced.

Overall, I believe these two papers provide us with a template for future data collection and analysis, and are extremely valuable in informing key policy discussions over the coming decade. I appreciate the opportunity to comment on the papers and urge NCES to take the authors' recommendations seriously.

JIM MCKENNEY

Historically, postsecondary data collection has focused on traditional college-age students and the structures and procedures of the traditional 4-year college/university environment. As American community colleges have evolved, those historical definitions of postsecondary education have been assessed as increasingly dysfunctional by the 2-year college sector. Yet those definitions persist with surprising tenacity, which is surprising, since the size and growth of the community college enterprise would seem to inherently require a more customized approach. The national network of community colleges today numbers approximately 1,100 institutions in every state. In 1992, these colleges enrolled 5.7 million credit students and conservatively another 5 million non-credit students. The colleges enroll 44 percent of the nation's undergraduates and 49 percent of all first-time freshmen. The average age of a community college student is 29, with females constituting 58 percent of the college enrollment. About 47 percent of all minorities in college attend community colleges, and more than half of higher education students with disabilities attend public community colleges.

It is with that perspective and skepticism that this discussant reviewed the papers by McPherson and Schapiro and Breneman and Galloway. One can almost appreciate that researchers may have chosen to cling to the traditional definitions out of sheer fear of the complexity of the 2-year college sector. By comparison, community colleges may appear to be the moral equivalent to the Balkans for many postsecondary researchers. There is no separate typology for community colleges. Thus, everything is lumped together, undercutting substantially the ability to make finite distinctions. Using the tradition-bound National Center for Education Statistics (NCES) data sources, McPherson and Schapiro assessed the usefulness of these data in tracking the cost/benefit of postsecondary education. These authors argued persuasively that a direct correlation between cost benefits of postsecondary education and classroom activities is impossible to frame—especially, using the national data sets as they are presently constituted.

Under the tight parameters of traditional social research, McPherson and Schapiro suggest that you cannot tease out all of the other possible behavioral explanations for post-graduate performance. They are right, of course. And, their suggestions for merging data sets and seeking voluntary research contributions from individual institutions are great ideas—ideas that would seem to have merit due to the ability of volunteers to drive research to greater detail and at no great cost to NCES. Again the researchers point out the difficulties that come with merging existing data sets—sets that were created for different purposes. Thus, voluntary contributions from institutions or state systems might provide a better picture of the connections between certain desirable causes and effects. We might not have a national picture, but we would have a limited one. This reviewer would only add that such an endeavor should not be attempted without a substantial effort to enlist a respectable sampling of community colleges. Such states as Florida, California, North Carolina, and Illinois have historically maintained extensive data on their 2-year college systems.

A final point needs to be made about the McPherson and Schapiro paper and the concern regarding the uneven fit of traditional research approaches to higher education. During most of the discussion regarding the economic benefits of higher education, one was left with the impression that the researchers had in mind traditional liberal arts majors. What about the measurement of economic benefits that might correspond to graduates of occupational and technical programs at community colleges? For that matter, one could ask the same question about graduates of the professional programs at the university level. One would think that there would be a great payoff in looking at these questions with engineers, nurses, electronics technicians, and accountants. Also, while the authors speak of merging IPEDS and NPSAS, they have given no thought to the potential use of the National Assessment of Vocational Education (NAVE). Again, the railroad tracks might not make an even match, but these data are aimed at assessing occupational education at the secondary and postsecondary level. There just may be some value in looking beyond the traditional college student when it comes to seeking correlations between causes and effects.

Breneman and Galloway have attempted to improve the data collection abilities of NCES by focusing attention on the following six issues: institutional finance, postsecondary assessment, loans and student indebtedness, the school-to-work transition, technological change and distance learning, and the proprietary sector. Again, we have a very compelling argument for new ways of looking at data with an eye to cost effectiveness. For example, the suggestion is made again that IPEDS and NPSAS be connected and that NCES attempt a shorter 2-year cycle.

The authors make the case that institutions of higher education are under substantial financial stress with growth essentially being flat. There is the stated concern that this circumstance has led private institutions to engage in price discrimination through student aid discounting. Thus, private institutions have found a way through public student financial aid to defray their escalating costs in a flat market. This may be okay, but it is a public policy issue that can only be massaged with the existence of confirming data. Alternately, the case is made that financial duress has led the public sector to move toward varying forms of privatization, such as heavier reliance on tuition and private fund raising. It would seem to this reader that this is such a fundamental issue with respect to the true intent of financial aid use and misuse that NCES could hardly ignore the challenge. Heretofore, all of the attention has been focused on student abuse of aid, but the authors are raising a more subtle, but equally important issue.

On the other hand, it is doubtful that financial aid will be a good gauge for financial stress in community colleges. This is not to say that this sector is beyond economic duress. Rather, low tuition and high numbers of part-time students will mean that student aid will be a less robust intervening variable. Community colleges react to economic stress by lowering the full-time/part-time faculty ratio, cutting marginal curriculum/courses, seeking local bond market relief, and seeking infrequent and modest tuition increases. Of course, some community college students will seek recourse in some form of financial aid. The more likely student reaction will be reduced course loads, increased working hours, and the extension of years in college. Hence, the community college and the non-traditional student behave in very different ways from their 4-year counterparts.

The argument is proffered that there is a need for better outcome measures in order to permit individuals and society the ability to calculate rough cost-benefit ratios in making educational selections. The point is made that NCES does a good job in collecting data on degree earners and those with some college experience. However, the data are inadequate for those students enrolling for just a few postsecondary courses. The authors correctly point out that those courses taken for job-related reasons do have a value-added component that is not presently captured. From the standpoint of community colleges, Breneman and Galloway are beginning to spotlight a very large area of concerns surrounding the mapping of community college impact—the tracking of the growing number of non-traditional students, most of them adults, that drop in and out of college as if 2-year institutions were convenience stores. These students, some already having a degree, attend the college for the purpose of attending one or a series of classes in order to achieve a particular skill. Some of these students may achieve a degree over time, but rarely in 2 years and many never intend to earn a degree. Yet, they are there using these institutions in a value-added manner. Hence, community colleges and the night programs at 4-year institutions are becoming as important to the burgeoning number of adult students as they are to traditional college-age students. It seems that NCES must find a better way to map this terrain or forego the ability to comment with authority on this major growth sector of higher education.

Breneman and Galloway make a strong plea that the rising costs of higher education and the concomitant rise in loans and student indebtedness have implications for the larger economy and for student choices. The concern is raised that increasing levels of educational debt mean that students will likely defer other lifetime purchases and that they may, in fact, alter their occupational choices or lifetime goals as a result of debt incurred as students. It is pointed out that the data issue here is

one of timing rather than that of an information vacuum. On the other hand, the authors suggest that the rapid changing nature of work calls for the development of more precise information on career path selection and on follow-up data with both college graduates and high school students moving directly into the work world. In short, the nature of the school-to-work transition has become the subject of such increased concern that NCES should not ignore the need to enhance the database for this purpose.

The point is valid as far as the authors take it. Community college professionals would point out the additional need to track the work-to-school behavior of adults. For these students, starting salary is less relevant than salary increases or job and occupation movement. For that matter, job retention may depend upon the acquisition of a new set of skills. Breneman and Galloway are correct in suggesting better follow-up data for working high school students and college graduates. But, the major story in higher education may be the work-to-school transition of the 25- to 40-year-old cohort.

Breneman and Galloway raise the specter that technological change and distance learning loom on the horizon with major implications for the delivery of instruction, the quality of instruction, and the financing of education. Ironically, it is the speed with which technology is influencing our world and the rapid response of consumers that raises questions about policy making that is dependent on NCES data collection. As stated, it is true that there is a need to assess quality and effectiveness among new modes of delivery, but this reader had the sinking feeling that we were all watching an avalanche in process and no one was sure what to do to avoid being run over.

Finally, the issue of the paucity of data surrounding the "burgeoning" world of the for-profit colleges is of major concern. The authors suggest that these institutions owe much of their financial success to the existence of federal student financial aid, but that these same institutions are not always very forthcoming with the requested information on their effectiveness. As in their earlier point about student aid discounting and privatization, Breneman and Galloway have raised another major policy issue with respect to the complex web of growing interdependence between financial aid and higher education. Federal policymakers cannot begin to address this issue effectively without better data from NCES. It would seem to this reader that this ought to receive the highest priority from NCES as student financial aid is the major federal investment in higher education.

The authors are to be congratulated for their penetrating look at the data sets and their suggestions regarding the applicability of these data to future issues in higher education. Clearly, the issues surrounding student financial aid are critical given budgetary constraints. Moreover, it appears that both papers call for a merging of data sets and a more user friendly timing of data access. All researchers were mindful that the desire to measure the benefit of higher education must be contrasted to the finite resources of NCES. Thus, most suggestions were made with an eye toward activities that sought economies of scale as well as new data yields. This reviewer thinks that NCES has received excellent suggestions from both sets of researchers. The major caveat is a reservation about the applicability of data generalizations to community colleges.

PAULA KNEPPER

I would like to thank the authors of these two papers for providing very thoughtful and complementary perspectives on improving NCES data in the area of postsecondary education. Mike McPherson and Morty Schapiro have presented a very thoughtful and expansive view of the need for longitudinal data at the postsecondary level and beyond. They have suggested several areas in which more information is needed about the college experience itself. They point out that it is necessary to illuminate the "black box" experience in order to more accurately relate education to outcome measures. However, their primary emphasis has been on the need for longer studies of any single cohort, and on the need for "good" longitudinal data.

Similarly, Dave Breneman and Fred Galloway have pointed out six specific areas where additional information is needed concerning postsecondary education. As in the McPherson and Schapiro paper, many of their data needs can only be met with additional longitudinal information. Breneman and Galloway have also provided a set of recommendations on how to achieve much of what is needed with limited resources.

As was mentioned yesterday, education is a very complex process at the K–12 level. Postsecondary education is even more complex; although it serves fewer people, it provides many more diverse experiences and serves a much more diverse population in terms of age and past experiences. Many people continue directly from high school and simply see it as more schooling. These are typically thought of as traditional students. But others continue after a hiatus from education only when they have perceived the need for additional education for a variety of reasons, not the least of which is to enhance their ability to acquire a better job. Some return because they want additional education, though not directly tied to obtaining a specific job. Still others do not complete degrees in the traditional order. For instance, they may return after completion of a bachelor's degree for vocational training of some type, often at the local community college or a private trade school. Others complete a second bachelor's degree instead of, before, or even after completing a master's degree or higher. And these non-traditional students are increasing in number.

Postsecondary education itself also has a split personality of sorts—vocational schools emphasize getting the skills required for a specific job, while collegiate education emphasizes expansion of knowledge not directly tied to a specific job. Galloway pointed this out in his presentation, and it was further emphasized by Jim McKenney in his discussion. The majority of postsecondary students attend either a private trade school or a community college sometime in their educational careers, but not necessarily as the first institution as is so often thought. As Breneman and Galloway have also pointed out, the transition from education and work is no longer neat and clean. McPherson and Schapiro further complicate the picture by pointing out that people in postsecondary education are for the most part there voluntarily, not because the state requires that they attend until a certain age.

These non-traditional patterns are not new—early in my professional career, I hired a programmer who had just completed work for a computer programming certificate at a proprietary vocational school in Northern Virginia. The previous spring, he had completed a bachelor's degree magna cum laude in psychology at a prominent state university, but had not found job prospects particularly promising. Several years after that, he started work on a master's degree in computer

science, and when that was finished, he moved on to become one of the chief developers of the computerized database system used by one of the national grocery chains. This clearly was not the traditional path through postsecondary education, even though he started college right after high school. As a statistics agency, we cannot ignore these different paths through education and work.

We have been encouraged to think in broad terms without regard to money or other constraints. Both of these papers stress the importance of long-term longitudinal data, and both recommend that there be more longitudinal surveys with more frequent re-interviews, and that these be conducted over longer periods of time and include more subjects. But I think our real challenge is to consider the broad data needs and how we might begin to meet them within realistic resources. The suggestion to follow multiple high school or earlier cohorts more often and further through all of the possible education paths is unrealistic—sample size alone would be prohibitive when you think how many 8th graders, for example, you would need to sample to ensure that you had a representative sample of people taking each of the diverse paths through high school and later into and through postsecondary education, some as far as a Ph.D. or similar degree. Even in High School and Beyond, the numbers are too small for reliable analysis even into, much less through, the Ph.D. levels. Thus, it becomes clear that there are two real challenges to NCES:

1) We must keep in mind that even though we think of education as a continuum, in reality, K–12 is very different from postsecondary education in terms of both the "black box" process and in terms of goals and purposes. These differences must be reflected in the data we try to collect, how we collect them, and how we evaluate these data. Completing one level of postsecondary education no longer leads just to either the work force or the next higher level on the education continuum.

2) We must find ways to do more with less. This includes finding ways to reduce the time lag between data collection and data availability, while at the same time ensuring accuracy and completeness.

In order to enhance our surveys, we need to continually be aware of the changes that are rapidly occurring, several of which have been pointed out specifically by Breneman and Galloway—e.g., distance learning, increasing use of and capabilities of PCs, the ever faster expanding knowledge base and related curriculum concerns of what to teach in the time available. New issues are emerging almost daily. Because of these rapid changes, as several speakers pointed out yesterday, information is now also needed by others more quickly if it is to be useful. As a statistical agency, NCES provides data. But we need to make sure we do so in a manner that is consistent enough to evaluate change over time, yet is flexible enough to include new emerging areas and to provide information on their impact.

One way to do this is to be more imaginative in the use of technology. We saw a short demonstration yesterday of how the Web could be used. While this has obviously been put into effect in limited areas, how many of us have thought about its use in these or similar terms?

It has been suggested that we link into databases such as IRS for both student and institutional financial information. This has considerable appeal and could greatly reduce burden and

cost to NCES. In this case, the link would be relatively easy; although the legal hurdles are higher, they are not insurmountable. A minor change in the laws governing IRS release of data and interagency cooperation could make this doable at reasonable cost. However, again we cannot ignore reality—the major impetus to change these laws will have to come from outside of NCES. But doing so would free up time and resources for other data collection and dissemination efforts. A similar case can be made for linkage with other databases, and in fact we do link to IPEDS for institutional information, and to ED Student Aid records for student aid and loan information. As other relevant databases are identified, the feasibility of linkages for data collection efficiency and accuracy should be investigated and implemented as appropriate.

In the area of longitudinal data collection, two seemingly opposing strategies have been suggested: fewer cohorts over longer time periods, and more frequent and overlapping cohorts. Both suggest more frequent re-survey intervals. This later point is the key to obtaining what McPherson and Schapiro refer to as "good" data. As they indicate, longer term surveys provide what no other type of surveys can, an indication of the impacts of experiences over longer time periods. However, given the reality of constant change and the non-homogeneity of postsecondary education, this approach can lead to misinformation as well as "old" information not seen as useful by the time it is available (years after the actual experience of interest). Part of the problem is the constricted sample of postsecondary attenders (a single age cohort rather than the full age mix of postsecondary attenders). The other problem with a longer term study starting in or before high school is that it cannot provide the sample size necessary for accurate evaluation of the various postsecondary experiences, a problem exacerbated by the continual reduction in participation at each higher level. However, this type of survey does help to tie the pieces of more segmented surveys together (as suggested by Breneman and Galloway).

Overlapping surveys, as recommended by Breneman and Galloway, while not providing long-term background information about individuals, would include all types of students at each level. However, this puts much more of a burden on NCES to develop sample designs that allow linkages between the unique surveys. For instance, a relatively small high school graduate sample with a 2-year followup could provide good access and choice information for immediate entrants, but would be too small a sample for postsecondary progress, completion, and postsecondary outcome information. However, a coordinated beginning postsecondary student survey such as BPS includes recent high school graduates as well as late entrants, and provides a full range of undergraduates (both traditional and non-traditional) and information on the undergraduate education. Again, however, too few will continue on to postbaccalaureate education to provide good information concerning education and outcomes at that level. Therefore, a new sample of only recent degree completers, as in Baccalaureate and Beyond, is necessary to provide a sufficient number of students who continue their education in order to obtain information about their experiences and outcomes.

For a system such as this to be useful, however, there needs to be a continuing commitment to conduct these various surveys on an appropriate schedule that, in fact, allows these comparisons between overlapping data collections. In addition, these overlapping surveys need to continue for sufficient time to make outcome comparisons, as McPherson and Schapiro suggest. However, at the frequency recommended by Breneman and Galloway, the data at each stage would be recent enough to not be considered "old," and the more frequent comparison cohorts would provide useful information concerning change.

Also, Breneman and Galloway have made it clear once again that NPSAS is vital and should be more, not less, frequent if it is to be useful to policymakers. They also recommend that the number of students within an institution be increased so that the data could provide useful information at the institution level as well as at the sector and national levels. Currently, only very gross statistics (such as the percentage receiving student aid or a student distribution by family income) can be calculated at the institution level, and not at all institutions. (NCES standards require each calculation be based on at least 30 individuals.) To be able to accurately calculate something like aid packages by class level within an institution would indeed result in a significant increase in student sample size. However, if structured properly, it could also allow both a BPS and a B&B cohort off of the same base-year survey, which they also recommended.

The larger samples that would be required by the recommendations of both sets of authors may not be as onerous as they appear on first blush. With current advances in technology, institutional computer assisted data entry (CADE), department record merges, and so on, this should become easier and less costly. In the same vein, I hope that the next Postsecondary Education Transcript Study (PETS) can be done more electronically than has been the case for the previous ones, and as a result will be more thorough and less costly.

To summarize briefly, these two papers have given NCES a great deal of guidance for the postsecondary longitudinal studies program. Though we have been working toward their suggested goals to some degree, they do provide additional support and guidance in terms of importance, frequency, size, length, content, and use. It is up to us to keep these goals in mind as we refine our data collection activities, and to creatively determine ways to make this set of surveys as useful and timely as has been suggested in these papers.