DOCUMENT RESUME

ED 418 148 TM 028 246

TITLE	Schools and Staffing Survey (SASS): 1995. Selected Papers Presented at the Meeting of the American Statistical Association (Orlando, Florida, August 13-17, 1996). Working Paper Series.
INSTITUTION	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-WP-96-02
PUB DATE	1996-02-00
NOTE	94p.; For related document, see ED 417 222.
AVAILABLE FROM	U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Avenue, N.W., Room 400, Washington, DC 20208-5652.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Elementary Secondary Education; Least Squares Statistics; *Longitudinal Studies; National Surveys; Private Schools; Public Schools; *Research Design; Research Methodology; Responses; *Statistical Analysis; Tables (Data)
IDENTIFIERS	American Statistical Association; *Schools and Staffing Survey (NCES)

ABSTRACT

The papers were presented at the Social Statistics Section, the Government Statistics Section, and the Section on Survey Research Methods. The following papers are included in the Social Statistics Section and Government Statistics Section, "Overcoming the Bureaucratic Paradigm: Memorial Session in Honor of Roger Herriot": "1995 Roger Herriot Award Presentation" (Daniel Kasprzyk, Fritz Scheuren, and Dan Levine); "Space/Time Variations in Survey Estimates" (Leslie Kish); "Out of the Box: Again and Again, Roger Herriot at the Census Bureau" (William P. Butz). The Section on Survey Research Methods is divided into two parts. The first part, "Design and Estimation Issues for School Based Surveys," includes: "Improving the Coverage of Private Elementary-Secondary Schools" (Betty J. Jackson and Richard J. Frazier); "Improved GLS [Generalized Least Squares] Estimation in NCES Surveys" (Steven Kaufman, Bonnie Li, and Fritz Scheuren); "Optimal Periodicity of a Survey: Alternatives under Cost and Policy Constraints" (Wray Smith, Dhiren Ghosh, and Michael Chang); "Properties of the Schools and Staffing Survey's Bootstrap Variance Estimator" (Steven Kaufman); "Discussion" (Charles H. Alexander). The second part, "Data Quality and Nonresponse in Education Surveys" includes: "Assessing Quality of CCD (Common Core of Data) Data Using a School-Based Sample Survey" (Sameena Salvucci, Sandeep Bhalla, Michael Chang, and John Sietsema); "Documentation of Nonresponse and Consistency of Data Categorization across NCES Surveys" (Steven Fink, Mehrdad Saba, Michael Chang and Sameul Peng); "Multivariate Modeling of Unit Nonresponse for 1990-91 Schools and Staffing Surveys" (Sameena Salvucci, Fan Zhang, David Monaco, Kerry Gruber, and Fritz Scheuren); "Evaluation of Imputation Methods for State Education Finance Data" (David Monaco, Stanley Weng, and Frank Johnson); "Discussion" (David L. Hubble). The Section on Survey Research Methods contains "Variance Estimates Comparison by Statistical Software" (Stanley Weng, Fan Zhang, and Michael P.

Cohen). The Social Statistics section contains "Teacher Supply and Demand in the U.S." (Richard M. Ingersoll). Each paper contains references. (Contains 37 tables and 6 figures.) (SLD)

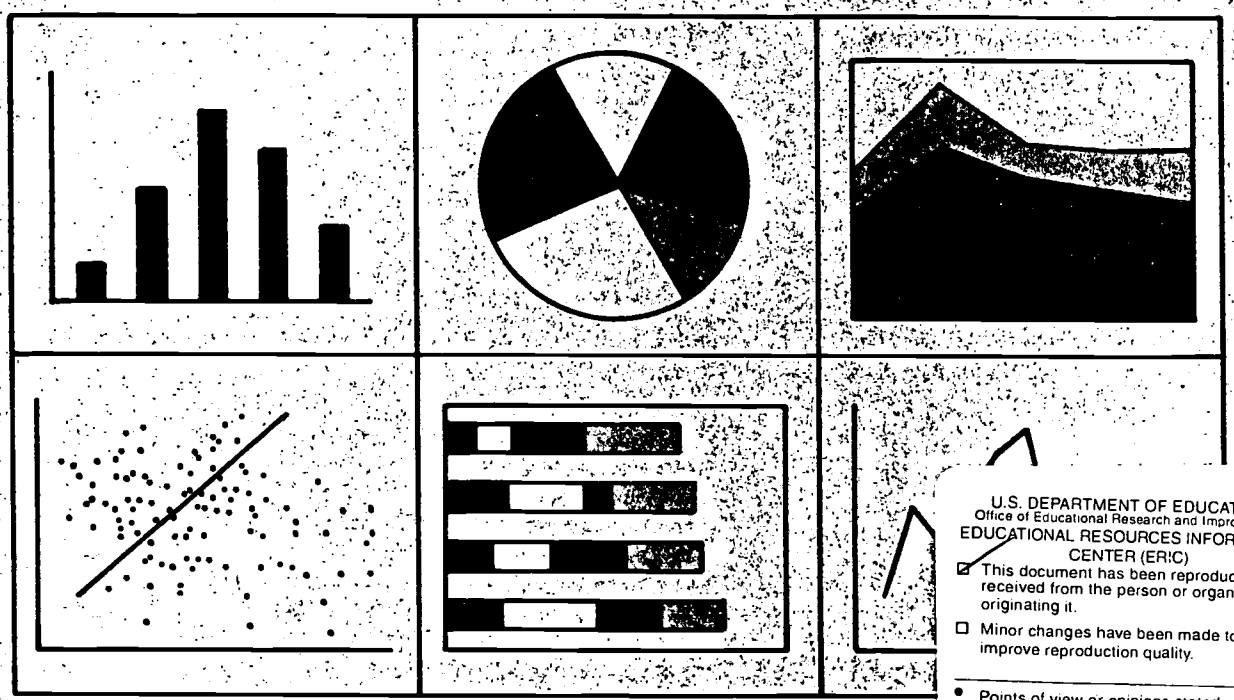
NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Schools and Staffing Survey (SASS): 1995

Selected papers presented at the 1995 Meeting of the American Statistical Association

Working Paper No. 96-02 February 1996



U.S. Department of Education
Office of Educational Research and Improvement

Schools and Staffing Survey (SASS): 1995

Selected papers presented at the 1995 Meeting of the American Statistical Association

Working Paper No. 96-02

February 1996

Contact: Dan Kasprzyk
Elementary/Secondary Education Statistics Division
(202) 219-1588

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Jeanne E. Griffith
Acting Commissioner

Elementary/Secondary Education Statistics Division
Paul D. Planchon
Associate Commissioner

National Center for Education Statistics

The purpose of the Center is to collect and report "statistics and information showing the condition and progress of education in the United States and other nations in order to promote and accelerate the improvement of American education."—Section 402(b) of the National Education Statistics Act of 1994 (20 U.S.C. 9001).

February 1996

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Suellen Mauchamer at (202) 219-1828 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5652.

Susan Ahmed
Acting Associate Commissioner
Statistical Standards and
Methodology Division

Samuel S. Peng
Statistical Services and
Methodological Research

Table of Contents

SOCIAL STATISTICS SECTION AND GOVERNMENT STATISTICS SECTION

Overcoming the Bureaucratic Paradigm: Memorial Session in Honor of Roger Herriot

SECTION ON SURVEY RESEARCH METHODS

Design and Estimation Issues for School Based Surveys

Chair: Paul D. Planchon
National Center for Education Statistics

Data Quality and Nonresponse in Education Surveys

Chair: Daniel Kasprzyk
National Center for Education Statistics

SECTION ON SURVEY RESEARCH METHODS

SOCIAL STATISTICS SECTION

Preface

The fifteen papers contained in this volume were presented at the 1995 American Statistical Association (ASA) meeting in Orlando, Florida (August 13-17). This is the third collection of ASA papers of particular interest to users of NCES survey data published in the ***Working Papers Series***. The two earlier collections were Working Paper 94-01, which included papers presented at ASA meetings in August 1992 and August 1993 and the ASA Conference on Establishment Surveys in June 1993, and Working Paper 95-01, which included papers from the ASA meeting in August 1994.

1995 Roger Herriot Award Presentation
Daniel Kasprzyk, Fritz Scheuren, and Dan Levine
Roger Herriot Award Committee

Last year, after the sudden and unexpected death of Roger Herriot, substantial support emerged to develop ideas to honor his memory. Roger, as most of you know, was an unusually creative person. The old ways and methods of the federal statistical system were never safe from his easy but persistent style.

Last summer, the organizations in which Roger was most active — the Social Statistics Section and Government Statistics Section of the American Statistical Association, and the Washington Statistical Society — established "the Roger Herriot Award for Innovation in Federal Statistics." This award was intended not only to honor his memory but also to recognize individuals who develop unique approaches to the solution of statistical problems in federal data collection programs. The award consists of an honorarium and a framed citation.

In particular, the sponsoring groups intended the award to reflect the special characteristics that marked Roger's career: dedication to

- issues of measurement;
- improving the efficiency of data collection programs; and
- improving and using statistical data for policy analysis.

The Award Committee was composed of three members, each representing one of the sponsoring organizations:

- Dan Levine, representing the Government Statistics Section;
- Fritz Scheuren, representing the Washington Statistical Society; and
- Dan Kasprzyk, representing the Social Statistics Section.

As part of the Committee's work, there was a need to publicize the Award and solicit nominations. The means the Committee employed included:

- announcing the award in the AMSTAT News, Section and Chapter newsletters, the COPAFS newsletter, the Association of Public Data Users newsletter, the Society of Government Economists newsletter, and the Population Association of America newsletter;
- asking the heads of federal statistical agencies to make nominations; and
- consulting with the members of the Office of Management and Budget's Federal Committee on Statistical Methodology (since Roger had been a member of this group).

The Award Committee was pleased to receive nominations for a number of highly qualified individuals, each of whom made a significant contribution to the federal statistical system. We were very pleased with the interest shown in the award — especially, since it is in its first year.

Our selection for the first Roger Herriot Award is a person who has more than met the Committee's expectations:

- as an "innovator in federal statistics," and
- for his personal characteristics and quiet force.

Our recipient is a person who has had a distinguished career both within government, and also within the private sector. His career epitomizes all the qualities set forth for consideration of this award. Many examples exist in his career: implementing coverage improvement research and census data quality evaluation projects; initating methodological projects to improve the Current Population Survey; and developing random digit dialing methods, to mention just a few examples.

The award being given, though, is not a lifetime achievement award, although if anyone deserves such an award our recipient surely does. Rather, this award is being given for not one, but several, recent contributions to federal statistics.

Our recipient this year is Joe Waksberg. Joe is a Senior Vice President and Chair of the Board of Directors for Westat. Joe joined Westat in 1973; before that, he was with the Census Bureau for 33 years, where in his last two years he was Associate Director for Statistical Standards and Methodology.

Joe is being recognized this year for his innovative contributions in three specific areas:

1. improving procedures for sampling rare populations;

2. improving our understanding of random digit dialing methods; and
3. improving our understanding of recall error.

To help you see the value of his contributions, let me briefly describe them.

Procedures for oversampling rare populations — Joe has had a strong interest in improving the efficiency of sampling rare populations. He made innovations in the sample design of the National Health Interview Survey and the National Health and Examination Survey III, where he developed special strategies to more efficiently oversample the minority population. His recent work (Judkins, Massey, and Waksberg, 1992) provides important information on residential concentrations by race and ethnic origin, essential to assessing the usefulness of oversampling geographical areas for minority populations. At this year's meetings he studies the problem of oversampling minority children (DiGaetano, Judkins, and Waksberg, 1995) and he extends his research to investigate the residential concentration of another subpopulation for which oversampling is often required — persons in poverty (Waksberg, 1995a).

Random digit dialing methods — As the developer of the Mitofsky-Waksberg method of two-stage sampling of telephone households(Waksberg, 1978), the standard approach for RDD sampling in the United States, Joe did not have to pursue modifications and efficiencies of the method, but he did so anyway (Waksberg, 1983; Waksberg, 1985; Brick and Waksberg, 1991). Recently, he has contributed to a completely different method of RDD sampling by examining the bias from list-assisted samples (Brick, Kulp, Starer,and Waksberg, 1995). This recent work, in conjunction with his previous re-examination of RDD methods, clearly shows why Joe was chosen for this award. He exemplifies an underlying premise of the Herriot Award—the desire to constantly re-examine standard approaches and find new ones. Joe did this, even though he developed the standard approaches.

Understanding of recall error — The Waksberg and Neter efforts to study recall error in expenditure surveys(Neter and Waksberg, 1964; Neter and Waksberg, 1965) were a landmark undertaking that shed light on the magnitude of various types of memory recall problems, and indicated, as well, procedures for reducing the effects of the recall problems. Joe has continued his interest in this topic; for example, recently he helped design and analyze results from an experiment to measure the direction and magnitude of possible biases from a one year recall for a survey sponsored by the U.S. Fish and Wildlife Service (Chu, Eisenhower, Hay, Morganstein, Neter, and Waksberg, 1992). The results of that experiment had a substantial effect on the redesign of the survey; perhaps, more importantly the work significantly adds to our knowledge regarding respondent bias, when respondents are asked to recall the frequency of activities under varying lengths of the recall period.

While the Committee cites these three instances where Joe's contribution is evident, the statistical community should note that at the 1994 statistics meetings Joe co-authored three papers. This year Joe is also co-authoring 3 papers. Clearly, his contributions to the profession and the federal statistical system grow and grow.

Through all his achievements, Joe has retained his quiet, interested, and unassuming nature - much like Roger. When informed, he was to be given this award, Joe exclaimed: "What did I do to deserve this? " **In our opinion, Joe, you did quite a lot!**

Joe has also been generous of his time to the profession through his work with the ASA Board, to the federal statistical system through his participation in a number of advisory panels, and to his colleagues in their various collaborations. His role as the originator and energy behind the distinguished Morris Hansen Lecture Series is just a recent example(Waksberg, 1995b).

It is a privilege to know Joe Waksberg; I am honored and pleased to present the First Roger Herriot Award for Innovation in Federal Statistics to Joseph Waksberg.

Bibliography

Brick, J.M. and Waksberg, J.(1991). "Avoiding Sequential Sampling With Random Digit Dialing," **Survey Methodology**, 17(1), 27-41.

Brick, J.M., Kulp, D., Starer, A., and Waksberg J., (1995). "Bias in List-Assisted Telephone Samples," **Public Opinion Quarterly**, 59, 218-235.

Chu, A., Eisenhower, D., Hay, M., Morganstein, D., Neter, J., and Waksberg, J.(1992). "Measuring the Recall Error in Self-Reported Fishing and Hunting Activities," **Journal of Official Statistics**, 8(1), 19-39.

DiGaetano, R., Judkins, D., and Waksberg, J. (1995). "Oversampling Minority School Children," **Proceedings of the Section on Survey Research Methods**, American Statistical Association, Alexandria, Virginia, forthcoming.

Judkins, D., Massey, J. and Waksberg, J.(1992). "Patterns of Residential Concentrations by Race and Hispanic Origin," **Proceedings of the Social Statistics Section**, American Statistical Association, Alexandria, Virginia, 51-60.

Neter, J. and Waksberg, J. (1965). **Response Errors in the Collection of Expenditure Data from Household Interviews: An Experimental Study**(Bureau of the Census Technical Paper No. 11). Washington, DC: US Government Printing Office.

Neter, J. and Waksberg, J.(1964). "A Study of Response Errors in Expenditure Data from Household Surveys," **Journal of the American Statistical Association**, 59, 18-55.

Waksberg, J. (1995a). "Oversampling the Low-Income Population," **Proceedings of the Section on Survey Research Methods**, Alexandria, Virginia, forthcoming.

Waksberg, J.(1995b). "Opening Remarks: second Morris Hansen lecture," **International Statistical Review**, 63(2), 119.

Waksberg, J.(1985). "Discussion of Some Research Issues in Random Digit Dialing and Estimation," **Proceedings of the First Annual Census Bureau Research Conference**, U.S. Bureau of the Census, Washington, D.C., 87-92.

Waksberg, J.(1983). "A Note on Locating a Special Population Using Random Digit Dialing," **Public Opinion Quarterly**, 47, 576-579.

Waksberg, J.(1978). "Sampling Methods for Random Digit Dialing," **Journal of the American Statistical Association**, 73, 40-46.

SPACE/TIME VARIATIONS IN SURVEY ESTIMATES

Leslie Kish, Institute for Social Research
The University of Michigan, Ann Arbor MI 48106

1. Introduction

To lay the ground and construct the framework for my main proposals we must first discuss some basic similarities in two major sources of statistical variations: over space and over time. Variations over space serves as the chief justification for the complete decennial censuses, whereas monthly surveys are designed to cover temporal variations, but each of these two neglects variations in the other dimension. I shall try to bridge that chasm, without having to assume exact similarity between the two sources of variations.

Here I advocate the design of cumulated rolling samples with the chief objectives of obtaining and publishing good annual estimates with adequate detail both in the temporal and the spatial dimensions. I expect such annual estimates to provide most of the details for spatial and other domains that decennial censuses now give, but to do it annually with much enhanced temporal effectiveness and usefulness. On the other hand, I also expect that annual estimates and the rolling samples will also satisfy most needs for current data that we now demand from monthly and quarterly releases. Furthermore, quarterly, monthly, and even weekly estimates will also be available. For the central methodological operation I propose weekly national samples designed to be cumulated into monthly, quarterly, annual, and decennial coverages of the entire country and all its constituent domains, spatial and others. The design for temporal cumulation is the novel aspect of the design, because the monthly Current Population Survey (CPS) even now collected within approximately one week. However, for more efficient cumulation the usual overlaps between months would need to be changed to non-overlaps. Overlaps between years need separate consideration.

I have written six papers about rolling samples since 1979, and also designed two of them long before then [Kish 1961, Mooney 1956], but this is my first for the ASA. Those papers were mainly focused on cumulated rolling samples as replacements and substitutes for decennial censuses, or as additions to them. I have not abandoned that goal, but it is not my main emphasis here, because I now prefer to call attention to more frequent estimates, and especially to annual estimates that can provide many of the kind of current details needed for policy decisions.

Even more emphatically, my chief focus is not on the U.S. Census of 2000, which has become a field more suited to legal and public relations talents than to statisticians. Furthermore, I recognize that rolling samples may be even better suited to a smaller country like Canada with 25 million people and 10 million households, with its large monthly labor force surveys; or to Sweden or Hungary with 10 million persons and 4 million households. As for the mere head count of persons in any of those countries, and whether decennial, annual, or weekly, that function may well be inherited by ever improving administrative registers [Scheuren 1991; Redfern 1995]. Registers can provide data that are both timely and detailed, when they are good, as in the Nordic countries. But they cannot provide rich data now or in the future.

We must also consider another basic and important aspect without having time for the attention it truly deserves. Opposite the temporal, I focus here on spatial domains for the sake of brevity and because of the popular attention they receive in official statistics. However, please consider "spatial" as shorthand for all kinds of domains, such as age, sex, occupation, economic, behavioral and all the many domains and subclasses used and presented in surveys. For example, the teen-aged males and females, white and black, are among the most important domains for surveys of unemployment. The cumulations for these other domains may be even more effective than for spatial domains.

2. Statistical Variations Over Time and Over Space

Let me alert you to two important departures from our customary ways of thinking. First, we must admit that these two kinds of variations are not entirely similar either in their essential structures or in survey practice. On the other hand, I shall also note and emphasize some of their similarities and upon those similarities I shall base proposals for altering our views and our treatment of temporal variations.

Second, I ask you take different views of both time and space from those we normally use for our physical world. We think of time as flowing forward evenly and unidimensionally. And we think of variations as occurring chiefly monotonically in a secular straight line or along a logarithmic growth curve; or perhaps in a cyclical variation, governed either by the Earth's daily turns around its own tilted axis; or by its yearly path, on its tilted axis, around our Sun. These diurnal and seasonal cyclical variations are seen in many averages. But in statistical and survey data taken over time intervals, we actually observe mostly random or haphazard variations. This is true of individual blood pressure and blood counts, stock market averages, unemployment rates, air pressure and temperature, etc., etc. The cyclical and secular trends are typically removed by either model-based adjustments, or by taking small time segments (like "strata") or by both. Thus the variations actually observed and used over time intervals is similar to the variations also measured in sampling over space.

Space also has a different meaning here for surveys than the three (or more) dimensional space of physics. It refers chiefly to partitions of the earth's surface into administrative domains like provinces and districts, and into areal sampling units like strata, blocks, and segments. Furthermore, the same concepts can be applied to domains and partitions created by statistical analysis and treated similarly to spatial domains, for example, social economic classes, etc. So that, in contrast with the temporal dimension, "space" and "spatial" can stand as shorthand for other domains covered in cross-section surveys [Kish 1994].

Permit me two side remarks in this admittedly deep discussion. If my "time" and "space" differ so much from the traditional physical concepts, why did I not choose some other terms? Frankly, because I could not think of any. (Perhaps T and S would do better, and this could be called a TS theory!). Second, I have long made a distinction between "proper" and "design" domains and subclasses, used in sample designs (like provinces and districts in area sampling); and "crossclasses" (like age, sex, occupation, behavior, etc) that cut across sample designs [Kish 1987, 2.3]. But I need not expand here on those familiar distinctions.

Thus, despite the physical and philosophical differences between the temporal and "spatial" dimensions (and other domains) we find and can use the statistical similarities we find for most variables in survey situations for spatial aspects also for the temporal aspects. However, we should examine those similarities from four distinct points of view. First, with regard to smooth continuity versus sudden discontinuity: they both exist in both the temporal and spatial aspects. Against the smooth temporal growth curves of peaceful nations, we can counter- pose epidemics (influenza, AIDS), stock market crashes, and sudden weather changes. Against the smooth spatial changes of the Midwest, we pose drastic changes along the Andes and the Rockies, or the drastic social changes found when crossing the Rio Grande between Mexico and the USA.

Second, most people seem to perceive a conceptual difference between temporal and spatial variations. For example, adding regional, provincial statistics into national aggregates and averages appears "natural," but rolling monthly samples into annual or decennial averages seems to run against perceptual walls. We may need a "paradigm shift" to hoist ourselves over that wall [Scheuren 1991]. I believe that this conceptual block is truly less philosophical than psychological and social, conditioned by our long acquaintance with the images of censuses and of monthly survey data.

Third, understanding the similarities may depend strongly on the time interval involved. For example, annual income is a readily accepted aggregation not only for steady incomes but also for occupations with high variations (seasonal or irregular). Averaging weekly samples for annual statistics will prove more easily acceptable than decennial averaging. Nevertheless, many investors in mutual stock funds prefer their ten-year or five-year average earnings (despite their obsolescence) to their up-to-date prior year's earnings(with their high "random" variations). Most people would also prefer a 50 year average "normal" temperature to last year's exact temperature for planning a picnic. There are many similar examples of sophisticated averaging over long periods by the "naive" public. They would also learn fast about rolling samples, given a chance.

Fourth, rolling samples will encounter formidable problems of feasibility. These will differ so much between countries, resources, and the nature of statistics that I cannot discuss this topic both generally and usefully. One difficult example is the "continuing censuses" (decennial) for the USA in 2000 [Alexander 1993]. On the other hand, designing rolling samples for annual statistics for most countries without monthly surveys, may be simple compared to its alternatives.

3. Major Surveys of the National Populations

Figure 1 lists the major types of population surveys now conducted in the USA and in many industrialized countries. Some of these are also conducted in the "less industrialized countries" (LDC's), and decennial censuses cover almost all countries today.

countries. But complete reliance cannot yet be placed on telephones, and therefore area segments are used for frames, or as supplements.

The sampling frames and resources needed for these periodic statistics have also been used as resources and vehicles for other statistical needs (line 7). For example, annual surveys of statistics of education, income, and crime victimization. Also *ad hoc* one time cross section surveys have been collected on many topics. With some modification, weekly samples of 1,000 households, and their 2,500 occupants, are collected and cumulated to 52,000 households, about 130,000 persons yearly [National Center for Health Statistics 1958].

A great gap exists between the complete focus of decennial censuses on geographic / administrative and other domains, with great sacrifice of timeliness, and on the contrary, the complete focus of monthly samples on timeliness, with great sacrifice of domain details. Between these extremes, most statistical needs which are now missing, could be filled with large annual samples. Cumulation and rolling samples are proposed to fill this gap in Sections 4 and 5.

4. Rolling Samples for Annual Statistics

Annual statistics seem neglected now by surveys which concentrate chiefly on decennial censuses at one extreme and on monthly labor force surveys on the other. This seems to be a historical curiosity, due to the success that those two great inventions have enjoyed in our times; and we placed our trust in them -- more or less.

Annual statistics play leading roles in many endeavors: in economic data, in accounting practice, in weather reports, in demographic reports, etc. There are annual social and demographic statistics released in some countries, based on the last decennial censuses with "postcensal" adjustments based on vital and other registers. There are annual fertility and population samples of 1/2000 in China [Li 1985]; and Germany had annual 1 percent counts of the population; but I have made no study of these efforts. An annual sample of 1 percent was advocated for the USA long ago by Hauser [1942]. However, I believe that these yearly snapshots would be more costly, less useful and feasible than rolling samples. The yearly data from 52 weekly samples of 1000 dwellings in the National Health Interview Surveys come somewhat closer to rolling samples [National Cener for Health Statistics 1958], but are not quite that.

To avoid confusion with other methods, I define rolling samples as: *a combined (joint) design* of *k separate* (nonoverlapping) periodic samples, *each a probability* sample with the selection fraction $f = 1/F$ of the *entire population, so designed* that the cumulation of k periods yields a *detailed sample* of the whole population with $f=k/F$. Several feasible modifications can be accommodated within the definition [Kish 1990]:

a) When $k=F$, the cumulated sample yields a complete census with $f=F/F=1$; perhaps decennially.
b) The fixed, constant sampling fraction can be changed from $1/F$ to P_h, perhaps to accommodate with larger P_h, small domains or because of frame problems, etc.
c) Changing the periods and the sampling fractions $1/F$ between periodic waves are both possible, but the population weights for the periods must be considered.
d) It is implicitly assumed that the reference periods of the waves are "mutually exhausting," so that weekly (or monthly) samples refer to the entire weeks (or months). But the reference periods can also be only stematic samples of the periods; for example, one week in the month, as in the CPS sample [Kish 1987, 6.1].
e) For simple and efficient combining we assumed separate samples that are "mutually exclusive" (not overlapping), but overlapping designs can be accommodated with special care and methods.

I propose rolling samples to be collected weekly (or perhaps) monthly to serve simultaneously several major objectives:

I. They can replace the present monthly and quarterly surveys of labor force and/or current population surveys. Countries that have not yet adapted these may now have added incentives for starting them. The multiple objectives of rolling samples can be built into the designs from the start. Countries that have good, large surveys can use those budgets, but may face problems of conversion, because of two main obstacles. Some have large month-to-month overlaps, which may yield some modest gains for some change statistics, such as changes in unemployment. Furthermore in countries with many telephones, later interviews may be cheaper than the first doorstep interviews.
II. Annual statistics based on 52 weekly rolling samples may be the chief product.
III. Decennial (and quinquennial) samples will be based on combinations of annual samples.

IV. Panel studies may also be attached, as discussed in Section 6.

V. The entire operation can also serve as basis for other periodic or one-time surveys.

Thus, the budget of the rolling samples should be compared to the combined cost of all these operations, rather to the cost of only one of these, such as the CPS or the complete census.

5. Rolling Samples for Decennial Censuses

a. I may disappoint some of the audience because I shall say little about rolling samples for decennial censuses, about which I already have several publications. There exists good, current treatments specifically about the "continuous census" for the US Census of 2000 AD, whereas my interests are more general [Alexander 1993, Herriot 1988, Bounpane 1986].

b. Also my methods are less relevant for the simple population counts on the "short" form of the complete census. These concern mostly such problems as the "undercount," the "constitutional requirements," the feasibilities of administrative registers as censuses; and I am no expert in any of these topics.

c. If we aim at "long" form only, the rolling samples need to compete not with the complete count of the entire population over ten years, but only with 5, 10, or 20 percent samples, depending on the country.

d. One aim of the rolling samples will be geographical detail over 10 years of cumulation. For national and large provinces the current annual sample will be preferred usually.

e. The 10 years cumulations can be performed annually, I suppose, and not have to wait for ten year gaps, as at the present.

f. The weights for the 10 years need not all be 0.1, nor 1.0 only for the last year and 0 for others, but monotonically nondecreasing over time:

$$\geq w_t \geq w_{t-1} \geq 0 \text{ and } \Sigma w_t / 10 = 0.$$

g. Will year-to-year overlaps be excluded? Will within year overlaps be excluded? Will they be both replaced by a Split-Panel-Design? (Section 7)

6. Asymmetrical Cumulations

This topic may serve to best distinguish the rational statistical designs that rolling samples can offer from the traditional designs that pass for "common sense." However, I want to emphasize that asymmetrical cumulation (AC) does not depend on rolling samples, and can be applied to other sample designs [Kish 1986]. I refer mainly to the strategy of balancing sampling errors against biases due to obsolescence of data from temporal changes. Take for example the justly famous Current Population Survey of the USA, with monthly samples of about 60,000 households, with twice as many adult persons. Many judge that sample too large because its sampling precision is swamped by structural, temporal nonsampling errors, due to the vagaries of the weather, or the calendar, or other haphazard factors that appear in its monthly news releases. On the contrary, for the statistics of important small domains the sample is too small and the sampling variability is much too great for reliable statistics. Small domains may be either geographical-administrative, such as a state; or they can consist of "crossclasses," such as the Black teenage girls and boys in the labor force. Sampling variability of the statistics is even greater for the many comparisons between statistics of small domains. This is a general problem with applications in many countries and in many subjects and variables [Kish, 1987, 2.1-2.3].

The same periodic surveys must serve both for overall (national) statistics and for domain statistics. Asymmetrical cumulations can best satisfy both needs: frequent (monthly) statistics for the total (national) statistics, but less frequent (e.g., quarterly or annual) statistics for smaller domains. And for these multipurpose aims, rolling samples can serve best.

Three main reasons should lead to asymmetrical cumulations. 1) The principal divisions of most countries tend to vary greatly in size, with ranges of 50 or even 100 to 1; e.g., the states of the USA and Australia, the provinces of Canada and China. Similar variations also exist for other social organizations, like firms, universities, and hospitals. 2) Below the level of the principal divisions, statistics are also wanted for their subdivisions; e.g., counties, districts, etc., which are much smaller and more numerous. 3) Cumulations are often needed for rare items, which can be of three kinds [Kish 1965, 11.4].

7. Panels and Correlations for Rolling Samples from Split-Sample-Designs (SPD)

Panels have nothing to do with censuses, but have a great deal to do with the use of overlapping samples for periodic surveys. Panels denote samples in which the same elements (persons, families, households)

are measured on two or more occasions for the purpose of obtaining *individual* changes. From the mean of these individual changes the net mean population change can be estimated. However, from the net changes of means we cannot estimate (directly) the gross change of individuals. This contrast of population/element change has been variously designated by individual/mean, or gross/net, micro/macro, or internal/external.

Only panels can reveal the gross changes behind the net changes generally (exceptions can be found with strong models) [Kish 1987, 6.2D, 6.4-6.5]. The periodic labor force surveys fail to yield it, because the samples are rotated, and also because households and people change and move.

"Split Panel Designs" (SPD) may be added to rolling samples, as I have proposed [Kish 1987, 1990]. This would displace partial overlaps with two samples: a panel *p* added to the independent rolling samples *a, b, c, d....* Thus the periodic samples will consist of *pa-pb-pc-pd* etc. The size of the panel *p* relative to the independent samples can be varied, but a small ratio, $p/a<1/3$ will usually suffice. This SPD has two critical advantages over the classical partial overlaps. First, it provides true *panels* of elements (e.g. persons or households), which are missing for the moving elements in designs of mere overlaps. However, panels involve following the movers, and thus they can uniquely yield most valuable statistics, which mere overlapping samples of sampling units (e.g. segments, PSUs) fail to yield. Second, in SPD the correlations are present for *all* periods, not only for the pairs arbitrarily designed in the classical symmetrical rotation designs. These overlaps are mostly designed for successive monthly and yearly changes. However, often the most desirable comparisons may not be foreseen in the design, hence the benefits of correlations are absent for them. These comparisons would benefit from the correlations of SPD designs.

Figure 2

Possible Modifications of Rolling Samples

1. Overlaps between samples. Excluded from rolling samples?
2. SPD-Split Panel Design. Panels and overlaps for all periods
3. Oversampling some small domains
4. Undersampling some expensive domains
5. Weighting, e.g. moving averages to favor recent data
6. Over (under) sampling for some periods
7. Synthetic estimation for small areas and periods (SPREE)
8. Other cumulations of F periods; e.g. 52 weeks = 1 year. IS

8. In Conclusion

Periodic surveys are becoming much more widely and commonly used and valued, and I see them as the wave of the future, especially for official national statistics, but also for "unofficial" social surveys.

Up to now, they have been designed especially to defeat trends and particularly for short term differences between collecting periods, such as month-to-month differences.

However, here I urge that periodic surveys should also be considered and designed for cumulations over time to provide more and better data for spatial and domain details. Also that "rolling samples" would provide the best bases for such cumulations. I also urge that annual statistics should be the principal aims of rolling samples, because they can give the best compromise between the needs for better temporal and spatial variations. Finally, decennial censuses, either as samples or as complete counts, can also be based on rolling samples.

References

Alexander, Charles H. (1993). **A Continuous Measurement Alternative for the U.S. Census,** Report to US Census Bureau, also presented at the 1993 meeting of the American Statistical Association.

Bounpane P (1986). How Increased Automation Will Improve the 1990 Census, **Jour. Official Stats.**, 4, pp. 545-553.

Hansen, Morris H., and Hurwitz, W.N. (1946). Sampling Methods Applied to Census Work, in **U.S. Bureau of the Census, The History, Operations and Organization of the Bureau of Census,** Washington: Government Printing Office, pp. 83-94.

Hansen, Morris H., Hurwitz, W.N. and Madow, W.G. (1953). **Sample Survey Methods and Theory,** Vol. I, New York: John Wiley and Sons.

Hauser, Philip M. (1942). Proposed Annual Census of the Population, **Journal of the American Statistical Association,** 37, pp. 81-88.

Herriot R, Bateman DJ, and McCarthy WF (1988). **The Decade Census Program**, US Census Bureau, Internal draft.

Kish, L. (1965). **Survey Sampling**. New York: John Wiley.

Kish, L. (1979a). Samples and Censuses. **International Statistical Review**, (47), pp. 99-109.

Kish, L. (1979b). Rotating Samples Instead of Censuses. **Asian and Pacific Census Forum**, (6), pp. 1-2, 12-13.

Kish, L. (1981). Using Cumulated Rolling Samples. U.S. Government Printing Office, No. 80-52810; 78 pages.

Kish, L. (1983). Data Collection for Details Over Space and Time, *in* T. Wright, ed., **Statistical Methods and the Improvement of Data Quality**, New York: Academic Press, 73-84.

Kish, L. (1986). Timing of Surveys for Public Policy. **Australian Journal of Statistics**, pp. 1-12.

Kish, L. (1987). **Statistical Research Design**, New York: John Wiley and Sons, Chapter 6, Sample Designs Over Time.

Kish, L. (1990). Rolling Samples and Censuses. **Survey Methodology**, (16), pp. 63-79.

Kish, L. (1994). Multipopulation Survey Designs, **Int. Statistical Rev.**, 62, 167-186.

Kish, L., Lovejoy, W., and Rackow, P. (1961). A Multistage Probability Sample for Continuous Traffic Surveys. **Proceedings of the Social Statistics Section, American Statistical Association**, pp. 227-230.

Kish, L., and Verma, V. (1983). Censuses Plus Samples: Combined Uses and Designs. **Bulletin of the International Statistical Institute** 50(1), pp. 66-82.

Mooney, H.W. (1956). **Methodology in Two California Health Surveys**, U.S. Public Health Monograph No. 70.

Moser, C.A. and Kalton, G. (1971). **Survey Methods in Social Investigation**. London: Heineman Educational.

National Center for Health Statistics (1958). Statistical Design of the Health Household Interview Survey, **Public Health Services**, 584-A2, pp. 15-18.

Platek, R., Rao, J.N.K., Sarndal, C.E., and Singh, M.P. (1987). **Small Area Statistics**, New York: John Wiley and Sons.

Redfern, P. (1995). Chapter in this volume.

Scheuren, F. (1991). Paradigm Shifts: Administrative Records and Census Taking. **Statistical Policy Working Paper 2, Seminar on the Quality of Federal Data**. Office of Management and Budget, Washington, D.C.

U.S. Census Bureau (1978). **The Current Population Survey: Design and Methodology**, Technical Paper 40, Washington, D.C.

OUT OF THE BOX: AGAIN AND AGAIN
Roger Herriot at the Census Bureau

William P. Butz [1]
U.S. Bureau of the Census

Upfront, I admit I am a fan of Roger Herriot. I admire who he was and what he did. And, by and large, how he did it. So this discussion is a personal one. I don't offer an exhaustive review of Herriot's life work, nor even a summary. I do offer these observations, though, as an analysis of how he accomplished so much. It is a critical analysis. As I say, I admire how he did it...by and large. I hope in addition to being critical, it is also productive and funny. For Roger was always productive and often funny, although I almost never heard him be personally critical of anyone.

Table 1 shows a list of items. It is noteworthy in three respects. First, it is a diverse list: scientific and technical, statistical and economic, organizational and programmatic.

Second is the importance of many of these items as "hot topics" on today's National policy agenda. This relevance is obvious in many cases. Some specialized knowledge is required to see it in others.

Third, Roger Herriot was instrumental in inventing, discovering or implementing, as the case may be, each item. Not only instrumental: I will develop the argument this morning that Herriot's involvement--No, let me strengthen my case a little--Herriot's engagement in a problem was a sufficient condition for the discovery of a solution and for the programmatic implementation of that solution. This is an extraordinary claim to make about anyone while perusing a list of his or her life activities. To claim it about a federal civil servant working always with and through coworkers and always with or against bureaucratic structures and rules, may be considered foolhardy. In this rare case, I think the statement is supportable. I'll repeat it: Roger Herriot's engagement in a problem was a sufficient condition for the discovery of a solution **and** for the programmatic implementation of that solution.

It is not useful to try to trace Herriot's contributions to these items by way of his publications. Although these run to four pages of citations, I find documentation of only a handful of his contribution in this way. Neither have I discovered how he accomplished so much by considering how Roger thought about a problem. By and large, this is a mystery to me as it was, I think, to him on the occasions we talked about it. James Gleick, in his recent biography of Richard Feynman, quotes Murray Gell-Mann on how Feynman solved problems. Gleick reports that a physicist studying with Gell-Mann at CalTech in the 1950's asked Gell-Mann whether Feynman's own problem-solving methods were the same as the methods Feynman proposed in unpublished lecture notes that were circulating. "Gell-Mann says no, Dick's methods are not the same as the methods used here. The student asks, well, what are Feynman's methods? Gell-Mann leans coyly against the blackboard and says, Dick's method is this. You write down the problem. You think very hard (Gell-Mann shuts his eyes and presses his knuckles parodically to his forehead.) Then you write down the answer." This seemed to be Herriot's general method as well. I don't learn much from it.

What, then, made Roger Herriot so immensely productive? What distinguished him from most everyone else? Certainly, he had a solid background in economics and statistics as well as rich organizational experience, having worked at the Census Bureau 22 years. And he was real smart. But some other people also have these attributes. I see the answer to the puzzle of how Herriot did it less in how he thought about problems than in his attitude toward problems and in his behavior with others in dealing with problems. Herein, I propose, is the key to his sufficiency. These are the attitudes and behaviors that, in my view, fundamentally set Herriot apart:

[1] Associate Director for Demographic Programs. These lightly edited remarks were delivered at the Memorial Session in Honor of Roger Herriot at the Annual Meeting of the American Statistical Association in Orlando, August 16, 1995. I am grateful to many of Roger's colleagues for their enthusiastic ideas for this paper.

1. **He had unflagging optimism** that a solution could be found and an unwillingness--a categorical unwillingness in my experience--to admit that something could not be done, or to accept that assessment from others. Several instances:

- Consider the shambles of federal support for the Survey of Income and Program Participation that threatened hard to leave the new survey stillborn in 1982 after seven years of development. Roger remained optimistic that a package of design features, questionnaire content and early deliverables could be forcefully put on the table by the Census Bureau Director and bought by the other Cabinet departments, OMB and Congress. This did come to pass.
- Herriot was always optimistic that a new system of household income data could be built, based on the three legs of the Current Population Survey, The Survey of Income and Program Participation and IRS records, with CPS providing the timeliness, SIPP the detail and IRS records the accuracy for some items, all modeled together. This has not yet come to pass and shows little life at the moment, but, were he here, Roger would smile and say, "It'll happen" ...then sketch for us how the science, politics and budget will come together.
- One thing Roger and I disagreed about: whether data items not provided by a respondent in a survey but available on some administrative record for that same person should be substituted directly into the statistical record and subsequently analyzed and released. I thought not, Roger thought so, and he was optimistic that it would happen. I had to be constantly vigilant to keep him from getting it sneaked in somewhere.
- Scores of times I have heard Roger drawl out in meetings, "I probably don't know all the details and complications here, but I don't understand how that could take so long (or be so complicated). It seems like all you have to do is..." Boy, a lot of people didn't like to hear that!

2. **Herriot was unwilling to choose between two good things** when there appeared to be insufficient time or resources to do both, or when the two things seem technically inconsistent. Instead, he refused to consider either/or, and displayed a remarkable talent for finding a way to do the essential parts of both, or to do a third and different thing that accomplished the first two plus more. An example:

- Consider, for example, the continuing conflict between the expanded provision of survey microdata to users--especially survey data linked to related administrative records--on the one hand, and the protection of respondents' confidentiality and privacy, on the other hand. This conflict arose 40, maybe 50 specific times in Roger's career, but he refused to view it as a conflict! He proposed and implemented solution after solution in case after case--some solutions statistical, some technical, some bureaucratic, some legal[2]. This reached its zenith at the National Center for Education Statistics where Herriot developed protocols of informed consent, data collection, data matching, record formatting and user contracts that, in my estimation, just barely crawled through the available space between all the constraints.

3. **He continually searched for low-cost spinoffs** from ongoing activities. He asked a whole lot of questions like, "Now that we have TIGER[3], what else can be done with it?" or "As long as we're doing these Statistical Briefs, why not give them to the field representatives so they and the survey respondents can know what's happening with the data?" or "If we're going to use SAS for data analysis, why can't we try it for generalized data processing?" And he formulated answers to such questions.

4. **He didn't wait for a mandate to do something**, or even for permission. "If you see a vacuum of power, expertise, accomplishment, go ahead and fill it." Roger had little patience for those who complain that they can't do something because they don't have the grade or the title or it isn't in their job description. "Don't ask permission," he drawled. "Do it. People will probably be pleased and if some aren't, you can apologize later." If you do need some authority you

[2] The Orlando audience laughed knowingly at this word. I insisted that I meant "legislative," not "legal," but some of them knew how far Roger's ingenuity could extend!

[3] Topologically Integrated Geographic Encoding and Referencing System, the digital mapping system that supported the 1990 census.

don't have, all that is required, Roger would remind me, is for the boss to support you twice in a row on a decision. I have found this a subtle but powerful technique for shifting power around in the bureaucracy.

5. **He begged or stole the time of collaborators** to work out solutions--usually along lines suggested by himself--and to get these solutions implemented in actual programs. Roger Herriot was legendary at getting people who were supposed to be working for someone else to work informally for him instead. Paula Schneider, who succeeded Roger as Chief of the Population Division, told me once, "Roger goes up and down the hall getting Pop Division people to work on his stuff. (But don't worry, I can handle it.)" Jay Waite, Chief of the Demographic Statistical Methods Division, mentioned to me about three months into his then new job, "It says here on the Division roster that Bob Fay is working for me, but as nearly as I can tell, he's really working for Herriot." Dan Weinberg, Chief of the Housing and Household Economic Statistics Division, replied to a request, "Jack McNeil, Enrique Lamas and Chuck Nelson can't do that. They're working on that idea of Herriot's ."

In these dealings, Herriot was the master of unstructured assignments. He would sit down with someone and give them a scrap of paper that made little sense. He would start talking about it. And big things would eventually result. The method of multiple synthetic imputation of occupations between the 1970 and 1980 census occupation classifications got its start this way.

Roger was a productive collaborator (some 20 different co-authors appear on his CV), but he was a truly great instigator, applying to this endeavor prodigious ingenuity and stealth. Two years ago several high officials at the National Center for Education Statistics, important customers of the Census Bureau, let us know that they weren't altogether pleased with the work we were producing for the money they were sending us. Much of the trouble, when together we figured it out, was, as I understood it, that some of our people were actually working hard on a project of Roger's, rather than on what they were being paid to do by his colleagues at the Department of Education!

6. **He was totally unconcerned about who gets the credit**, uncommonly modest. He liked to say of the Population Division staff: "My people make me look good." I must have heard this 30 times over the years. Roger Herriot was a personification of the adage, "There is no end to what can be accomplished if you don't care who gets the credit."

7. **He always did some real work**. Herriot feared losing track of the details. This was a very real concern, often expressed to me. He told me, "You have to know how the work is actually done before you can be effective in changing it." And on other occasions, "They can argue you down, if you don't know what they're actually doing."

By the summer of 1988, Roger had led his small team of co-designers to a fleshed-out description of the revolutionary Integrated System of Area Statistics, which would replace the census long form in 2000 or 2010. He had conspired with Bruce Johnson to use this model as the basis for a series of staff retreats to think about the 2000 census. With each retreat, new criticisms of Roger's plan emerged. The sample couldn't be controlled, the phone numbers couldn't be matched, the estimated coefficients of variation were biased. Several days later, a new paper would emerge. Even the name changed, to the Decade Census Program. Roger took Census Bureau alumni to lunch to try out his program. More revisions followed. Soon, critics and kibitzers couldn't keep up with the flow of paper. Their critiques became passé shortly after they were distributed. For example, in late September 1988, an excellent critique delivered the conclusions of one technical review committee. The critique states up front, "Our comments pertain to the August 16, 1988 draft. Later proposals, including the possible use of mail questionnaires, are not discussed." Multiple proposals in just 45 days! The author of this critique, Chip Alexander, became the principal designer of the Decade Census Program's current powerful incarnation, the Continuous Measurement Program

Charles Darwin, in an 1871 letter to his son, wrote this, which reminds me a good deal of Roger Herriot:

> "I have been speculating at night, what makes a [person] a discoverer of undiscovered things, and a most perplexing problem it is. Many [people] who are very clever--much cleverer than discoverers--never originate anything. As far as I can conjecture, the art consists in habitually searching for causes or meaning of everything which occurs. This implies sharp observation and requires as much knowledge as possible of the subject investigated."

Roger got this knowledge by keeping his hand in. There were always at least a couple of things that Roger knew more about than anyone else.

8. **He worked a problem in his mind continuously**--in the office, in meetings on other subjects, commuting, at home--until he had a solution, and had checked it out with a few other people. Herriot did use Feinman's method: "You write down the problem. Then you think very hard. Then you write down the answer." But that second stage could last days or sometimes weeks.

So there you are. Eight characteristics of Roger Herriot's attitude toward problems and of his behavior with others in dealing with problems that, in my view, account for his success in solving problems and implementing solutions.

His record wasn't perfect. Table 2 lists the items I can come up with that fully engaged his energy but did not come to pass. He liked to tell the stories about these few things that didn't work out. In these stories, the culprit was invariable himself. It was what he overlooked or presented to the wrong person that messed things up. I heard about these failures far more often than the many successes. He was apparently rolling them over in his mind... learning from them.

His methods weren't perfect either. Some of the eight characteristics I've discussed aren't always productive. Here are four others that certainly are not:

1. **Roger was not an effective communicator.** He had only average writing ability and was quite poor in front of a large group. He spoke softly and slowly, frequently haltingly. When he was Director of the Census Bureau, Jack Keane called me into his office one day and said, "Roger is outstanding in so many ways. Why don't you get him to take a workshop in public speaking so he can express himself better in front of a group?" I suggested to Roger that he do this, and he agreed. Every now and then Jack asked me if Roger had taken the course. Finally we gave up. I don't think Roger ever took it. If he did, it didn't do him much good.

In one important circumstance, though, Roger was quite an effective communicator. This was when an argument was developing, when voices rose and tempers frayed. At these times, Roger's voice grew even softer, the words coming even more slowly. He had a calming influence. And it was then, sometimes, that he would float the germ of an idea that might possibly satisfy both sides of the argument at once.

For other occasions, Roger knew his weakness and built up around him colleagues, notably Gordon Green, who were strong public speakers and press briefers.

2. **He was not an effective administrator**. He certainly wasn't good at budget monitoring, administrative reporting or meeting administrative deadlines. What's worse, I suspect he didn't care about these important matters, either, because too often he failed even to arrange for others to do these things and keep him in line.

Altogether, it can be said that Herriot disregarded and frequently disdained bureaucratic procedures. I got fair warning in my first week on the job. My boss was explaining how to write my performance plan for 1983. "Look at Tom Walsh's plan," he advised. "It's a good example of an excellent plan: the right number of elements and good specificity for each. For the other extreme, look at Roger Herriot's." A year later, the Chief of the Budget Division pleaded with me. "Can't you do something about Herriot?!" I got more than a few such requests.

Now it is true that, in those days, we had colleagues who seemed to believe that the core mission of the U.S. Census Bureau was to submit budget documents on time and in prescribed format; to control expenditures carefully; to undertake no hire, no capital procurement, no diversion of work time, unless the budget for these items was absolutely guaranteed to perpetuity. For these people, Roger Herriot did not contribute to the Agency's mission!

Even by legitimate and essential administrative criteria, though, Roger came up short. It hurt his effectiveness and it hurt him, within the organization.

3. **His style and results didn't always please people**. Certainly they didn't please many of our administrative colleagues. Moreover, they could displease technical staff who had worked long and hard on a solution to a problem and on a decision strategy for adopting the solution, only to hear Herriot drawl out some different--and better--solution off the cuff at a meeting late in the game. I well remember one irritated colleague following me to my office after a meeting. "Who the hell does he think he is," I heard, "coming in at the last minute with an idea like that after staff have been working on it for four months?!"

After World War II, General Eisenhower stated three principles he tried to follow in dealing with the enemy. First, never question his motives; in his own mind he thinks he's right. Second, never embarrass him in public; that will only make him fight harder. And third, never cut off his escape route. If you do, you must destroy him and he might destroy you. Give him a way out. It seems to me that Roger Herriot followed this prescription. He was a gentle guy. But some of his ideas were not gentle, and they could and did make their originator unpopular.

4. **His methods don't scale up**. Herriot was like a successful public school principal. He got results by bending rules and stealing the best staff. If everyone tried this, I don't think it would work. (We could certainly do with five or ten times as many trying it, however!)

A physicist colleague remarked about Richard Feynman, "There are lots of people who are too original for their own good, and had Feinman not been as smart as he was, I think he would have been too original for his own good." Likewise, Herriot was original and he was smart. It worked.

Roger Herriot's engagement in a problem was a sufficient condition for finding a solution and getting it implemented. This does not mean that those of us who worked with Roger on something weren't productive, even vital. It means only that if we hadn't been on that job, Roger would have found, somehow, a way around our absence. Herriot was sufficient in this sense, which, I submit, is quite a meaningful one in a bureaucracy. If, then, Herriot's engagement was sufficient, the challenge for those around him became: "How can I get Roger engaged in my problem?" or "How can I keep him from getting engaged in my problem?"--the latter when one already has a solution of one's own or a process of one's own to find a solution.

For many years, some of the most stimulating, the most productive, the most dangerous, the most fun moments at the Census Bureau came unexpectedly when Roger Herriot stood in your office doorway, without knocking, and drawled apologetically, "You gotta few minutes? I wanna show you something interesting."

TABLE 1

ROGER HERRIOT'S MAJOR INNOVATIONS

At the Census Bureau

- Model to simulate taxes in The Current Population Survey.
- Analytical and content aspects of the Survey of Income and Education.
- Methodology for updating income estimate for revenue sharing.
- 1973 CPS - IRS - SSA Exact Match Study.
- March CPS income supplement.
- Matching - based imputation system for growing CPS non-response.
- Income Survey Development Program (ISDP).
- Probability coding for IRS address match to improve population estimates.
- Multiple synthetic imputations method to bridge occupation classifications between 1970-80 censuses.
- Longitudinal definition of households.
- Non cash income estimates.
- Rescue and initiation of The Survey of Income and Program Participation.
- SIPP recurring report series.
- SIPP Executive Committee and decision structure.
- Statistical Briefs.
- After-tax income estimates.
- Publications Opportunity Committee.
- SIPP Longitudinal Research File.
- Decade Census Program (which became Continuous Measurement).
- Idea of a continuously maintained address list for surveys and census.
- SAS analysis of spells and transitions in SIPP.
- SAS as a processing system tool.
- State income estimates by combining CPS years.
- Decennial Access and Profiling System (DAPS-90).

At the National Center for Education Statistics

- Modernized statistical standards review procedure.
- Greatly expanded micro-data dissemination.
- User-friendly data base on the nation's school districts.
- Innovative electronic data dissemination.

TABLE 2

ROGER HERRIOT'S FAILURES

- Proposal for a Census Bureau Visiting Scholars Program.
- Matrix Sampling in the 1990 Census.
- Longitudinal Analysis of Mature Persons ("LAMP") - a long-term panel retirement and aging survey
  Modeling SIPP, CPS, and IRS data in a new system of income and poverty statistics

Improving the Coverage of Private Elementary-Secondary Schools

Betty J. Jackson, Richard L. Frazier
Betty J. Jackson, Bureau of the Census, Washington, D.C., 20233

Key Words: Data, Collection, Evaluation, Education

I. GENERAL

In the mid 1980's, the National Center for Education Statistics (NCES) undertook a critical review and redesign of its elementary and secondary school surveys. This redesign program resulted in the creation of the Schools and Staffing Survey (SASS), an integrated network of surveys that provided data on schools, school principals, teachers, and school districts. The SASS is complimented by the Teacher Followup Survey (TFS), which collects information for a sample of SASS teachers on such topics as the teacher's employment and teaching status, educational activities, and future plans. In addition, the Private School Survey (PSS) was developed as a universe of private schools in the United States.

The SASS consisted of two frames of elementary and secondary schools: public schools and private schools.

A. Definitions:

Private schools are institutions that include any of grades 1-12, have one or more teachers, are not administered by a public agency, and are not operated in a private home.

List Frame is a national coverage improvement operation designed to locate private schools not listed on the private school universe.

Area Search Frame is a coverage improvement operation consisting of an independent search, in a sample of counties in the country, to locate private schools not listed on the private school universe.

School Birth is any school added as a result of updating the universe.

School Death is any school found to be closed as a result of the updating process.

II. HISTORY

A. Private School Universe Creation

Between 1987 and 1994 the Census Bureau conducted four List Frame and four Area Search Frame operations to update the private school universe.

The Private School Universe was created in 1987 to select the private school sample for the 1988 Schools and Staffing Survey (SASS). The base for the private school universe is the Quality Education Data (QED) Inc. list. It is a commercial list of private schools compiled from various sources.

The National Center for Educations Statistics (NCES) purchased the QED list and provided it to the Census Bureau. In an attempt to improve coverage of private schools, the Census Bureau conducted two coverage improvement operations, (1) the "List Frame" and (2) the "Area Search Frame".

B. 1987 Updates to the Private School Universe

Definition: Affiliation Lists are lists of private schools on the rolls of a specific private school association.

1. 1987 List Frame

The first "List Frame" operation began in January 1987. NCES provided the Census Bureau with 22 private school associations. The Census Bureau sent a letter explaining the survey and requesting lists of schools. Four associations requested nominal payment for their lists. The Bureau received 17 of the 22 lists requested.

Once the Bureau received the lists, they were clerically matched to the QED list. This operation resulted in 1,437 adds to the private school universe.

2. 1987 Area Frame

The first area search frame operation was conducted in March 1987 by field representatives (FRs). Ten sources plus the FR's own personal knowledge of the area were used to make independent lists of private schools in the sample counties. The sources were: Yellow Pages (Schools and non Roman Catholic Churches), Catholic Local Archdiocese, Local Government Offices, Local Education Agencies, Milk Companies, Real Estate Agencies, Chamber of Commerce, Fire Inspector, and Health Department.

Next the RO unduplicated the lists within county and matched them to the universe. All new schools were then contacted to determine eligibility.

C. 1989-90 Private School Survey and Updates to the Private School Universe

The first Private School Survey (PSS) was conducted in 1989-90. To prepare for it, the Census Bureau conducted a second coverage improvement operation on the private school universe.

The PSS is a **CENSUS** of private elementary and secondary schools in the country. The purpose of the survey is to:

(1) build a universe frame of private schools that is of sufficient accuracy and completeness to serve as a sampling frame for other NCES private school surveys; and

(2) to generate biennial data on the total number of private schools, teachers, and students.

Approximately 25,000 private schools were contacted in the first PSS.

1. 1989-90 List Frame Operation

The second List Frame operation began in March of 1989. The Census Bureau contacted QED Inc. to obtain an updated list of their schools. Also, the Census Bureau contacted 23 private school associations. Due to budget constraints, we only asked 12 of the 23 associations to send in their lists. The decision on which lists to request was based on the size of the lists. Eight of the 12 associations that sent lists had also sent us their list in 1987. The remaining four associations sent lists for the first time.

This list frame operation was conducted similar to the one in 1987 with some minor changes. For the eight affiliations that provided lists in 1987, we first asked for updates (births and deaths) instead of the complete list. If they could not provide updates, then we took the complete list.

2. 1989-90 Area Search Frame

The 1989-90 Area Search Frame was conducted in October of 1989. It differed from the 1987 Area Search Frame in three distinct ways.

(1) Only five of the ten sources from 1987 were contacted. These sources are: Yellow Pages (Schools and Non-Roman Catholic Churches), Catholic Diocese, Local Education Agency, and Local Government Offices.

(2) The unduplication process (to the universe) was not conducted in the RO.

(3) Schools were screened over the telephone and, if eligible, interviewed at the same time.

D. 1991-92 Private School Survey and Updates to the Private School Universe

The second PSS was conducted starting in the fall of 1991. To prepare for it, the Census Bureau conducted a third coverage improvement operation on the private school universe beginning in the spring of 1991.

1. 1991-92 List Frame

The 1991-92 list frame operation was more extensive then the first two. In 1991 we contacted 44 private school associations, 50 states and the District of Columbia, QED, Inc. and a private vendor, Jostens Education Data, to obtain lists of private schools.

The 44 associations included the associations from 1987 and 1989. Twenty-six of the 44 associations provided lists. We matched and unduplicated all 26 association lists and the lists from the 50 states and the District of Columbia as well as the lists from QED, Inc. and Jostens.

Some lists were available as electronic files while others were in book form or a printout. As in the first two list frame operations, we had to purchase some lists. As in the 1989 List Frame, we requested only births and deaths of schools from the associations. However, all associations sent complete lists.

This operation yielded 7,552 adds to the universe before mailout (6,267 from states, 959 from the affiliations, 20 from QED Inc. list and 306 from the Jostens). There were 385 schools that overlapped between the four sources.

2. 1991-92 Area Search Frame

The 1991-92 Area Search Frame began in September of 1991. This provided more time to gather and unduplicate lists of private schools and to match the schools to the universe. We wanted to have the operation completed in time for the birth schools to be interviewed during the nonresponse followup phase.

As in 1989 five sources were used to obtain lists of private schools. The difference between the two years was mostly in the check-in and keying procedures.

E. 1993-94 Private School Survey

The third PSS was conducted starting in the fall of 1993. To prepare for it, the Census Bureau conducted a fourth coverage improvement operation that began in the spring of 1993.

1. 1993-94 List Frame

The 1993-94 list frame operation was done in two parts. Association and QED Inc. list updating was done in time to use for the 1993-94 SASS sampling operation. We matched and unduplicated these lists with the 1991-92 PSS universe. These lists yielded 927 births before mailout: 919 from association lists and 8 from the QED Inc. list.

The state list updating operation was done in time to get the birth schools on the private school universe for the 1993-94 PSS. We matched and unduplicated these lists with the 1993-94 SASS universe. This yielded 2,172 births before mailout.

2. 1993-94 Area Search Frame

As in the previous area search frame the FRs contacted five sources plus used their own knowledge to obtain lists of schools in sample counties in their area. The matching, keying and unduplicating operations were centralized in the Indiana processing office, enabling us to maintain better control.

In addition to obtaining the lists of private elementary and secondary schools the FRs also sent in lists of nursery schools, daycare centers and pre-kindergarten schools. These schools/programs were used to help develop an early childhood care frame. The remainder of this paper will discuss the analysis of this 1993-1994 list frame and area frame updating operation, but it will not discuss the early childhood care frame.

III. GOALS/OVERVIEW OF THE 1993-1994 FRAME UPDATING ANALYSIS

We will determine the characteristics of the list frame and area frame by religious orientation (Catholic, Other Religious, Nonsectarian), school level (elementary, secondary, combined), and total student enrollment, school type, and minority student population percentage.

We will determine the effect of the adds on private school characteristics, such as religious orientation, school level, and enrollment, school type, and minority student population percentage. The statistic of interest in this analysis is the percentage of the universe estimate of each characteristic that is represented by the adds (i.e., the numerator will be either the list frame or area frame adds estimate of the characteristic and the denominator will be either the list frame universe (original universe plus adds) or the entire PSS universe estimate of the characteristic). We will show how the universe benefits from the adds in general and by school characteristic.

By answering the following questions, we will identify which sources (states, associations, and QED) of lists provided us with the most up-to-date and complete information about the types of school births we need.

(a) Which source was most effective?

(b) Which source provided the largest quantity of eligible or in-scope additions to the private universe?

(c) Which source provided the eligible or in-scope additions with the highest interview rate?

(d) Which source provided the largest quantity of ineligible or out-of-scope additions?

(e) Which source had the highest out-of-scope rate?

(f) How did these results compare the results with those from the 1991 analysis?

IV. ANALYSIS OF LIST SOURCES FOR ADDITIONS TO THE 1993-94 PRIVATE SCHOOL UNIVERSE

There were three main sources of lists that we contacted when it was time to update the private school universe. These sources are the states (including the District of Columbia), twenty-four of the largest private school associations, and QED, Inc.

A. HIGHLIGHTS

(1) All birth schools on the QED list were found on other lists. We could have eliminated the QED list for the 1993-1994 operation.

(2) The fifty states and D.C. provided 70% of the total additions to the private universe during the 1991 update. Among the individual state lists 60% of the state additions came from Utah, Georgia, Nevada, Wyoming, California, Connecticut, North Carolina, North Dakota, Arizona, Vermont, District of Columbia, Delaware, Florida, Michigan, and Alabama. These states were listed in order of effectiveness (highest rate of in-scope births to lowest rate of in-scope births compared to what was on each list).

(3) Twenty-one of the twenty-four association lists requested provided additions to the private universe. Their contribution to the private universe is on a smaller scale than the state lists.

B. State Lists

At the national level, the state lists have contributed more to the in-scope, out-of-scope, and interview rates than either the association or QED lists. Sixty-five percent of the 2,288 in-scope adds came from the state lists. Eighty-five percent of the 811 out-of-scope adds also came from the state lists. The two main out-of-scope reasons from state lists are "School Closed" and a category that included reasons such as duplicate, PK only, and school merged. The interview rates for the individual schools for the in-scope additions coming from the state lists was 83% (a decrease of 12% from 1991).

The contributions made by the updating operation differed by state. When we rank the states from most effective to least effective, we find the following results. At least 7% of the schools from each of the top 16 states were in-scope births. After the lists were matched to the current private universe, the top sixteen states account for 55% of the state additions. Approximately 2/3 or more of the schools from each of these 16 states' additions were eligible or in-scope with four exceptions: Maine at 46%, Arizona at 33%, Delaware at 37%, and Alabama at 59%. Of these in-scope schools, each state had approximately an 85% interview rate with three exceptions: Maine at 50%, California at 70%, and Delaware at 55%. Thus, in general these states provided quality additions as well as a large quantity of additions.

For the remaining 35 states, their contribution was less relative to the overall total of state additions. Less than 7% of the schools from each of these lists were in-scope births.

C. Association Lists

35% of the 2,288 total in-scope adds are from association lists. 15% of the 811 total out-of-scope adds came from this source. The two main out-of-scope reasons for affiliation lists are "School Closed" (47%) and "Don't Know" (30%).

The top five association lists are the most effective ones. They alone account for 75% of the association

additions. The lists from these associations provided good quality additions as well as a large quantity.

Each of the remaining fifteen association lists were less than 10% effective (i.e., less than 10% of the schools from each of these lists were in-scope compared to the total on the list). However, the importance of these lists to these associations outweighs the fact that they provided a small quantity of additions.

D. Quality Education Data List

The original QED list only provided school births. There were 39 school births. Only 8 were left after clerical unduplication with the existing universe.

Less than 1% of the 2,288 total in-scope adds are from QED. Similarly, a small percentage of the 811 total out-of-scope adds come from this source. The only out-of-scope reason is "Don't know".

E. List Overlap

We updated the private school universe with affiliation and QED lists for the 1993-1994 SASS private school sample. We then updated the universe with state lists for 1993-1994 PSS. Thus, there is no evidence of overlap between state and affiliation lists.

For example, suppose that "ABC" elementary school was added to the universe as a result of the affiliation updating operation for SASS. Now suppose that "ABC" elementary school was on a state list. Because this school was already on the universe, it would not have been counted as a birth from the state list updating operation.

F. Summary

In general, the 1993-94 interview rate among the individual states and affiliations is lower than that for 1991-92.

The total number of births from the association lists in 1993-94 is slightly smaller (919) than that of 1991-92 (959).

The total number of births from the state lists in 1993-94 is drastically smaller (2,172) than that of 1991-92 (6,267). The difference in these figures could be attributed to the way in which the updating operation was done (refer Section II.E for an explanation).

V. ANALYSIS OF THE CHARACTERISTICS OF LIST FRAME ADDS AND THEIR IMPACT

A. HIGHLIGHTS

Other Religious adds make up the largest percentage additional students, teachers, and graduates across all religious orientation categories. The exception is for schools where Nonsectarian adds make up the largest percentage.

Combined school adds make up the largest percentage of additional schools, students, teachers, and graduates across all school levels.

Updating had a big impact on Nonsectarian and Other Religious schools, but very little impact on Catholic schools.

Updating had the biggest overall impact on combined schools although the impact on elementary and secondary schools was significant as well.

Updating had the biggest impact (on all variables) on the smallest schools. With the exception of graduates in Catholic schools, impact decreased as the size of the school increased.

B. Characteristics of Adds

1. General

Other Religious adds contributed 1,169 schools (58.6% of all school adds). This was followed by 709 Nonsectarian school adds (35.6%) and 116 Catholic school adds (5.8%). This pattern for schools across religious orientation is similar for students, teachers, and graduates.

Elementary school adds contributed 936 schools (46.9% of all school adds). This was followed by 854 combined school adds (42.8%) and then 205 secondary school adds (10.3%).

This pattern for schools is different across school level for students, teachers, and graduates (when valid). Combined schools contribute more than elementary schools to the total of the adds.

2. Enrollment

Small schools contribute more significantly to the list frame adds than the larger ones. The overall percent contributions for schools for each of the size categories for the list frame adds schools are as follows: 0-75 students: 68% (68% of the adds are schools with less than 75 students), 76-150 students: 18%, 151-225 students: 6%, 226 + students: 8%.

In general these percents hold true (in magnitude and direction) for each religious orientation and school level. The exception is the Catholic schools where the larger schools contribute a greater number or adds than the smaller schools.

3. Minority Student Percentage

The overall percent contributions for schools for each of the minority student percentage categories for the list frame adds are as follows: less than 6%: 33% (33% of the adds are schools with less than 6% minority students), 6% to less than 21%: 28%, 21% to less than 51%: 18%, 51% or more; 21%.

In general, the above pattern holds true (in magnitude and direction) for each religious orientation and school level. The exceptions are secondary schools where each category for the adds contributes approximately 25% and nonsectarian schools where the schools with a larger minority student percentage contribute more significantly to the adds.

4. School Type

Regular elementary/secondary schools make up the vast majority of the list frame adds at 61% (61% of the list frame adds are regular schools). Alternative school adds contribute 17% to the total adds followed by Special Education schools at 12%. Each of the other three school types (Montessori, Special Program Emphasis, and Voc. Tech.) contribute less than 5% each.

The exceptions to the above pattern are secondary and nonsectarian schools where special education schools contribute the most.

C. Impact of Adds on Private School Characteristics

1. General

The list frame adds represented 8.3% of schools, 3.7% of students, 5.2% of teachers, and 2.7% of graduates on the universe. Nonsectarian led the way with 14.6% for schools on the universe, followed closely by Other Religious at 10.7%, and Catholic considerably smaller at 1.4%. These percentages were reduced somewhat for each religious orientation when you look at students, teachers, and graduates. However, the general relationship seen for schools still holds. These percentages ranged from 5% to 9% (of students, teachers, and graduates on the universe) for Other Religious; 5% to 8% (of students, teachers, and graduates on the universe) for Nonsectarian; 0.5% to 1.5% (of students, teachers, and graduates on the universe) for Catholic.

The school grade level percentages indicated that the list frame updating had a substantial impact on improving the coverage for all three school grade levels. Combined schools led the way with 12.2% for schools, followed by 8.6% for secondary schools and 6.4% for elementary schools. As was seen for religious orientation, these percentages were reduced somewhat when looking at the other statistics (i.e., students, teachers, and graduates).

2. Enrollment

The enrollment percentages showed variation and reflected a strong inverse relationship between the size of the school and the impact of the updating operation on improving the coverage for the different enrollment categories. The smallest schools (0-75 students) led the way at 16.8% indicating that the small schools were greatly impacted by the updating operation. The second smallest group (76-150 students) of schools showed a 7.3% impact, followed by 3.4% for the group of schools that had 151-225 students and 2.2% for the largest schools (226 + students). The pattern for enrollment percentages for students, teachers, and graduates is very similar in both magnitude and direction to that for schools.

3. Minority Student Percentage

The minority student population percentages showed a slight variation between the percentage of minority students at the school and the impact of the updating operation on improving coverage of the universe for the different categories. Schools with a large population of minority students (51% or more) led the way with an 11.0% impact. In other words, the updating operation resulted in 11% of the schools on the 1994 PSS universe having a minority student population of at least 51% that would not have been on the universe if the updating operation had not been done. As the percentage of minority students at a school decreases, so does the impact on the universe.

In general, the same pattern can be seen for secondary and combined schools as well as other religious and nonsectarian schools.

4. School Type

Regular elementary/secondary school adds contribute more to the list frame adds (61%) than the other five school types combined. Their impact (6.2%), however, on the list frame universe of this school type is the smallest of the six school types. In contrast, Vocational/Technical schools make the smallest contribution (.3%) to the list frame adds, but they have the largest impact (51.1%) on the list frame universe of this school type.

VI. ANALYSIS OF THE CHARACTERISTICS OF AREA FRAME ADDS AND THEIR IMPACT

A. HIGHLIGHTS

Other Religious adds make up the largest percentage of additional area frame schools across all religious orientation categories.

Combined school adds make up the largest percentage of additional area frame schools across all school levels.

Area Frame updating had a big impact on Nonsectarian and Other Religious schools, but very little impact on Catholic schools.

Area Frame updating had the biggest impact on combined schools although the impact on elementary and secondary schools was also significant.

Area Frame updating had the biggest impact on the smallest schools.

B. Characteristics of Adds

1. General

Other Religious adds contributed 1,286 schools (63.5%) of all school adds in the 1994 PSS area frame updating operation. This was followed by 671 Nonsectarian school adds (33.1%) and then 69 Catholic school adds (3.4%).

Combined school adds contributed 1,003 schools (49.5%) of all school adds in the 1994 PSS area frame updating operation. This was followed by 904

elementary school adds (44.6%) and then 119 secondary school adds (5.9%).

2. **Enrollment**

Small schools contribute more significantly to the area frame adds than any of the larger ones. The overall percent contributions for schools for each of the size categories for the area frame adds schools are as follows: 0-75 students: 74% (74% of the adds are schools with less than 75 students), 76-150 students: 16%, 151-225 students: 5%, 226 + students: 5%.

In general, these percents hold true (in magnitude and direction) for each religious orientation and school level. The exception is the Catholic schools.

3. **Minority Student Percentage**

Schools with a low minority student population (less than 6%) contribute more significantly to the area frame adds than any with larger ones. The overall percent contributions for schools for each of the minority student percentage categories for the area frame adds are as follows: less than 6%: 46% (46% of the adds are schools with less than 6% minority students), 6% to less than 21%: 27%, 21% to less than 51%: 14%, 51% or more; 13%.

The above pattern holds true (in magnitude and direction) for other religious schools and elementary and combined schools.

4. **School Type**

Regular elementary/secondary schools (60%): contribute more significantly to the area frame adds than the other school types combined. Alternative/nontraditional schools follow distantly with a 17% contribution. The other four school types (Montessori, Special Program Emphasis, Special Education, Vocational/Technical) each contribute less than 10% to the area frame adds.

C. Impact of Adds on Private School Characteristics

1. **General**

The area frame adds represented 8% of the schools on the 1994 PSS universe. The area frame updating had a substantial impact on improving the coverage of Nonsectarian and Other Religious schools -- increasing them by 12% and 11% respectively. The impact on Catholic schools was minimal at 1%.

On the other hand, the area frame updating had an impact on improving the coverage for all three grade levels -- combined schools: 13%, elementary schools: 6%, secondary schools: 5%.

2. **Enrollment**

The enrollment percentages showed variation and reflected a strong inverse relationship between the size of the school and the impact of the updating operation on improving the coverage for the different enrollment categories. The smallest schools (0-75 students) led the way at 15.6% indicating that the small schools were greatly impacted by the updating operation. The second smallest group (76-150 students) of schools showed a 6.7% impact, followed by 2.6% for the group of schools that had 151-225 students and 1.2% for the largest schools (226 + students).

3. **Minority Student Percentage**

The impact for each of the minority student population percentage categories is similar. Schools with a small population of minority students (less than 6%) led the way slightly with a 9% impact (9% of the schools on the 1994 PSS universe having a minority student population of less than 6% would not have been on the universe if the updating operation had not been done). The impact for schools in the remaining categories is as follows: 6% to less than 21%: 8%, 21% to less than 51%: 7%, 51% or more: 7%.

4. **School Type**

The area frame adds were made up mostly of regular elementary/secondary schools (see Section V.A.4). However, their impact on the private school universe was only 6%. In other words, 6% of the schools on the 1994 PSS universe were represented by regular elementary/secondary area frame adds schools. Area frame updating had a substantial impact on improving the coverage of Montessori, Special Program Emphasis, Vocational/Technical, and Alternative/Nontraditional schools increasing them by 21%, 21%, 38% and 17% respectively.

VII. CONCLUSION

We should continue to collect lists of private schools from all the states in the future.

We should also continue to collect lists of private schools from the associations in the future. The association lists do contribute to the universe on a smaller scale than the state lists. Requesting these lists may do more than just update the universe. List requests from associations may promote good public relations with the association heads and they in turn may encourage participation among their member schools.

The list frame updating operation continues to be effective in improving the coverage of private schools.

Since area frame updating estimated that we're missing 8% of the universe, we need to continue this area frame updating to achieve a more complete private school universe.

Updating operations are especially needed for improving coverage of small schools, Other Religious and Nonsectarian schools, and non regular types of schools.

IMPROVED GLS ESTIMATION IN NCES SURVEYS

Steven Kaufman, National Center of Education Statistics: Bonnie Li, Synectics: Fritz Scheuren, The George Washington University
Bonnie Li, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd. #305, Arlington VA 22201

KEY WORDS: Generalized Least Squares, Winsoring Weights, Reducing Mean Square Errors

For the first time, in 1993-1994, the private school component of the Schools and Staffing Survey (SASS) and the Private School Survey (PSS) are being fielded in the same year. Even though these two surveys measure some of the same variables, the results between the surveys will not agree.

PSS and SASS both measure numbers of schools, numbers of teachers, and numbers of students. Conventional simple or raking ratio adjustment procedures could be used to adjust sample weights so that the SASS estimates agreed with the much larger PSS for each of the three totals separately. Such approaches do not work, though, if the weights are to be adjusted so that **all** three SASS estimates agree simultaneously.

As we reported at last year's meetings (Holt et al., 1994), Generalized Least Squares (GLS) techniques is an alternative that offers promise. While the asymptotic properties of GLS and GLS-like estimators are attractive, their finite sampling properties are not necessary desirable. To avoid some of the operational concerns with GLS procedures found in the 1990-1991 experiment, our plan for the 1993-1994 surveys is to follow a three-step process:

1. For the largest schools, GLS reweighting will not be carried out; instead, a direct use of the PSS cases is to be attempted where, through statistical matching of SASS with PSS, the SASS data will be added to one or more of the PSS observations in what is called a "mass imputation" procedure.

2. To improve further the adjustment process, a multivariate ratio adjustment (like in Olkin, 1958) is to be made within moderately sized domains of SASS -- before the GLS procedure is undertaken.

3. Only then will the resulting new SASS weights be carried forward to a GLS estimation step along the lines described in the next section.

Our expectations of these modified procedures were both that they would lead to improvements in SASS mean square error and that operational difficulties would be lessened. The partial results obtained so far bear this out.

Generalized Least Squares (GLS)

For NCES Private School Surveys, the Generalized Least Squares (GLS) techniques advocated by Deville and Särndal (1992) can be used, as in Imbens and Hellerstein (1993).

To discuss the basic algorithm employed in Generalized Least Squares, it is necessary to define some notation; in particular --

w_i is the original SASS Private School base weight for the ith SASS observation, $i=1,\ldots,n$.

t_i is the SASS total of teachers for the ith SASS observation, i=1,...,n.

s_i is the SASS total of students for the ith SASS observation, i=1,...,n.

N is the total estimated number of schools, as given by PSS.

T is the total estimated number of teachers, as given by PSS.

S is the estimated total number of students, as given by PSS.

In reweighting SASS, three constraints are imposed on the new weights u_i,

$$\sum u_i = N$$

$$\sum u_i t_i = T$$

$$\sum u_i s_i = S$$

For our application, the new weights u_i, subject to these constraints, are to be chosen (as in Burton 1989) to minimize a loss function that can be written as the sum of squares

$$\sum (u_i - w_i)^2$$

This is perhaps the simplest and most straightforward loss function that might be chosen. Motivating it here is outside our present scope; except to say that the sensitivity of the results to the loss function chosen (e.g., Deville and Särndal, 1992 and Deville et al., 1993) seems not to be too great (but this is, in part, an application issue and will be among the areas for future study). Anyway, the usual Lagrange multiplier formulation of this problem yields after some algebra that the new weights are of the form

$$u_i = w_i + \lambda_1 + \lambda_2 t_i + \lambda_3 s_i$$

where the λ_i are obtained from the matrix expression

$$\underline{d} = M\underline{\lambda}$$

with the vector $\underline{d}$ consisting of three elements, each a difference between the corresponding PSS and SASS totals for schools (first component), teachers (second component), and students (third component); in particular

$$N - \sum w_i$$

$$T - \sum w_i t_i$$

$$S - \sum w_i s_i$$

where the summations are over the SASS sample observations and the quantities: N, T, and S are known PSS totals for schools (N), teachers (T), and students (S) respectively. The matrix M is given by

$$\begin{matrix} n & \sum t_i & \sum s_i \\ \sum t_i & \sum t^2_i & \sum t_i s_i \\ \sum s_i & \sum t_i s_i & \sum s^2_i \end{matrix}$$

and $\underline{\lambda}$ is the vector of unknown GLS adjustment factors obtained from

$$\underline{d} = M\underline{\Delta}$$

Notice that the M matrix is based solely on the unweighted sample relationships among schools, teachers, and students. This is not an essential feature of our approach; a weighted version of the M matrix could have been used -- with, of course, a corresponding change in the loss function to be minimized.

Olkin-like multivariate ratio estimation

An old idea of Olkin(1958) forms a starting point here. Assume we have a total τ, to be estimated from a sample. Olkin proposed a multivariate ratio estimator of the form Y composed of a sum

$$Y = \sum a_i R_i X_i$$

where the a_i are positive and add to 1, the X_i are known outside totals and the R_i are conventional ratios estimated from the sample of τ and X_i.

How cast the Olkin procedure in the PSS and SASS setting? The multivariate ratio Olkin proposed could, in principle, consist of any number of ratio estimates being added together and averaged in some way by the a_i. Note that in our application there are only three outside totals: X_1 for schools, X_2 teachers and X_3 students -- so the expression has been simplified for this analysis.

For this paper, the a_i are simply chosen to be equal to one-third; however, a more natural approach would be to select them so as to minimize the variance of Y. Given the complex sample design of SASS, though, this has been left for the future.

In principle, an Olkin adjustment to the original weights could be produced within whatever domain is desired; then in order to determine the "new" weight for that domain, all the cases would be adjusted such that they would have new weights

$$u_i = Rw_i$$

where the overall ratio R is obtained by taking Y and dividing it by the corresponding estimate obtained from the original sample.

The intuition is that if the Olkin estimation was carried out for small (appropriate) subdomains, then there would be a direct benefit from this step in those subdomains. Further, the overall PSS/SASS differences would shrink appreciably, minimizing any harm that GLS might do. To try something to check these intuitions, it might be enough to use our greatly simplified Olkin-like approach over suitable subdomains (leaving for later, as already mentioned, a way to choose the a_i to minimize the variance of the estimator).

PSS and SASS Data for 1993-94

As noted earlier, it seems natural to use the PSS figures for schools, students, and teachers as the standard and to adjust the SASS estimates correspondingly; that is what we have done here. To fix ideas and to simplify our discussion, only private Nonsectarian Regular Schools will be examined. There were two basic steps taken which are listed below.

1. Based on an initial visual inspection, we identified about a dozen large schools for which some form of mass imputation, rather than reweighting might be the adjustment of choice.

2. With the remaining SASS sample schools and the remaining universe of PSS schools (See figure 1 below), we then calculated Olkin-like factors by school size to begin the adjustment process.

Figure 1. -- PSS and SASS originally weighted estimates compared

	SASS	PSS
Schools	2524	2186
Teachers	52868	49587
Students	514569	463263

3. Then, a GLS adjustment followed, using the adjustment formula, shown below of

$$u_i = w_i - 0.758 + 0.04006t_i - 0.0032s_i$$

The large negative value for $\lambda_1 = -0.758$ meant that for small schools (with only a few teachers) the possibility of very small weights existed. Similarly, for large schools (with many students) the possibility of negative weights existed (since λ_3 is negative too). Our examination of the weights showed, in fact, that three were negative and three very small.

Another look at data plots identified these schools as cases that were away from the basic scatter -- so we excluded them as well. Another 20 or so small schools had weights (between about 0.2 and 0.7). For these schools, we employed a simple winsorizing routine (and added +0.5) -- so that when subjected again to the GLS algorithm they would not be unacceptably small.

Redoing the Olkin and GLS steps with this slightly smaller set of SASS sample cases yielded an acceptable result -- no negative cases and none that were judged to be too small.

Evaluation of Adjustments

There are two ways we will evaluate our results. Each represents an alternate course of action:

1. One possible course of action might be to do nothing. Here we will compare our method to the original SASS weighted results.

2. Another course of action might be to carry out a simple GLS adjustment, without also introducing Olkin-like factors. Here we will be comparing the Olkin-GLS weights (and estimates) with what would have happened if only a GLS estimate had been attempted.

To be consistent with what has been done already, we look only at the SASS sample cases that were finally subjected to an Olkin GLS weight adjustment. Figure 2 displays the original, unadjusted GLS and Olkin GLS weights in the form of a scatterplot matrix. As can be seen, for these cases

- Visually, the three sets of weights appear close; however, at the bottom of the standard GLS weight distribution, there are about 30+ negative weights.

- Notice also that the regression means differ overall too. The standard GLS mean is closest to the original, since it does not adjust the weights separately by school size; also the fit between the Olkin GLS is somewhat poorer than is true for the unadjusted GLS, again for the same reason.

The real test of the methods is how close they come to improving not only overall totals but also the totals by school size. To examine this, a comparison was made for SASS schools, teachers, and students by school size as a percent of the corresponding PSS total. While not

uniformly better, the Olkin GLS method demonstrated considerable superiority, suggesting we are on the right track.

Conclusions and Areas for Future Study

The work done so far on intersurvey consistency is gratifying in that a clear improvement has been obtained. There are many issues to face, though, as we try to learn more. Among these are

- Can we find a better, more systematic way of handling outliers (e.g., negative and small weights) ahead of time?
- Using mass imputation is only mentioned in the paper. How would that work in these two NCES surveys?
- Can we unite the Olkin and GLS techniques into a single adjustment (as the theory seems to suggest)?
- What about integrating still other information from PSS into SASS (say, information on Community type)? Via a raking version, perhaps?
- Is there a way to calculate variances for an Olkin GLS estimator that is not any more computationally intensive than for the current SASS estimator?
- What about other GLS loss functions? Minimizing percent differences in the weights rather than absolute differences?

The above gives you an idea of some of the issues that will be on our "What next" list. So stay tuned!

Afterword

We would like to thank Chip Alexander for his insightful discussion comments on this paper. His own research, albeit in another setting, certainly parallels ours. We are also grateful for the two references he mentioned that we had not seen: To his own 1990 work appearing in the ARC Proceedings and to the paper by Jayasuriya and Valliant, given in Orlando Thursday, after our paper was delivered. Both will be of help in handling our list of "What Nexts."

References

Burton,R.(1989), Unpublished Memorandum, National Center for Education Statistics.

Deville, J.C., and Särndal, C.E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association,* 87, 376-382.

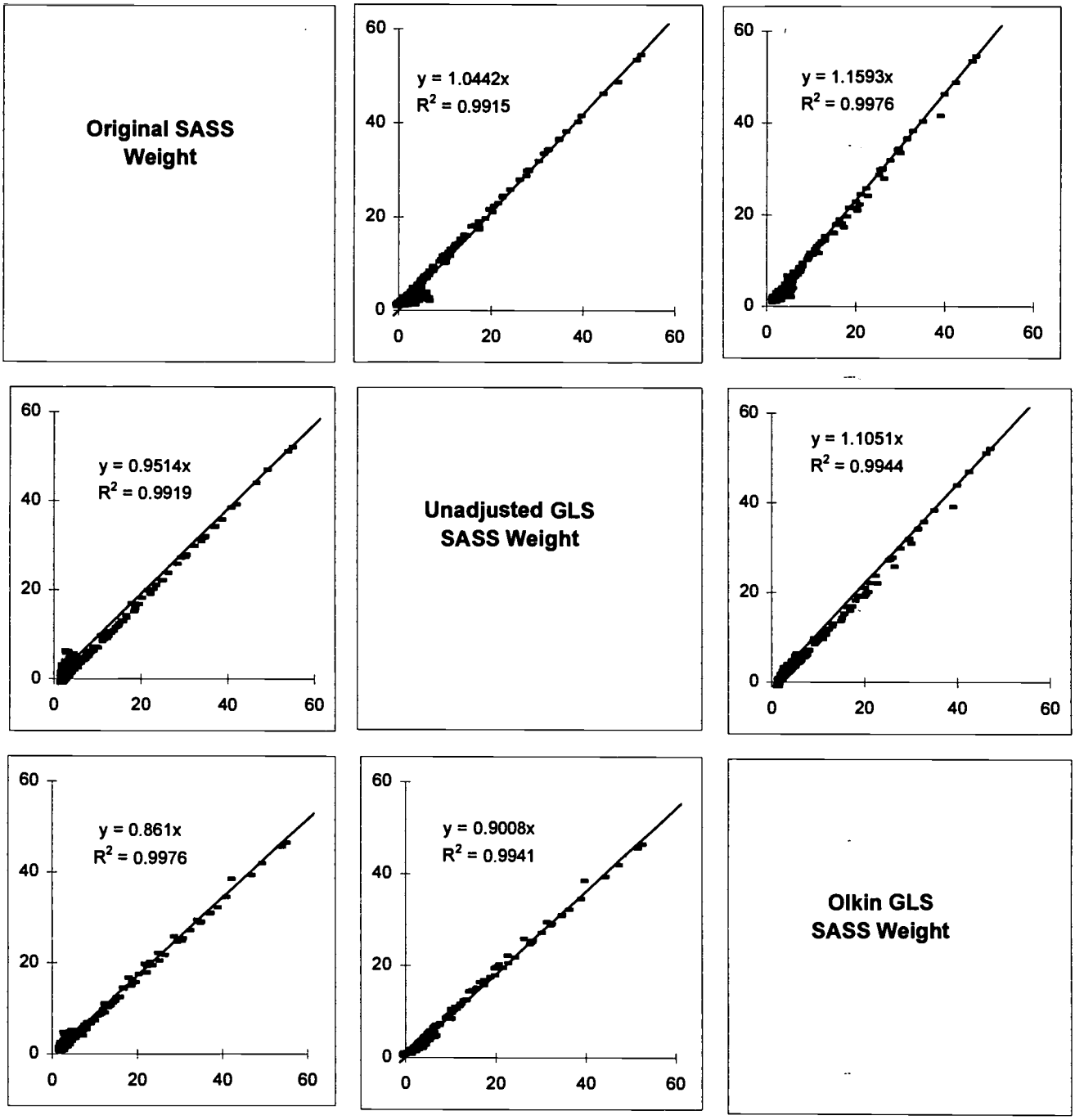
Deville, J.C., Särndal, C.E. and Sautory,O. (1993), "Generalized Raking Procedures in Survey Sampling," *Journal of the American Statistical Association,* 88, 1013-1020.

Holt, A., Kaufman, S., Scheuren, F. and Smith, W. (1994), "Intersurvey Consistency in School Surveys," Paper presented at the 1994 Joint Statistical Meetings, Toronto, Canada.

Imbens, G.W. and Hellerstein, J.K. (1993), "Raking and Regression," *Discussion Paper Number 1658*, Cambridge, MA, Harvard Institute of Economic Research, Harvard University.

Olkin, I. (1958), "Multivariate Ratio Estimation for Finite Populations," *Biometrika*, 45, 154-165.

Figure 2--Nonsectarian Regular
School weights, Unadjusted GLS, and Olkin-GLS SASS Compared



SOURCE: U.S.Department of Education, NCES, Private School of Schools and Staffing Survey:1993-94, Private School Surveys, 1993-94

OPTIMAL PERIODICITY OF A SURVEY: ALTERNATIVES UNDER COST AND POLICY CONSTRAINTS

Wray Smith, Dhiren Ghosh, Michael Chang, Synectics for Management Decisions, Inc.
Wray Smith, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd #305, Arlington, VA 22201

KEY WORDS: Data user needs; Indirect estimators; Loss functions; Probable-error models; Repeated surveys; Small area estimates; Statistical policy issues; Structural time series modeling

This paper is a progress report from a series of ongoing studies related to constrained optimization of the periodicity of school-based surveys -- that is, considering a range of choices of sample size and intersurvey timing intervals subject to a set of external constraints and programmatic goals for the fulfillment of data user needs. Ghosh *et al.* (1994) presented our general approach to these questions via a family of "probable-error" models with joint consideration of sampling error, data deterioration, and cost. There we addressed some of the tradeoffs, under a given multi-year budget for fixed and variable survey costs, between more frequent data collections with smaller sample sizes at each collection and less frequent data collections with larger sample sizes at each collection.

We now give more explicit attention to the statistical policy issues that arise when a set of survey redesign options confronts the policymaker with the possible adoption of "indirect estimation" methods for some subnational or subdomain estimates while, say, retaining "direct estimation" methods for national-level statistics and for the larger states and larger analytic domains of interest. The statistical policy framework we adopt here is in the spirit of the "recommendations and cautions" set forth in *Indirect Estimators in Federal Programs* (Subcommittee on Small Area Estimation, 1993).

Schools and Staffing Survey

Our work has been specifically directed toward techniques that may lead to future redesign options for the Schools and Staffing Survey (SASS). SASS has been developed and sponsored by the U.S. National Center for Education Statistics (NCES) and is conducted for NCES by the U.S. Bureau of the Census. As stated in Bobbitt *et al.* (1995), "SASS is an integrated survey of public and private schools, school districts, principals, and teachers. It was conducted first during the 1987-88 school year, again in 1990-91 and 1993-94, and will be conducted at five-year intervals thereafter. SASS is a mail survey that collects public and private sector data on the Nation's elementary and secondary teaching force, aspects of teacher supply and demand, teacher workplace conditions, characteristics of school principals, and school policies and practices...."

The shift from three-year intervals to five-year intervals is understood to be the result of current and foreseeable budgetary resource constraints for federally-sponsored education surveys and does not rule out consideration of a range of design or redesign options for SASS in the 21st century. Electronic recordkeeping, new data collection technologies, and near real-time data processing capabilities may well open up new design options for school-based surveys.

Partial Redesign of a School-Based Survey

The first three rounds of SASS were conducted at three-year intervals and the intention was that each data collection would have a sufficient sample size to permit statistical estimates to be made for most public school variables and school types at the geographic level of individual states. After data collection and analysis of the 1987-88 SASS, it became evident that "(1) state estimates from the states with smaller populations had higher than expected standard errors, (2) state estimates from the states with larger populations had lower than expected standard errors, (3) state elementary and state secondary estimates could not be made except for the largest states, and (4) the overall national estimates had much lower than expected standard errors" (Kaufman and Huang, 1993). In view of these findings, the design for the 1990-91 SASS was changed to reduce the sample sizes for the largest states and increase the sample sizes for the smallest states. The result was that direct estimates for 1990-91 (and 1993-94) are available for individual states for most school and teacher variables for elementary and secondary schools separately -- and, in most cases, for combined public schools (with grade spans of grade 6 or less to more than grade 8). The quality of national-level estimates was not degraded appreciably by these reallocation steps. Producing separate estimates for elementary and secondary schools was a major objective and hence a major change in the sample allocation was felt to be justified.

Direct and Indirect Estimators

This example serves to illuminate a design and estimation challenge for school-based surveys such as

SASS. The present policy-and-practice setting for SASS is that only "direct estimates" (and their associated estimated standard errors) will be published by NCES in its official publications.

NCES has a broad legislative mandate to "collect, analyze, and disseminate statistics and other data related to education in the United States and other nations." Other federal statistical agencies operate under somewhat different or additional legislative mandates. For example, the Bureau of Labor Statistics (BLS) prepares monthly employment and unemployment estimates for some 5,300 geographic areas, including "...subcounty areas for which data are required by legislation." Since 1989, using data from the Current Population Survey (CPS), BLS has been publishing monthly *direct* sample survey estimates of employment and unemployment for the 11 largest states as well as for Los Angeles and New York City. BLS also publishes monthly *indirect* estimates for the 39 smaller states and the District of Columbia.

"The method used to provide [these] monthly state estimates [for the smaller states] is based on the time series approach to sample survey data. Originally suggested by Scott and Smith (1974), this approach treats the population values as stochastic and uses signal extraction techniques developed in the time series literature to improve on the direct survey estimator." ... "The signal is represented by a time series model that incorporates historical relationships in the monthly CPS estimates along with auxiliary data from the Unemployment Insurance (UI) and Current Employment Statistics (CES) programs. The time series model is combined with a noise model that reflects key characteristics of the sampling error to produce estimates of the true labor force values. This estimator has been shown to be design consistent under general conditions by Bell and Hillmer (1990) and is optimal under the model assumptions." See Chapter 5 in (Subcommittee, 1993); also see Tiller (1992).

A similar approach was taken in Ghosh *et al.* (1994) which assumed, for one model, that there is an underlying stochastic process that is observed periodically by the repeated survey data collections and that this process can be modeled as an ARIMA(0,1,1) time series process observed with sampling error. The formulation of the model is based on a general modeling procedure set forth in Smith (1980) and Smith and Barzily (1982) using Kalman filter concepts." Average cost as a function of sample size and intersurvey time interval (in years) is minimized by a numerical search procedure for a hypothetical survey with given cost coefficients and known noise covariances, yielding a jointly optimal solution for sample size and intersurvey interval. Available methods for the analysis of repeated surveys are summarized in Appendix A of the present paper. The Smith-Zalkind-Barzily (S-Z-B) approach is described in Appendix B. An extension of Ghosh's probable-error model paradigm to an assumed random walk process is outlined in Ghosh (1995).

Possible Enhancement of SASS Estimates

Assume a simple vector autoregressive process that evolves in discrete time at one-year accounting intervals. The vector process may involve a potentially large number of variables that may be observed through data collections at the level of local public schools. A few core variables are selected for observation through two different series of repeated surveys. The first observation series is assumed to be the ongoing annual data collection that is known as the Common Core of Data (CCD). The CCD system covers all public schools in the U.S. and is carried out within States by State education agencies (SEAs). The second observation series is assumed to be the public school component of SASS, for which three rounds of data have now been collected at three-year intervals. SASS covers a sample of public schools with some overlap schools in successive rounds of the survey. SASS also covers a sample of private schools, but these are not considered here.

Both the CCD and SASS series collect data from individual schools on such school variables as grade-by-grade enrollment, number of teachers, ethnic and gender components of enrollment, and number of students eligible for or receiving free lunches. In addition to such common or "core" variables, SASS collects data on such variables as the number of students served by Chaper 1 services, the number of K-12 (Kindergarten through grade 12) teachers who are new to the school this year, the number of K-12 teachers who left the school between October 1 of last year and October 1 of this school year, and the number of K-12 teachers who have a degree beyond the bachelor's degree.

We are currently exploring the possible dependence of components observed in the SASS series, but unobserved in the CCD series, on the observed components in the CCD series. For this purpose we may fit a set of equations in structural time series form (cf. Harvey, 1989) with a signal modeled with components (possibly time varying) that include a Regressor component, a Trend component, and an Irregular component. There is no Seasonal component

since the established accounting period for school-based reporting is annual. If the explanatory power of the CCD regressor variables turns out to be weak, such a finding would support a more frequent SASS data collection. If the dependence turns out to be nontrivial, this finding would support, within limits, a less frequent SASS data collection.

In the course of this work we expect to apply the estimation methodology for short time series set forth in Anderson (1978) for AR(1) processes and extended by Azzalini (1981), and Shumway (1988). We refer to this foundation as the A-A-S approach and will be attempting to connect it to the S-Z-B approach summarized in Appendix B.

Ghosh *et al.* (1994) demonstrated how to determine the optimum periodicity of a survey if the process model is known and is fairly simple (e.g., AR(1), ARIMA(0,1,1) or the Random Walk model). SASS data has been collected only three times; therefore, it is not feasible to fully determine the process model from the SASS data alone. But CCD, which is collected annually, has been in operation for several years and is a complete census. For selected SASS variables not included in CCD, we intend to develop linear models consisting of CCD variables as the candidate explanatory variables for the selected SASS variables in each year of SASS data collection. Such a linear model is like a newly constructed variable; let us call it M. The variable M is constructed entirely of CCD variables and thus is defined for each unit (school) in CCD. We may then use Anderson's method to obtain estimated autocorrelation and partial autocorrelation functions over appropriate subdomains of units of CCD. From these, we can estimate the process model for M. We can then use the available SASS data and the model for M to estimate a model equation for SASS. If this model turns out be a simple process we can then apply the following cost/error principles in our search for an optimal periodicity.

Direct and Imputed Costs in Choice of Periodicity

Any formalization of the problem of seeking an "optimal" choice of survey interval and survey size must account for the fixed and variable costs of operating a system of repeated surveys, such as SASS, as well as imputed costs due to increasing errors in the estimates as sample size is reduced and out-of-date estimates are used. In a recent book on survey errors and survey costs, Groves (1989) provides an up-to-date review of the kinds of considerations which should go into creating cost-and-error models for surveys, with particular emphasis on household surveys. Currently there is no comparable work on cost-and-error modeling for surveys of institutions such as schools.

In the case of SASS, there is an ongoing, more-or-less fixed annual cost of maintaining the core elements of the SASS system whether or not a survey is conducted in a particular year. Some costs might be regarded as either fixed or variable. Among these are the costs of updating list and area frames, with special emphasis on updates immediately preceding each wave of data collection. In this paper we lump such costs with the fixed annual costs of maintaining institutional memory for all aspects of SASS, making evolutionary design changes in coverage and content to be incorporated in the successive waves of data colletion, and conducting ongoing research in support of SASS processing and estimation procedures.

In addition to the directly measurable dollar outlays associated with maintaining and operating the SASS system, it is possible to include imputed dollar costs to represent the loss or penalty which is incurred by public and private users as a result of using outdated survey data. Smith and Zalkind (1978), Smith (1980), and Smith and Barzily (1982) used such an approach, formulating an imputed loss associated with the use of imprecise estimates from an observed economic process where the objective was the allocation of public funds on the basis of such estimates. This approach of Smith, Zalkind, and Barzily involves a framework in which knowledge of the state of a socioeconomic process is characterized as the level of a stock of information (an equivalent sample size on hand). The S-Z-B approach requires a policymaker to select a scale factor or equivalence to characterize the "cost of not knowing" in dollar terms so that the imputed cost or loss can be combined in the same formulas with the dollar outlays. See Appendix B for additional discussion of the S-Z-B approach.

Appendix A: Methods for Repeated Surveys

Since the early papers of Scott and Smith (1974) and Scott, Smith, and Jones (1977), there has been a renewed and growing interest in the application of time series methods to survey data. An excellent review article, Binder and Hidiroglou (1988), may be found in volume 6 of the *Handbook of Statistics*. This review and the papers by Bell (1984), Bell and Hillmer (1990), and Tam (1987) provide a balanced account of time series approaches, including state-space modeling and Kalman filter techniques, for use with data from

repeated surveys. Although most statisticians are now aware of the time series methods of Box and Jenkins (1970), who provided an understandable, systematized approach to model identification, estimation, and forecasting, many survey statisticians are still unaware of the potential of the time series methods for improving estimation with survey data in the sense of minimizing mean squared error. The key principle in the time series approach is that there is information in the time series structure of an observed process which may be used to make better estimates by combining information from past data collections with the new information from a current data collection than would be made if the current data were to be used alone.

Signal Extraction, Kalman Filters, and State Space

Classical survey estimates are made under assumptions that the observed variables, whether of labor force, or school enrollment, or other socioeconomic phenomena, have values that are fixed but are observed with sampling error (and possibly nonsampling error). The time series approach regards the process variables as stochastically varying over time and the identification problem is to find a parsimonious time series model evolving in discrete time such as one-year intervals, that will capture the main features of the underlying process sufficiently well. For univariate processes it has often been found to be quite satisfactory to fit a model with a very simple structure, such as an autoregressive model of order 1 or 2.

In the case of surveys of schools and similar institutions, the natural accounting period is the school year, so that within-year changes are of secondary interest and seasonal effects do not arise. Linear trends are easily incorporated in Box-Jenkins type models and can, if desired be factored out by taking first differences of successive observations. Thus the trend component may be accounted for separately. One useful model that is related to classical exponential smoothing is the Box-Jenkins ARIMA(0,1,1). It is mildly nonstationary and can "wander" up and down; in one sense the current process value serves as a "local mean" for the process as time moves ahead one step and the process noise term kicks the process up or down a bit. In classical Box-Jenkins modeling it is assumed that the process is observed without observation error.

Borrowing from the contributions of R. Kalman in the control engineering literature in the 1960s, the Scott-Smith time series approach utilizes a two-equation setup in which there is a process equation which represents the evolution of the underlying (unobserved) process through time. The second equation, the observation equation, consists of the sum of the underlying process variable and an observation noise term. The noise term in some simple models may just represent the sampling error. In other cases it may have a structure of its own. The state of the process may be represented by a vector with two or more components, representing, for example, the levels of two or more process variables such as number of teachers and number of students at a school, or in an aggregate of schools within a state or other subnational grouping.

The classical Kalman approach assumes that the variance-covariance (V-C) matrices are known and time invariant. In real world settings, the V-C matrices will not be known and will have to be estimated from the data. Furthermore, they will not necessarily be time invariant. These complications have led to the formulation of extended Kalman filters which, although theoretically sound, place an estimation burden on the available data and may lead to inconclusive results. Also, it is somewhat awkward to try to accommodate nonlinear features such as the presence of level-dependent variances. For example in a set of elementary schools arranged by size within one state, the variance in enrollment or in number of teachers will typically depend on the size of the school and hence the number of teachers. This is easier to capture using one of the model types known as state-dependent models and in particular with a class of models known as bilinear models (see Smith, 1994).

Cost Models with Fixed and Variable Costs

We assume that data users will keep on using the data obtained from the most recent past survey until a new survey is undertaken and the newly collected data are processed and released to data users. Thus, if the inter-survey period is long, "deterioration" of the data, if it is of considerable magnitude, could affect the quality of decisions made by users. On the other hand, if the survey is undertaken frequently, the costs of conducting the survey, of analyzing the data, and of response burden may be judged to exceed the benefits achieved in using fresh data.

Typical analyses of cost-benefit tradeoffs tend to focus on the best use of a fixed resource amount over a time period that would include two or more survey data collections. The present budgetary restrictions for the 1990s are such that the "fixed" resource amount may be arbitrarily depressed and may

overconstrain any realistic formulation of the optimization problem.

The usual cost model for a sample survey assumes a start-up cost ($= C_0$) and a per unit (ultimate sample unit) cost ($= C_1$). Thus, the total cost is represented as $C = C_0 + nC_1$. However, the start-up cost may be dependent on the periodicity. We represent it as C_0^k (where k is the periodicity) which may be regarded as increasing with increasing intersurvey interval; i.e., the start-up cost is higher if the interval is three years than if the interval is two years.

We usually assume that total resources for a multi-year time period are fixed. The different possible periodicities spend these total resources in different ways. This assumption then determines the possible sample sizes every time the survey is undertaken corresponding to different periodicities. A modified approach would be to use similar models but to attempt to take explicit account of the fact that total resources may be arbitrarily reduced by external constraints and formulate the decision problem within that framework.

Appendix B: The S-Z-B Optimization Tools

In Smith (1980) the concept of "equivalent sample size" was adapted to a reformulation of the optimal filter theorem for a scalar (single variable) model of an evolving process observed at discrete points in time. The development was as follows:

Consider a repeated survey system in which the process state is represented by the scalar state variable $x(j)$ evolving as a scalar random walk in discrete time, $x(j) = x(j-1) + w(j)$, where $w(j)$ is the process noise term, with scalar survey measurements $y(k)$ given by $y(k) = x(k) + b(k)$, where $b(k)$ is the measurement noise term and the sample size at each survey time k is the scalar quantity $n(k)$ and the sample noise variance $B(k)$ is given by $B(k) = R / n(k)$ with R as the assumed known constant unit measurement noise variance. The Kalman gain in the optimal filter theorem then becomes

$$K(k) = C(k|k\text{-}t)\,[\,C(k|k\text{-}T) + B(k)\,]^{-1}$$

$$= [C(k\text{-}T)|k\text{-}T) + TQ]\,/\,[C(k\text{-}T|k\text{-}T) + TQ + R/n(k)]\,,$$

which is of the same form as the exponential smoothing parameter in a development due to Harrison; see Harrison and Stevens (1976). The error variance equations in the optimal filter theorem are now of the form

Between surveys

$$C(k+j \mid k) = C(k \mid k) + j\,Q\,,$$

At surveys

$$C(k \mid k) = [1 - K(k)\,]\,C(k \mid k\text{-}T)\,,$$

where $C(0 \mid 0)$, Q, and R are positive scalars and so are $K(k)$, $C(k|k)$, and $C(k+j|k)$. In this development the scalar quantity $n^o(k|k)$ was then defined by $n^o(k|k) = RC^{-1}(k|k)$ and referred to as the updated equivalent sample size after surveying at survey time k with no processing delay. It was further interpreted in inventory terms as the level of a "stock of information" on hand immediately after ordering $n(k)$ additional units (with no leadtime); that is, as an inventory "order level." The scalar quantity $n_r(k+j|k)$ was defined by $n_r(k+j|k) = RC^{-1}(k+j|k)$ and referred to as the equivalent sample size remaining at time $k+j$, j time units after the survey time k. It was interpreted in inventory terms as the "stock on hand" j time units after ordering and receiving new stock. For a fixed interval T between surveys, assuming the system is in steady state, $n_r(k \mid k\text{-}T)$ was interpreted in inventory terms as the "reorder point" and T as the "scheduling period." The Kalman gain becomes

$$K(k) = n(k) / [n_r(k|k\text{-}T) + n(k)]$$

and the updated equivalent sample size becomes

$$n^o(k|k) = n_r(k|k\text{-}T) + n(k)\,.$$

A further interpretation of $n^o(k|k)$ was that it is the size of a survey that would be required to have the same degree of precision as that provided by the combined amount $n_r(k|k\text{-}T) + n(k)$. This development led to a set of equivalent sample size relations in place of the error variance equations in the optimal filter theorem:

Between surveys

$$n_r(k+j|k) = n^o(k|k)[1\text{-}jQR^{-1}\,n^o(k|k)\,]^{-1}\,,$$

At surveys

$$n^o(k|k) = n(k) + n^o(k\text{-}T|k\text{-}T)\,[1 +$$

$$TQR^{-1}\, n^{o}(k\text{-}T\,|\,k\text{-}T)\,]^{-1}\,.$$

Smith and Barzily (1982) gave a numerical example for a two-item process assumed to be a vector random walk with scalar sample size n_d and integer sampling interval T (T = 1, 2, ..., 10 years) . With assumed cost coefficients for start-up cost and unit costs of interviewing, they demonstrated that the cost function J was convex in (n_d , T) and found a minimum for J by a numerical search. They noted that a survey administrator who was "concerned that the underlying process parameters may take unexpected jumps or exhibit turning points, which are not modeled by the simple time-invariant random walk models, would presumably opt for sampling more frequently than the optimal interval found by this method."

References

Anderson, T.W. (1978), "Repeated Measurements on Autoregressive Processes," *Journal of the American Statistical Association,*73, 271-278.

Azzalini, A. (1981), "Replicated Observations of Low Order Autoregressive Time Series," *Journal of Time Series Analysis,* 2, 63-70.

Bell, W. (1984), "Signal Extraction for Nonstationary Time Series," *Annals of Statistics,* 12, 646-664.

Bell, W.R. and Hillmer, S.C. (1990), "The Time Series Approach to Estimation for Repeated Surveys," *Survey Methodology,* 16, 195-215.

Binder, D.A. and Hidiroglou, M.A. (1988), "Sampling in Time," in *Handbook of Statistics, Vol. 6,* ed. P.R. Krishnaiah and C.R. Rao, Amsterdam: Elsevier Sci. Publishers, 187-211.

Bobbitt, S.A., Broughman, S.P. and Gruber, K.J. (1995) "Schools and Staffing in the United States: Selected Data for Public and Private Schools," E.D. TABS NCES 95-191, U.S. Dept. of Educ.

Box, G.E.P., and Jenkins, G.M. (1970), *Time Series Analysis: Forecasting and Control,* San Francisco: Holden-Day.

Ghosh, D. (1995), "Periodicity of School Surveys: An Extension of Probable-Error Modeling for the Case of a Random Walk Process," unpublished technical note, Synectics for Management Decisions, Inc.

Ghosh, D., Kaufman, S.F., Smith, W. and Chang, M. (1994), "Optimal Periodicity of a Survey: Sampling Error, Data Deterioration, and Cost," *1994 Proceedings ASA Section on Survey Research Methods,* 1122-1127.

Groves, R.M. (1989), *Survey Errors and Survey Costs,* New York: Wiley.

Harrison, P.J. and Stevens, C.F. (1976), "Bayesian Forecasting" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 38, 205-247.

Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter,* New York NY: Cambridge University Press

Kaufman, S. (1991), "1988 Schools and Staffing Survey Sample Design and Estimation," Technical Report NCES 91-127, U.S. Dept. of Educ.

Kaufman, S. and Huang, H. (1993), "1990-91 Schools and Staffing Survey: Sample Design and Estimation," Technical Report NCES 93-449, U.S. Dept. of Educ.

Scott, A.J. and Smith, T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods," *Journal of the American Statistical Association,* 69, 674-678.

Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," *International Statistical Review,* 45, 13-28.

Shumway, R.H. (1988), *Applied Statistical Time Series Analysis*, Englewood Cliffs: Prentice-Hall

Smith, W. (1994), "Nonlinear-Modeling for Schools Data with Level-Dependent Variances," unpublished technical note, Synectics for Management Decisions, Inc.

Smith, W. (1980), "Sample Size and Timing Decisions for Repeated Socioeconomic Surveys," unpublished D.Sc. dissertation, The George Washington University.

Smith, W. and Barzily, Z. (1982), "Kalman Filter Techniques for Control of Repeated Economic Surveys," *Journal of Economic Dynamics and Control,* 4, 261-279.

Smith, W. and Zalkind, D. (1978), "Statistical Decision and Control Approaches for Allocation of Funds," *1978 Proceedings of the ASA Section on Survey Research Methods,* 108-113.

Subcommittee on Small Area Estimation (1993), *Statistical Policy Working Paper 21: Indirect Estimators in Federal Programs*. Washington DC: Statistical Policy Office, OMB

Tam, S.M. (1987), "Analysis of Repeated Surveys Using a Dynamic Linear Model," *International Statistical Review,* 55, 63-73.

Tiller, R. (1992), "Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey," *Journal of Official Statistics*, 8, 149-166.

PROPERTIES OF THE SCHOOLS AND STAFFING SURVEY'S BOOTSTRAP VARIANCE ESTIMATOR

Steven Kaufman, National Center for Education Statistics
Room 422d, 555 New Jersey Ave. N.W., Washington, D.C. 20208

Key Words: Simulation, Half-Sample Replication, Bootstrapping, Variances

Introduction

The National Center for Education Statistics' (NCES) Schools and Staffing Survey (SASS) conducted by the Census Bureau has a complex sample design. Public schools are selected using a stratified systematic PPS (unequal selection probabilities) sample design. From this design, data are collected at the school and school district level. The school district is an aggregation unit (i.e., the district selection probability is computed by aggregating school selection probabilities containing the district across the school strata). The probability is nonlinear with respect to the school sample sizes. A bootstrap variance estimator (Kaufman,93) has been developed that provides better variance estimates than the balanced half-sample replication (BHR) variance estimator for the SASS public school district component. A bootstrap variance estimator for the other SASS components was presented in 1994 (Kaufman,94). The bootstrap variance estimators reflects the finite population correction associated with the SASS high sampling rates, without using the joint inclusion probabilities. **A set of bootstrap replicate weights are generated that work like BHR replicate weights, so that the bootstrap variances can be generated from any BHR variance software package.** It has also been shown that the bootstrap variance estimator performs better than BHR with other designs with high sampling rates (Kaufman, 94). This bootstrap variance estimator has been implemented into the 1994 SASS survey.

The goal of this paper is to provide results from simulation studies that demonstrate the bootstrap variance estimator (Kaufman, 94) works better than BHR with designs with low sampling rates. In addition, a balanced bootstrap will be presented that works better than the non-balanced bootstrap variance estimator.

First, a motivation why the bootstrap variances estimator may perform better than BHR is presented. Next, the balanced and non-balanced bootstrap variance estimators are described, as well as, the BHR estimators. The methodology presented here is the same as what is presented in (Kaufman, 94), except for the balancing of the bootstraps. A description of the designs being tested in this study follows. Finally, the results are presented showing the bootstrap variance estimator's superiority for the designs tested.

Motivation

BHR assumes 2 PSUs are selected with replacement within each stratum. To fit PPS systematic sampling of n PSUs into this model, sampled PSUs are paired by the order in which they were selected. Each pair is then treated as a stratum for variance estimation (variance stratum). If a systematic equal probability sample of size 10 is selected from a frame of 100 PSUs, the BHR model would have more than 10 trillion possible samples. In reality, there are only 10 possible systematic samples. Without, further homogeneity assumption, BHR can be a very large overestimate, even if the sampling rates are low. For this example, since the bootstrap selection is done systematically, approximately 10 possible bootstrap-samples can be selected from the bootstrap frame with each randomization of the bootstrap frame. Unlike the BHR estimator, a homogeneity assumption does not appear to be required; so the bootstrap estimator may get closer to the true variance.

Public and Private School-Bootstrap Frame

The idea behind the bootstrap samples is to use the sample weights (W_i) from the selected units to estimate the distribution of the school frame. From the estimated bootstrap-school frame, B bootstrap samples can be selected. The bootstrap-school frame is generated in the following manner:

For each selected school i, W_i bootstrap-schools (bi) are generated. If W_i has a noninteger component then a full school is generated with a reduced selection probability and weight. As shown in the bootstrap weighting section, the bootstrap expectation of the bootstrap weights (W_{bi}) equals the full-sample weight (W_i). The bi^{th} bootstrap-school has the following measure of size (m_{bi}):

$$m_{bi} = I_{bi} * 1/W_i,$$

$$I_{bi} = \begin{cases} 1 & \text{if bi is an integer component of } W_i \\ C_i & \text{if bi is a noninteger component of } W_i, \\ & \quad C_i \text{ being the noninteger component} \end{cases}$$

Bootstrap Sample Size

The bootstrap sample size is usually chosen to provide unbiased variance estimates. When the

original sample is a simple random sample of size n then Efron (1982) shows a bootstrap sample size should be n-1. Sitter (1990) has computed the bootstrap sample size for the Rao-Hartley-Cochran method for PPS sampling. A variation of this result is used in this simulation. Sitter's bootstrap sample size (n^*) is the sample size which makes the following quantity closest to 1:

$$\frac{\sum_{g=1}^{n^*}(N_g^{*2}-N^*)}{\sum_{g=1}^{n}(N_g^2-N)} * \frac{(N^2-\sum_{g=1}^{n} N_g^2)}{(N^*(N^*-1))}$$

n^*: is the bootstrap stratum sample size
g: represents a sampling interval in the stratum
N_g^*: is the number of bootstrap-schools in the g^{th} sampling interval, where the bootstrap-schools are in a random order
n: is the sample size in the stratum
N^*: is the number of bootstrap-schools in the stratum
N : is the number of schools in the stratum
N_g: is the number of schools in the g^{th} sampling interval, where the schools are in their original order; either a random order for the Rao-Hartley -Cochran method or the specific nonrandom order for the SASS method

n^* can not be calculated directly. The quantity above is computed for each n^* from n-20 to n. The n^* that is closest to one is used in the bootstrap selection.

The variation to Sitter's formulation is in the computation of N_g^* and N_g. Two modifications are made. The first occurs when I_{bi} is not equal to 1. Instead, of using 1, as Sitter does when counting units; I_{bi} is used to calculate N_g^*. The second modification is due to the fact that a school or bootstrap-school can be in two sampling intervals. When this happens, N_g and N_g^* are not increased by one. Instead, they are increased by the proportion of the unit that actually goes into the sampling interval. If I_{bi} does not equal 1, and the bootstrap-school is in two sampling intervals then N_g^* is increased by the product of the two modifications described above.

Determining the Bootstrap Sort Order

If the bootstrap variance estimate is to work correctly, it is important that the school-bootstrap frame be randomized in an appropriate manner. In one extreme, when the bootstrap frame is sorted by the order of selection from the original sample and $n^*=n$, the variance estimate will be zero. In the other extreme, when the bootstrap frame is sorted randomly, the variance estimate ignores the original ordering and may overestimate the variance. Bootstrap variances will be computed using a number of sort orderings for each of the simulation samples. Coverage rates are computed for each ordering. The coverage rates are compared with estimates of the true coverage rate. The ordering associated with the coverage rates closest to the true coverage rates is the ordering that is used for the bootstrap estimator. These comparisons are made at a level where the coverage rates should have some degree of stability given the number of simulations. For the designs in this study, the comparisons are made at the general school association/region level. The bootstrap sort orders are described below.

School Sort Method j

Selected schools within a stratum are sorted by order of selection. Next, schools are consecutively paired within each stratum. Each pair is assigned a random number. The bootstrap-schools generated within each pair of schools are assigned bootstrap-school random numbers. If $n-n^* \leq j$, for a stratum, the bootstrap-schools are sorted by bootstrap-school random number. If $n-n^* > j$, for a stratum, the bootstrap-schools are first sorted by the school pair random number; within each school pair the bootstrap-schools are sorted by the bootstrap-school random number. In other words, if the difference between the original and bootstrap sample sizes is small, as defined by j, then ignore the original sort ordering when randomizing the bootstrap-schools. Otherwise, randomize within pairs that reflect the original sort ordering.

The bootstrap program used in these simulations requires an initial bootstrap sort. Given this sort, the program searches for the sort that minimizes the maximum absolute bias in the average, total and ratio coverage rates. If the maximum absolute bias is less than or equal to 0.07 for a bootstrap sort, then that bootstrap sort will be used as the final sort for the association. Otherwise, the program tries other logical sorts. After the sort searching has finished, the coverage rate biases are reviewed and a final bootstrap sort is determined for each general association/region group.

Rationale for School Sort Method j

Sitter shows that if the number of schools in a sampling interval is constant across the intervals, then n^* will be close to n-1. If schools are sorted randomly, then the expected number of schools in the intervals is constant and n^* should be close to n-1. Therefore, if $n^*=n-1$, the assumption is that the sort ordering is effectively random, so that the school pairing should be ignored. Sort method j=1, sorts bootstrap schools

randomly if $n^{*}=n-1$. The smaller n^{*} is relative to n-1, the more effective the ordering is (i.e., the ordering acts less like a random ordering) and the more important the school pairings are to the sort method. Again, this is the affect of sort method j, when j is small.

When the pairings are ignored, a bootstrap-school generated for a particular school is in more sampling intervals and therefore can be selected more often. All other things kept equal, this should increase the bootstrap variance estimate. One then expects the variance from sort method j to be $\geq$ the variance from sort method k, when $j \geq k$. This rule can be used to determine which sort to use to improve the variance estimate. The rule, however, does not always work. This might be due to random error or to the implicit bootstrap-school joint inclusion probabilities that are generated. The coverage rate from a particular sort that matches the true coverage rate is implicitly: 1) matching the effective randomness of the original sort (sort method j=1), adding variability as necessary (sort method $j > 1$), as well as, 3) matching the bootstrap-school joint inclusion probabilities to the true school joint inclusion probabilities.

Bootstrap Sample Selection

Given the bootstrap frame, m_{bi} as the measures of size, stratum bootstrap sample sizes and bootstrap-school ordering, select the bootstrap sample using the same sampling scheme as in the original sample. The bootstrap frame is randomized with each sample selection. Bootstrap-schools, generated from noncertainty schools, with measures of size larger than the sampling interval are not removed from the sampling process. If a bootstrap-school is selected more than once, the bootstrap-school weight is multiplied by the number of times it is selected.

Balanced and Non-Balanced Bootstraps

Since systematic sampling gives good sample size control by values of the first sort variable, the variance estimate may be improved if the bootstrap samples have the same control (balance). This can be achieved by ordering the bootstrap frame by the first sort variable. Then, the bootstrap-schools can be randomized as described above within each of the values of the first sort variable. If the first sort variable is continuous there may not be enough bootstrap-schools within the sort variable's values to accurately estimate the variance using balanced bootstraps. In this situation, it might help to categorize the sort variable. Both balanced and non-balanced bootstrap variances will be presented.

Number of Replicates and Bootstraps

Since the old SASS BHR variances are based on 48 replicates, 48 bootstrap samples are computed for each simulation sample. Given the time it take to select a set of bootstrap samples, only 60 simulation samples are used.

Bootstrap Weights

The bootstrap-school weight, W_{bi}, is:

$$W_{bi} = I_{bi} * M_{bi}/p_{bi}$$

M_{bi}: is the number of times the bi^{th} bootstrap-school is selected

p_{bi}: is the bootstrap selection probability for the bi^{th} bootstrap-school

$E_{*}(\Sigma_{bi} W_{bi}) = \Sigma_{bi} I_{bi} = \Sigma_{i} W_{i}$, as desired.

E_{*}: is expectation over the bootstrap samples

Since the available data are defined by the schools selected in the original sample, a bootstrap-school weight indexed by i (BW_i) is required:

$$BW_i = \Sigma_{bi \in S_{iB}} W_{bi}$$

S_{iB}: is the set of all $bi \in i$ selected in the B^{th} bootstrap sample.

Balanced Half-sample Replicates

The r^{th} school half-sample replicate is formed using the usual textbook methodology (Wolter, 1985) for establishment surveys with more than 2 units per stratum. Since the SASS half-sample variances are based on 48 replicates, the simulations will be based on 48 half-sample replicates.

Three BHR variance estimates will be presented based on the methodology described above. The first (**BHR no FPC**) is the variance estimates described above. This estimate does not make any type of Finite Population Correction (FPC) adjustments.

The other two make simple FPC adjustments. The second BHR variance estimate (**BHR Prob FPC**) adjusts the first variance estimator by $1-P_h$, where P_h is the average of the selection probabilities for the selected units within stratum h.

The third BHR variance estimate (**BHR SRS FPC**) adjusts the first variance estimator by $1-n_h/N_h$, where n_h is the number of sample units in stratum h and N_h is the number of units on the frame in stratum h.

Low Sampling Rate Design I and II

The sample frame is the list frame component of NCES's Private School Survey (PSS). The list frame is stratified by general School Association (4 groups),

within Association by Census Region (4 levels), and within Region by school level (elementary, secondary and combined). The school sample is selected using a systematic probability proportionate to size sampling procedure. The design I uses square root teachers as the measure of size, while design II uses teachers. Before sample selection, the school frame is sorted by Urbanicity.

Sample Estimate

For each of the simulation samples, totals, averages and ratios are computed. The variables used are all on the sample frame. Two averages, one ratio and three totals are computed using estimated schools, teachers and students. For each of the 60 simulation samples, the sample estimates and respective sample variances are computed. An estimate of the true variance for the sample estimates can be obtained by computing the simple variance of the sample estimates across the 60 simulations. The bootstrap and BHR sample variance can now be compared with the estimate of the true variance.

When determining the bootstrap sort order estimates are computed within general school association/region. The estimates used in the tables are publishable estimates other than the ones used determining the sort order. To maintain stability given only 60 simulation are used, the samples used in the tables are the same as thoses used to determine the sort order.

The analysis statistics used to evaluate the variance estimates is described below.

Analysis Statistics

Coverage Rates

To measure the accuracy of the variance estimates, a one sigma two-tailed coverage rate is computed by determining what proportion of the time the population estimate is within the respective confidence interval. If the estimates are approximately normal then the coverage rates should be close to 0.68.

Coverage Rate Bias (**Bias**)

$$\text{Bias} = R_e - R_t$$

R_e: is the coverage rate based or either a bootstrap or BHR variance estimate

R_t: is an estimate of the true coverage rate, based on the simple variance of the simulation estimates

The distribtion of the coverage rate bias will be presented two ways. The first way, looks at the distribution of various publishable estimates implied by the sample design, this treats each publishable estimate equally. The second way, sums independent sets of the publishable total variances to produce a overall national variance for a estimated total. This method gives larger totals more weight than smaller totals and allows for variance underestimates to cancel out variance overestimates. In other words, the second way, provides a method of judging how well the variance procedures works when estimates are aggregated to produce new estimates.

Results - Coverage Rates by Publishable Estimates

Design 1 (Tables 1-3)

BHR No FPC variance estimates can be very large overestimates (**BIAS GE .14**); 35, 37 and 11 percent of the simulated estimates are in this category respectively for averages, totals and ratios. These coverage rates are closer to what one would expect from a two sigma coverage rate than from a one sigma coverage rate. Applying simple FPC adjustments helps somewhat, but the FPC adjusted BHR estimates still have between 6 to 29 percent in the very large overestimate category.

The tables shows the balanced bootstrap has 8, 4 and 0 percent in the very high overestimate category, respectively for averages, totals and ratios. This is much better than any of the BHR estimators. The not balanced bootstrap estimator has 16, 35 and 0 percent in this category respectively for average, totals, and ratios. The 35%, for the not balanced totals comes from estimates whose domain are not functions of the stratification variables (e.g., urbanicity and region/urbanicity). Once the bootstrap sample sizes are controlled on urbanicity (balanced bootstraps) the 35% drops to 4%.

For averages and totals, the BHR No FPC has the fewest number of estimates in the low bias category ([-.07,0.0) and [0.0, .07) categories); 30 and 14 percent, respectively for averages and totals. Applying simple FPC adjustments helps only marginally. Within the small bias category the bootstrap estimators perform better than BHR for averages and totals. The balanced bootstrap has 57 and 61 percent, in the low bias category, respectively for averages and totals. The not balanced bootstrap has 61 and 53 percent, in the low bias category, respectively for averages and totals.

For ratio estimates, the BHR estimators performs better than the bootstrap estimators. The BHR estimators have between 61 and 72 percent in the low bias category, while the bootstrap estimators have 53 and 58 percent in this category.

The bootstrap estimators are the only estimators that have a few estimates in the very large underestimate category (BIAS LT-.14). These coverage rates are closer to what one would expect from a .5 sigma coverage rate than from a one sigma coverage rate. The

balanced bootstrap has 2, 2 and 6 percent in this category, respectively for averages, totals and ratios. The not balanced bootstrap has 10% in this category for ratios.

Design II (Tables 5-7)

Tables 5-7 don't give either BHR or bootstrap a strong advantage.

The only estimator that has a large problem in the very large overestimate category is the not balanced bootstrap, with 20% for totals. The 20% comes from estimates whose domain are not functions of the stratification variables (e.g., urbanicity and region/urbanicity). Once the bootstrap sample sizes are controlled on urbanicity (balanced bootstraps) the 20% drops to 0%. All estimators have some estimates in the very large overestimate category, but usually just a few percent.

For totals and ratios, the BHR No FPC has the fewest number of estimates in the low bias category ([-.07,0.0) and [0.0, .07) categories); 59 and 72 percent, respectively for totals and ratios. Applying simple FPC adjustments helps. For BHR SRS, 67 and 76 percent are in this category, respectively for totals and ratios. For BHR Prob, 74 percent are in this category, for both totals and ratios. Within the small bias category the bootstrap estimators performs about as well as the BHR Prob and SRS estimators, for totals and ratios. The balanced bootstrap has 74 and 76 percent, in the low bias category, respectively for totals and ratios. The not balanced bootstrap has 70 and 82 percent, in the low bias category, respectively for totals and ratios.

For averages, the BHR estimators performs better than the bootstrap estimators. The BHR estimators have between 55 and 65 percent in the low bias category, while the bootstrap estimators have 43 and 51 percent in this category.

Results - Coverage Rates for Overall Estimates

The results presented above treat each estimate equally. The results in this section, provide a measurement of how well the estimators work when estimates are aggregated. Table 4 shows that for designs I and II, the bootstrap coverage rate biases are much smaller than the BHR coverage rate biases. The bootstrap biases are between 0.8 and 2.6 percent, while the BHR biases are between 3.8 and 7.3 percent. The no balanced bootstrap biases are the lowest. This indicates, since the main purpose of the balancing is to improve variance estimate for domains defined by the first sort variable, that if the only variances required are where the domain is defined by the stratification then the no balanced bootstrap is better than the balanced bootstrap.

Conclusions

The overall conclusion is that the bootstrap methods are better than the BHR methods for the designs in this study. How much better one method is than another depends on the sample design and the estimates of interest.

Coverage Rates for Published Estimates

These coverage rates treat each estimate equally.

For design I, using square root teachers as the measure of size, all BHR procedures have serious problems overestimating the variance. The balanced bootstrap procedure has a much smaller problem overestimating the variance. The no balanced bootstrap procedure overestimate the variance a large percent of the time when estimating averages and totals. For totals, this overestimation is caused from domains that are not function of the sampling strata.

For design II, using teachers as the measure of size, **BHR Prob** and the balanced bootstrap are comparable. The no balanced bootstrap procedure overestimate the variance a large percent of the time when estimating totals, but the the overestimation is caused from domains that are not function of the sampling strata.

Coverage Rates for Overall Total Estimates

These estimates provide a measure of how well the different variances work when aggregaing estimates. For both designs, the no balanced bootstrap is better than the balanced bootstrap. This indicates, if the only variances required are where the domain is defined by the stratification then the no balanced bootstrap is better than the balanced bootstrap. Both bootstrap methods are superior to all of the BHR methods.

References

Efron, Bradley(1982). The Jackknife, the Bootstrap and Other Resampling Plans. SIAM No. 38, p62.

Kaufman, Steven(1993). A Bootstrap Variance Estimator for the Schools and Staffing Survey. ASA 1993 Survey Research Methods Proceedings.

Kaufman, Steven(1994). Properties of the Schools and Staffing Survey's Bootstrap Variance. ASA 1994 Survey Research Methods Proceedings.

Sitter, R.R.(1990). Comparing Three Bootstrap Methods for Survey Data. Technical Report Series of the Laboratory for Research in Statistics and Probability, No. 152, p9-10.

Wolter, K. M.(1985). Introduction to Variance Estimation. New York: Springer-Verlag, p110-145.

Table -- 1 Publishable Estimate Dist. of Coverage Rate Bias for Averages using Private Design I

Bias	Bootstrap		BHR Estimates		
Averages (% Freq.)	Bal	No Bal	Prob	SRS	No FPC
LT -.14	2	0	0	0	0
[-.14,-.07)	2	2	2	2	0
[-.07,0.0)	14	20	12	12	8
[0.0,.07)	43	41	29	25	22
[.07,.14)	31	21	39	43	35
GE .14	8	16	18	18	35

Table -- 2 Publishable Estimate Dist. of Coverage Rate Bias for Totals using Private Design I

Bias	Bootstrap		BHR Estimates		
Totals (% Freq.)	Bal	No Bal	Prob	SRS	No FPC
LT -.14	2	0	0	0	0
[-.14,-.07)	0	0	0	0	0
[-.07,0.0)	14	20	6	6	4
[0.0,.07)	47	33	26	22	10
[.07,.14)	33	12	41	43	49
GE .14	4	35[1]	27	29	37

Table -- 3 Publishable Estimate Dist. of Coverage Rate Bias for Ratios using Private Design I

Bias	Bootstrap		BHR Estimates		
Ratios (% Freq.)	Bal	No Bal	Prob	SRS	No FPC
LT -.14	6	10	0	0	0
[-.14,-.07)	29	22	10	10	8
[-.07,0.0)	27	27	33	31	20
[0.0,.07)	26	31	39	41	41
[.07,.14)	12	10	12	12	20
GE .14	0	0	6	6	11

Table -- 4 Average Coverage Rate Bias from Overall Estimates of Totals Generated from Independent Groups by Design and Variance Estimator

Percent	Bootstrap		BHR Estimates		
Private Design	Bal	No Bal	Prob	SRS	NO FPC
Design I	0.9	0.8	4.0	4.1	6.1
Design II	2.6	1.9	3.8	5.4	7.3

Table -- 5 Publishable Estimate Dist. of Coverage Rate Bias for Averages using Private Design II

Bias	Bootstrap		BHR Estimates		
Averages (% Freq.)	Bal	No Bal	Prob	SRS	No FPC
LT -.14	4	6	4	0	0
[-.14,-.07)	18	14	18	21	8
[-.07,0.0)	14	10	18	12	24
[0.0,.07)	29	41	47	43	33
[.07,.14)	31	25	11	18	25
GE .14	4	4	2	6	10

Table -- 6 Publishable Estimate Dist. of Coverage Rate Bias for Totals using Private Design II

Bias	Bootstrap		BHR Estimates		
Totals (% Freq.)	Bal	No Bal	Prob	SRS	No FPC
LT -.14	4	0	2	0	0
[-.14,-.07)	10	8	12	8	4
[-.07,0.0)	31	35	25	24	14
[0.0,.07)	43	25	49	43	45
[.07,.14)	12	12	10	23	33
GE .14	0	20[1]	2	2	4

Table -- 7 Publishable Estimate Dist. of Coverage Rate Bias for Ratios using Private Design II

Bias	Bootstrap		BHR Estimates		
Ratios (% Freq.)	Bal	No Bal	Prob	SRS	No FPC
LT -.14	0	0	0	0	0
[-.14,-.07)	12	10	20	10	4
[-.07,0.0)	53	45	41	45	25
[0.0,.07)	23	37	33	31	47
[.07,.14)	8	6	4	12	20
GE .14	4	2	2	2	4

[1] The increase bias relative to the balanced bootstrap is due to the bias in estimates that are not functions of the stratification variables (i.e., urbanicity and region/urbanicity). Balancing the bootstrap samples by urbanicity reduces the bias.

DISCUSSION

Charles H. Alexander, Bureau of the Census

The Jackson and Frazier paper reminds us that even in the Age of Automation, expensive clerical work is still often required to merge multiple lists into a single frame. The need for clerical work is caused by lack of standardization of names and data fields for lists that were prepared with little or no coordination, often for disparate purposes. The desire to save money, and simplify processes so they can be more readily automated, creates a constant pressure to see whether multiple lists are really needed.

The main result of the paper is that multiple list frames are indeed needed to provide adequate coverage of the universe for the Private School Survey. Neither the State nor Association lists give adequate coverage alone. Even the lists for smaller Associations can have a noticeable effect on the coverage for affiliation categories. The Quality Education Data list does not seem to have added anything. However, this result needs to be double checked. The implied number of schools added to the QED between 1991 and 1993 is much smaller than seems reasonable; the procedure for identifying adds in this study should be reviewed carefully.

Can the results shed light on the completeness of coverage of the Private School Survey, using the existing frames? Perhaps, but more details would need to be recorded during the clerical operation, as described below.

Of course, when all the list frames included in the coverage study are actually used by the survey, it's impossible to prove any coverage deficiencies. For example, if a given State list did a very poor job covering the Association lists, this doesn't imply any undercoverage. After all, the Association-list Schools are covered by the Association lists and who's to say that the non-Association-list Schools aren't perfectly covered by the State list? Obviously this argument is dubious; such poor coverage of Association list would raise suspicions about the State list.

In this vein, we could seek indications of coverage problems by looking at the following.
For each Association and State:

i) what proportion of the Association-list Schools from that State are on the State list;

ii) what proportion of the State-list Schools with the relevant affiliation are on the Association list.

For States and Associations where these proportions are not high, questions should be asked about how the lists were put together to try to find out what's wrong. The second proportion is affected by inconsistencies in linking the affiliation information from the State list to the school's Association membership, as well as by the coverage errors we are looking for.

Although this information could be valuable, it would add steps to the clerical operation, so let me call this a suggestion rather than a recommendation, until the cost can be estimated.

The Kaufman, Li, and Scheuren paper gives a good illustration of the value of the Generalized Least Squares (GLS) methods of deriving survey weights, and the need for caution in using it. Their experience is similar to what was encountered in applying GLS to weighting for the Consumer Expenditure Surveys (Luery (1986), Zieschang (1986, 1990), Alexander (1987, 1990).)

Generalized least squares is a flexible, elegant method for making weighted estimates from surveys agree with as another or with controls derived from independent sources. But as the authors mention, it can have problems.

The most obvious problem, negative or very small weights, has several solutions. At a later session Jayasuriya and Valliant will present an appealing way of controlling the size of weights using the calibration estimation approach.

The more serious problem mentioned by the authors is the potential for harmful effects on estimates not directly controlled. We need a more complete theory of "harm" and "good" from the GLS method. The authors' "harm" measure is a step in the right direction.

At least part of the problem is that the "attractive asymptotic properties" of GLS do not apply when:

i) the survey has systematic undercoverage (Alexander, 1990); or

ii) the variables used to define the "control cells" have measurement error or are defined inconsistently between surveys; or

iii) as the authors note, when finite sampling

properties apply, either because of a small total sample size or because of a few large sample units.

In these circumstances, the original weighted sample estimates may be very far from the controls, and the results can in fact be very sensitive to the "loss function" used. In the household weighting context, the loss function used by the authors responds to a large across-the-board undercoverage of households of all sizes by raising the weights of large households relative to small households. A different loss function increases all weights proportionally (Alexander, 1987, Table 1).

Kaufman, Li, and Scheuren propose a solution similar to what was ultimately used by the Bureau of Labor Statistics in applying GLS to the Consumer Expenditure Survey: adjust for "undercoverage " (or other systematic deviation from agreement with controls) before applying GLS to force agreement with controls. This is in effect what the Olkin method does. This makes sense on these assumption that the "attractive asymptotic properties" more nearly apply once this bias is reduced.

The authors are to be commended for looking hard at their data and not being awed by the elegance of GLS, nor frightened off by the need to use it carefully.

In his solo paper, Kaufman likewise looks closely at how new methods actually work for his data and his sample design. Kaufman proposes and implements a bootstrap variance method inspired by a discussion in Efron (1992). He has to extend Efron's treatment to handle the case of systematic sampling without replacement.

The paper describes an extensive evaluation via simulations based on real SASS and PSS data. As the author has explained, his method is to draw repeated samples and calculate confidence intervals from each sample, see what proportion of the intervals cover the simulated population parameter, and to compare these proportions to the nominal confidence level. The author's conclusion is that the bootstrap method does better than the balanced half sample method previously used for the PSS as well as the SASS, with a few exceptions.

There is an obvious concern about the evaluation method as described. The bootstrap variance depends very much on the sort order applied prior to selection of the bootstrap sample. The optimal sort order is chosen as the one that given the best results looking at data from the same simulation on which it is evaluated; this may not be a fair evaluation. However, I suspect that this problem does not affect the basic result, because the range of sort orders actually need in the simulation is fairly limited, and because of results in Kaufman (1993) that show the bootstrap's superiority does not seem to be much affected by the exact ordering.

This problem aside, there are still some unanswered questions:

i) why is Kaufman's method occasionally not better than the balanced half-sample replication method? When does this occur?

ii) how does the bootstrap method compare to other improvements to the basic BHR method, such as variations on the stratified jack-knife, or Bob Fay's idea of giving partial weight to the "excluded half-sample." Intuitively, Kaufman's method has some of the same beneficial effects as these methods. Could this be the reason it beats the relatively crude BHR method used for SASS and PSS?

We need a more comprehensive theory of when and why these methods work best, and why.

Smith, Ghosh, and Chang boldly sail into tempestuous waters. The choice of survey periodicity is usually made based either on explicit but overly simplistic models, or on ad hoc intuitive attempts to consider the full range of concerns. Their paper is a skillful attack on this hard, controversial problem, of systematically representing the complexity of the periodicity choice. They explore some innovative approaches, though they do not reach a final conclusion.

I'm particularly appreciative of the complexity of this problem because of my recent involvement in similar problems related to the Census Bureau's so-called "Continuous Measurement" survey. We decided on an every-year (indeed every month) periodicity based on a much less careful analysis than that of these authors, but now we find we do need to take their kind of care with respect to the choice of how many years' data to use in small area estimates.

Among the authors' alternative ideas, there are many I like a lot, and few I would question.

Things I liked a lot:

- using ARIMA models to describe possible "population" values;

- consideration of methods for "short time series";
- the analogy with the inventory scheduling problem;
- the authors' awareness of the ambiguity of the notion "total resources are fixed."

Things I'm less enthusiastic about:

- the assumption that if the survey designer assumes a particular ARIMA model to evaluate the best periodicity, then data users will use forecasts from that model to analyze the data;
- waiting for "reliable cost estimates of all relevant cost elements", even for periodicities never encountered in practice;
- focussing only on the unconditional properties of the estimates.

Users will do as they please regardless of the designer's assumptions. In some applications, such allocating funds, it may make sense to project ahead to the current year if a good model is available. For other applications, users will prefer the last direct cross-sectional estimate.

It is very hard to speculate how the operation would be organized for periodicities that have never been used in practice, and what it would cost. We'd be fortunate to get plausible ranges for the cost.

As do most statisticians, the authors focus on the unconditional properties of estimates. Some statisticians would disagree with this, as would many politicians. If the realized recent values of the time series for a State are such that the State estimates are adversely affected by a particular periodicity for the next few years, it is little consolation to explain that their recent values are the product of a process for which that periodicity works well on average.

My general suggestion about this problem is that the conclusion must consider various possible combinations of: i) ARIMA models; ii) independent variables; iii) analyses and data uses; iv) sets of cost parameters; v) loss functions; vi) approaches to the evaluation.

It is not reasonable to wait for a single final answer to the questions "what is the world like" and "what are the important uses of the data". Instead the best periodicity should be calculated for various combinations of the above considerations. Then for each periodicity, a statement could be made of the assumptions and uses which it best supports. This would help in focussing on exactly what time series measurement problems or rankings of priorities must be addressed to make the decision about periodicity.

References

Alexander, C. H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Alexander, C. H. (1990). Incorporating person estimates into household weighting using various models for coverage. *Proceedings of the 1990 Census Bureau Annual Research Conference*, 445-462.

Luery, D. M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association*, 325-330.

Zieschang, K. D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association, 64-71.*

ASSESSING QUALITY OF CCD DATA USING A SCHOOL-BASED SAMPLE SURVEY

Sameena Salvucci, Sandeep Bhalla, Michael Chang, Synectics for Management Decisions, Inc. and John Sietsema, National Center for Education Statistics
Sameena Salvucci, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd. #305, Arlington VA 22201

KEY WORDS: Administrative records, Sample surveys, Multiple reporting levels, Definitional differences, Record Matching

I. Overview

This paper describes the methodology problems and data quality issues associated with the 1990-91 Common of Core of Data (CCD), a national database of universe data reflecting three levels of aggregation (state, local education agency, and school) collected from state education agency (SEA) administrative records. It evaluates the feasibility of using external school-based sample survey data, the 1990-91 Schools and Staffing Survey (SASS), to assess the accuracy of the CCD. It also describes the results of record matching procedures used to explain some of the existing differences between CCD and SASS.

II. Context and Motivation

The National Center for Education Statistics (NCES) has been authorized by Congress, in part, to collect, analyze, and disseminate full and complete statistics on education in the United States. A primary way that NCES pursues this goal is through maintaining a comprehensive and timely national statistical database, the Common Core of Data (CCD). CCD is comprised of three separate nonfiscal surveys, the Public Elementary and Secondary School Universe (School Universe), the Local Education Agency Universe (LEA Universe), and the State Aggregate Nonfiscal Survey (State Aggregate). CCD provides general descriptive information, basic statistics, and fiscal data regarding all children in the United States enrolled in public schools, from prekindergarten through the twelfth grade, as well as staff, schools and local education agencies. However, participation in CCD is a voluntary activity of the states. NCES asks states to provide, from their administrative records, information they have secured from schools and LEAs. For the most part, the data requested by NCES are already collected by the states in the exercise of their responsibility for public education.

Given the variety among state definitions of the statistics being collected, there has been concern about how useful the national summaries of these data are. In response, NCES has for many years conducted activities to develop standard definitions and procedures, help states observe these standards, and improve data quality of the CCD. Efforts continue to be devoted to improving CCD data accuracy. Since CCD has recently become the sampling frame for all NCES school-based surveys, new efforts are focusing on measuring the accuracy of the data. The work described in this paper, in particular, concentrates on measuring the accuracy of key statistical information in the CCD, such as the total number of students, teachers, schools and school districts.

While much can be learned from analysis of CCD data itself, another useful approach to measuring the accuracy of CCD is to compare it to data from other surveys. NCES's 1990-91 SASS, a national sample survey of public and private schools, is one such source of comparable data. SASS is comprised of four interrelated surveys. Three of these surveys are sent to public and private schools : (1) School Survey, (2) Administrator Survey, and (3) Teacher Survey and the fourth survey is sent to LEAs and is called the Teacher Demand and Shortage Survey (TDS).

III. Consistency Within CCD

We began by examining the national estimates of student and FTE teacher counts from the three separate Nonfiscal surveys of CCD (School Universe, LEA Universe, and State Aggregate) for 1988-89, 1989-90, and 1990-91. The national level differences between the estimates from the three CCD survey components for a particular year showed improvement over time, and in 1990-91 no difference was larger than 1.2 percent. Even though the national level differences in 1990-91 were small, further examination identified some large differences between the three CCD components for some states. When student counts were summed up to the state level from the School Universe and compared to the student counts from the State Aggregate, two states had student count differences greater than 5 percent. When student counts from the School Universe were compared to the student counts from the LEA Universe, each aggregated to the state level, five states had student count differences greater than 10 percent. Even more striking, when FTE teacher counts were summed up to

the state level from the School Universe and compared to the FTE teacher counts from the State Aggregate, seven states had FTE teacher count differences greater than 20 percent.

Table 1 indicates the number of states where the student count differences exceeded one percent for the comparisons described above. The number of states with student count differences greater than one percent had discrepancies less than five percent consistently over the three years. The number of states with student count discrepancies greater than one percent decreased between 1988 and 1990, suggesting an overall improvement in the quality of CCD student count data over time.

Table 1 also lists the results of the FTE teacher count comparisons. Comparisons are not applicable between the school and LEA level for FTE teacher counts since the FTE teacher counts are not collected at the district level in CCD. Comparisons between the School Universe and the State Aggregate, however, do not show an improvement over the three years. In fact, the number of states with differences greater than 5 percent increased from 13 to 15. To add to the problem, some states have not been able to provide FTE teacher counts at the school level.

Across all the states, student count data exhibit more consistency across CCD survey components than FTE teacher counts, especially between the CCD School Universe and the State Aggregate Survey.

Table 1: Comparison of Estimates Aggregated to the State Level Between CCD Survey Components

Comparison	Number of States		
	1988-89	1989-90	1990-91
Student Count (Difference >= 1%)			
School vs LEA	16	12	12
School vs State	13	8	9
FTE Teacher Count (Difference >= 5%)			
School vs LEA	N/A	N/A	N/A
School vs State	13	12	15

Potential sources of discrepancies between these levels of reporting include:

- different interpretations by states of CCD definitions;
- variations in data collection and editing quality within states and in aggregating state reports to the national level; and
- external conditions that may limit the comparability of a reported item from state to state regardless of how well definitions and data processing standards are applied.

IV. CCD-SASS Consistency

The accuracy of the 1990-91 CCD estimates of interest were further examined by comparing this data to another survey, the 1990-91 SASS.

Both CCD and SASS provide estimates of student counts, FTE teacher counts, the number of schools and the number of school districts (LEAs). This section describes the multiple comparisons we used to assess the level of accuracy of these counts in CCD. Table 2 summarizes the eight different comparisons made.

Table 2: CCD-SASS Comparisons

1990-91 CCD	1990-91 SASS
Student Counts	
School Universe	Public School Survey
LEA Universe	TDS Survey
State Nonfiscal Survey	Public School
FTE Teacher Counts	
School Universe	Public School Survey
LEA Universe	TDS Survey
State Nonfiscal Survey	Public School Survey
School Counts	
School Universe	Public School Survey
LEA Counts	
LEA Universe	TDS Survey

For every comparison described in table 2 above, a two-step approach was used to make a decision on which states had the largest CCD-SASS differences for the estimates of interest. The first step identified the 95% confidence interval around the SASS estimate for

each state. Although the confidence intervals took sampling variance into account, some SASS estimates had very small standard errors or no standard error (e.g. states with only one LEA). The resulting confidence intervals for these SASS estimates had very small ranges which increased the likelihood that the corresponding CCD estimates would fall outside the interval. Therefore, we found that combining the confidence interval approach with examination of the actual percent difference between CCD and SASS was necessary. Examining the confidence intervals in conjunction with the percent difference provided a more realistic indication of large discrepancies.

The CCD state estimate was compared to the 95% confidence interval bounds around the SASS state estimate. For those states where the CCD estimate fell outside the 95% SASS confidence interval, the absolute value of the relative percent difference between CCD and SASS was calculated. When the CCD state estimate was both outside the 95% SASS state estimate confidence interval and the absolute value of the relative percent difference exceeded 10 then we identified the CCD state estimate as discrepant.

Table 3 lists the number of states where we identifed the discrepancies between CCD and SASS as large according to the criteria described above. The intent was to define the extent of the problems requiring further investigation and to understand some of the issues surrounding cross-survey comparisons before any adjustments were made to the CCD data for differences in definitions between CCD and SASS.

Table 3: Discrepant States Before Definition Adjustment

Estimate	# of states where the estimate was identified as discrepant
SASS School vs. CCD School	
Schools	6
Students	2
FTE Teachers	10
SASS TDS vs. CCD Agency	
LEA's	11
Students	1
FTE Teachers	10
SASS School vs. CCD Nonfiscal	
Students	4
Teachers	6

One possible source of the SASS-CCD discrepancies is that the definitions of what are nominally the same variables - LEA, school, student enrollment, and teacher count - substantially differ in their operational definitions between the SASS surveys and CCD. Recognizing that some of the discrepancies identified in table 3 were caused by these differences in definitions, we attempted to reconcile the data being compared.

The reconciliation of CCD data to SASS data was achieved through the creation of modified CCD files which more closely matched the SASS definitions of student enrollment count, FTE teacher count, number of schools and LEAs.

The process of producing revised estimates is described in two parts: steps common to all estimates, and steps devoted to producing revisions specific to the school, student, FTE teacher and the LEA counts.

SASS is only conducted in the 50 states and the District of Columbia (SASS Data File User's Manual 1990-91), whereas CCD consists of 50 states, the District of Columbia and the five U.S. outlying areas. As a result these outlying areas were removed for all the comparisons (Instructions for Completing the Nonfiscal Surveys of the Common Core of Data 1991). CCD defines a public school as an institution which provides educational services and has one or more grades prekindergarten through 12 or ungraded. On the other hand, SASS defines a school as an institution that provides educational services for at least one of grades 1 through 12 (or comparable ungraded levels). Since schools that offered only prekindergarten or kindergarten classes were not eligible for SASS, these schools were removed from CCD. A final adjustment was made by deleting all other schools in the 1990-91 CCD school universe which were not eligible for SASS.

In addition, prekindergarten enrollment was subtracted from the total student count for each school because SASS student enrollment at the school level includes only students from kindergarten through grade 12.

Adjustments were also made to compensate for the differences in the definition of teachers on CCD and SASS. On the SASS Public School Survey, a teacher is defined as any full-time or part-time teacher whose primary assignment was teaching in any of the grades kindergarten through grade 12. Itinerant teachers are included as part of the teacher count, as well as long-

term substitutes who were filling the role of a regular teacher on a long-term basis.[1] Short-term substitute teachers, student teachers, nonteaching specialists (e.g. guidance counselors and librarians), administrators, teacher's aides, and support staff are not included. These counts are head counts, NOT FTEs. In the CCD Public School Universe, however, the teacher count is stated in FTEs (full-time equivalents). This count includes only filled positions. The difference between a head count and a FTE teacher counts can be substantial. Also, CCD teacher counts include prekindergarten teachers, while SASS teacher counts do not. It is not possible to subtract the prekindergarten teachers from the total teacher counts as we did with the total student counts because CCD does not collect FTE teacher counts by grade level. Despite these problems we tried to match the two counts by converting the SASS head counts into FTE teacher counts. We created a derived FTE teacher count for SASS equal to the sum of the number of full-time teachers plus a weighted number of part-time teachers.

In SASS, an LEA is defined as a government agency that employ teachers. There are LEAs that employ teachers which do not operate schools. For example, some states have special education cooperatives that employ special education teachers who teach in one or more LEAs. In CCD, however, an LEA is defined as a government agency responsible for providing instructional services. The CCD definition does not mention teacher employment. In fact, the 1988-89 CCD frame included 1,352 LEAs which are not associated with schools, but hire teachers (Quality Profile for SASS, 4.2; The SASS Data File Users' Manual, p. 24). In order to include them in the SASS TDS population, a 1 in 10 systematic random sample of these districts was taken and included in the SASS sampling frame.

To replicate this design in our comparisons, all CCD districts linked to schools from the CCD school file were included in the comparisons. From this set we deleted all those LEAs which were only linked to those schools that had only prekindergarten or kindergarten enrollment. For those LEAs not linked to schools only a 1 in 10 sample was included (Quality Profile For SASS, p4.2).

[1]. An itinerant teacher is defined as a teacher who teaches at more than one school (for example, music teacher who teaches three days per week at one school and two days per week at another).

At the LEA level, the student count in CCD is reported as the sum of prekindergarten-12 plus ungraded students. To match the CCD enrollment figure with SASS School enrollment, the reported number of prekindergarten students for each school as provided in the CCD School Universe was aggregated for each LEA. The number of prekindergarten students was subtracted from the CCD LEA total student count.

For the state level comparisons, prekindergarten students are removed from the CCD State total student enrollment in order to derive a comparable estimate to the SASS survey.

The comparisons made for the FTE teacher counts followed the same adjustment as the school FTE count, but the data used was from the CCD State Aggregate Survey.

Table 4 lists the number of states where we identified the discrepancies between the reconciled CCD data and SASS data as large according to our two step criteria.

Table 4 : Discrepant States After Adjustment

Estimate	# of States where the estimate was discrepant
SASS School vs. CCD School	
Schools	3
Students	2
FTE Teachers	12
SASS TDS vs. CCD Agency	
LEA's	2
Students	0
FTE Teachers	3
SASS School vs. CCD Nonfiscal	
Students	2
Teachers	5

V. Sources of CCD-SASS Inconsistency

After the identification of the set of states which have large CCD-SASS discrepancies for the estimates of interest, we focused our efforts on determining potential sources of the discrepancies within these states. Since SASS is a sample survey, we matched the set of all SASS schools and districts in the discrepant states with the corresponding CCD school and district in all three years 1988-89, 1989-90, 1990-91. We

limited our examination to only those districts and schools within the discrepant states for which the 1990-91 SASS-CCD discrepancy was greater than 10 percent for the student enrollment counts and greater than 20 percent for the FTE teacher counts. Next, we developed a flag that determined whether the size of the 1990-91 CCD-SASS discrepancy was larger than any discrepancy between the CCD counts across the three years examined. If the between year CCD differences were each smaller than the 1990-91 CCD-SASS difference, we felt that the 1990-91 CCD was probably more accurate than if the 1990-91 CCD-SASS difference was larger than any of the between year CCD differences for a particular school or LEA. In the latter case, we further compared the characteristics of these schools and districts as identified in both SASS and CCD. These comparisons indicated that most of the error was not due to random processing error, but rather showed systematic differences caused by different interpretations of definitions when reporting for SASS versus CCD. The following paragraphs describe some interesting findings.

The subsequent discussion is restricted to discrepancies in student counts and FTE teacher counts. Although CCD does not classify schools by level (elementary, secondary, combined), there should be consistency between the counts reported on CCD and the level reported on SASS. This is not always the case. For example, schools classified as secondary on SASS have no student counts in Grade 3, however there are 701 students reported for these schools on CCD. When SASS defined the school type as elementary, secondary or combined it was consistent with the definitions. That is, when the school was declared as elementary there was no student enrollment in SASS from grades 9 through 12, and when the school was secondary, there was no student enrollment in grade kindergarten through 6. On the other hand, CCD always reported student counts across the board even when the state was defined as elementary or secondary. Another interesting aspect of the analysis was that among the discrepant states across the eight comparisons the number of ungraded students reported always differed by a considerable amount.

We also conducted two rounds of interviews with state coordinators who submit the CCD data to NCES. We pinpointed states with large discrepancies between CCD and SASS in terms of student counts and FTE teacher counts. The first round of interviews resulted in a better understanding of which states were not able to collect certain types of information at certain levels and this was documented and helped reduce the number of states we examined further. The second round of interviews elicited state coordinator input on more specific reasons for large discrepancies in school and district level discrepancies within the subset of states examined. Some very insightful details were recorded during the second set of interviews. CCD coordinators, in some cases, reported that State Education Departments were imputing these missing numbers, while others reported that the number was not correctly entered in the electronic version of the report. Some state coordinators reported that districts within their states were reporting teacher head counts instead of FTE teachers or were reporting FTE teacher counts for the LEA and state levels, while reporting teacher head counts at the school level. A major problem reported by states with large discrepancies was that they were double reporting the student counts. Specifically, one state coordinator explained that the schools in that state reported the vocational student counts, once as a part of the total student count and once as a part of the individual school count. As a result these numbers were counted twice in the total student population at the state. We also found during these interviews, that some states had students not enrolled in a school, but were enrolled in a district. As a result, districts in CCD were reporting a higher number of students compared to school level student counts when summed up to the district level. Errors were also reported when aggregating the LEA level enrollment in the CCD. Some states report the number of student counts in the LEA to a supervisory union. The schools also report their number to a supervisory union. The supervisory union in turn added the two counts and reported the sum (double count) as the LEA student count.

VI. Next Steps

Many of the data quality assessment methods described in this paper could be used to further assess other important data elements in the 1990-91 CCD, such as the number of students by race/ethnicity. The quality of the 1993-94 CCD data could also be assessed using these methods, because survey data is now available for the 1993-94 SASS. Additional variables were collected in both the 1993-94 CCD, such as number of dropouts, that could be assessed using the Census Bureau's October School Enrollment Supplement to the Current Population Survey.

Electronic versions of this dataset are widely used by education researchers and policy makers. Any data quality problems at the school, district or state level could be included either in mainframe tape

documentation as technical notes and as context-sensitive help in the CCD CD-ROM.

VII . Bibliography

Jabine T. (1992), *Quality Profile for SASS, Aspects of the Quality of Data in the Schools and Staffing Surveys*, U.S. Department of Education, Washington D.C.

Kaufman, S. And Huang, H.(1991), *Schools and Staffing Survey : Survey Design and Estimation*, NCES Technical Report, July 1993. U.S. Department of Education, Washington D.C.

National Center for Educational Statistics (1991-92), *Instructions for completing the Nonfiscal Surveys of the Common core of Data*, School Year 1991-1992, U.S. Department of Education, Washington D.C.

----------(1989), *Data Base documentation, Common Core of Data Public School Universe, 1988-89*, U.S. Department of Education, Washington D.C.

----------(1989), *Data Base documentation, Common Core of Data Public Education Universe, 1988-89*, U.S. Department of Education, Washington D.C.

----------(1990), *Data Base documentation, Common Core of Data Public School Universe, 1989-90*, U.S. Department of Education, Washington D.C.

----------(1990), *Data Base documentation, Common Core of Data Public Education Agency Universe, 1989-90*, U.S. Department of Education, Washington D.C.

----------(1991), *Data Base documentation, Common Core of Data Public School Universe, 1990- 91*, U.S. Department of Education, Washington D.C.

----------(1991), *Data Base documentation, Common Core of Data Public Education Universe, 1990-91*, U.S. Department of Education, Washington D.C.

----------(1990-91), *SASS Data File User's Manual*, U.S. Department of Education, Washington D.C.

----------(1990-91), *Schools and Staffing Surveys: Data File User's Manual, Volume I*, U.S. Department of Education, Washington D.C.

DOCUMENTATION OF NONRESPONSE AND CONSISTENCY OF DATA CATEGORIZATION ACROSS NCES SURVEYS

Steven Fink, Mehrdad Saba, Michael Chang, Samuel Peng
Steven Fink, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd #305, Arlington VA 22201

KEY WORDS: NCES survey documentation, Unit response rate, Item response rate, Data categorization

During the last two decades, interest and concern have been growing regarding nonresponse (unit and item) in federal surveys because of how this issue relates to survey data quality. This report provides a systematic review of past and ongoing research on availability and calculation of response rates (both unit and item), and uniformity of several response categories of several NCES surveys.

Unit nonresponse is vitally important to users of federal surveys. Several attempts have been made to standardize response rate definitions. For example, the Council of American Survey Organizations (CASRO) reviewed response rate definitions with the intent of trying to establish uniformity of definitions across surveys (CASRO 1982). More recently, the Subcommittee on Nonresponse, commissioned in 1991 by the Office of Management and Budget's Federal Committee on Statistical Methodology (FCSM) provided the following recommendations:

1) Survey staffs should compute response rates in a uniform fashion over time and document response rate components on each edition of a survey.
2) Survey staff for repeated surveys should monitor response rate components (e.g., refusals, not-at-homes, out-of-scopes, address not locatable, postmaster returns, etc.).
3) Response rate components should be published in survey reports; readers should be given definitions of response rates used, including actual counts, and commentary on how response rates affect survey data quality.
4) Some research on nonresponse can have real payoffs. It should be encouraged by survey managers as a way to improve the effectiveness of data collection operations.

Item nonresponse occurs when the person participates in the survey but fails to answer some of the questions. It may arise for several reasons, including lack of respondent information, refusals, and inconsistency with other responses. This last category may include an inconsistency arising from a coding or keypunching error occurring in the transfer of the response from the answer sheet to the computer data file (Kalton 1983).

This report also examines consistency in data categorization. We identify commonly used demographic variables in NCES surveys and explore question wording and response categories of nine demographic survey items.

NCES Databases

We chose a mix of 13 surveys from NCES sample populations. Among these surveys, NALS and NHES were non-school-based surveys.

Elementary / Secondary Education
- o Schools and Staffing Survey (SASS, 1990-91)
- o Teacher Follow-up Survey (TFS, 1991-92)

Postsecondary Education
- o National Household Education Survey - Adult Education Component (NHES, 1993)
- o National Postsecondary Student Aid Study (NPSAS, 1990)
- o National Survey of Postsecondary Faculty (NSOPF, 1993)
- o Recent College Graduates Study (RCG, 1991)

Educational Assessment
- o National Assessment of Educational Progress (NAEP, 1990)
- o National Adult Literacy Survey (NALS, 1992)

National Longitudinal Studies
- o Baccalaureate and Beyond Longitudinal Study (B&B, 1993-94)
- o Beginning Postsecondary Student Longitudinal Study (BPS, 1992)
- o High School and Beyond (HS&B, 1992-93)
- o National Education Longitudinal Study of 1988 (NELS:88, 1991-92)
- o National Longitudinal Study of 1972 (NLS-72, 1979-80).

Response Rate Information

Below we present technical issues affecting responses in NCES surveys.

Unit Response Rates

Unit response rate refers generally to how many survey instruments were returned/completed. Below we present responses to seven questions relevant to unit response

rates. A common definition of unit unweighted response rate is the ratio of the number of units with completed interviews (the units could be telephone numbers, households, or individuals) to the number of units sampled and eligible to complete the interview.

A. Are unweighted unit response rates calculated consistently?

Eight of the 11 NCES surveys which provided calculations in their documentation used the same basic formula to calculate unweighted response rates. However, the definitional terms and level of detail varied from one survey to another. Several surveys used different names to describe the same response rate calculation. For example, a "locating" response rate (BPS) appears similar to a "screener" response rate (NHES), and an "interview" response rate (BPS) appears similar to an "extended interview" response rate (NHES). In NAEP, the unit response rate was referred to as the participation rate (also called the cooperation rate). RCG provided figures to calculate the unweighted response rate without specifying the formula to use, and NALS presented the unweighted response rates without providing the formula. TFS did not provide unweighted response rates.

B. Are substitute schools used in the calculation of unweighted unit response rates?

Among the surveys we examined, the answer is basically "no." Survey methodology sometimes allows substitute schools to replace nonparticipating schools in a selected sample; for example, when a selected school does not respond to a survey, another school with similar characteristics was asked to fill in. However, of the 11 school-based surveys we examined, only three-- NLS-72, NELS:88, and HS&B (all National Longitudinal Studies)--used substitute schools when calculating unweighted unit response rates, and NLS-72 and HS&B also calculated what unit response rates would be without the substitute schools.

C. Are weighted unit response rates calculated consistently?

A weighted unit response, with the effects of the sampling design incorporated into the calculation gives more accurate response information than the unweighted rate. Therefore, the weighted unit response is often a better measure for deciding whether further nonresponse studies should be conducted. Among the eight surveys identified in A, four did not provide weighted response rates. The other four (NELS:88, NPSAS, SASS, and HS&B) used the same basic formula. NELS:88 also used an additional weight: student design weight or school design weight. RCG used the standard formula. NALS provided no formula in its documentation (but rates were tabulated), and BPS and NHES utilized the sampling design to compute rates. TFS used a subsample of a previously conducted survey for its sampling frame. The weighted response rate was calculated as the product of SASS teacher list response, the SASS teacher response rate, and the TFS teacher response rate. Although NALS did not provide the formula for calculating weighted response rates, the documentation does state that the weighted response rates were calculated by applying the sampling weight to each individual to account for his/her probability of selection into the sample.

D. Are school/institution level response rates weighted by enrollment?

School or institution may be used as the sampling frame variable because small schools may have unique characteristics not associated with larger schools. Enrollment may then be used to weight the data. Only three of the school-based surveys examined--RCG, HS&B, and NELS:88--provided information on school level response rates, with only two weighting the response rate directly by school enrollment. In RCG, institution weight takes into account the sampling probability of the institution, which is proportionate to enrollment size. In HS&B, the school sampling probability was also proportional to the estimated enrollment.

E. Is there an intensive follow-up of nonrespondents? If so, were results built into the response rates?

One of the most pervasive and challenging sources of nonsampling error in estimates from sample surveys is the bias associated with nonresponse. Respondents may differ significantly from nonrespondents. Most of the time, funds are not available to conduct respondent follow-ups and convert every nonrespondent. One way to reduce bias is to take a subsample of nonrespondents and conduct an intensive follow-up to get everyone to respond. Different modes of data collection are used to encourage respondents to return their survey. However, only NELS:88 took a subsample of nonrespondents and conducted an intensive follow-up. NCES usually attains relatively high response rates and quality data. This may explain why intensive follow-up is usually not conducted.

F. Are unit response rates tabulated by the frame variables?

Frame variables such as sector and school or institution type are often used to select the samples from the populations. Tabulating unit response rates by frame variables helps to identify low and high response in certain strata. This practice can help researchers identify and perhaps improve future response in low response

strata or frames. Frame variables were used to tabulate unit response rates for all but three of the surveys we examined (NSOPF, NALS, and NLS-72). Given that most of those surveys were school-based, institution/school type was the frame variable most commonly used in the tabulations.

G. How is the overall unit response rate (covering all stages of sampling) calculated?

An overall unit response rate is only discussed for surveys using multi-stage sampling designs. Typically, an overall unit response rate for a two-stage sample survey is calculated as follows:

Overall unit response rate =
(First stage rate * Second stage rate)

The seven surveys which did calculate overall response rates all used this basic formula, although language differs. Four of the surveys we examined (NSOPF, HS&B, NALS, and NELS:88 2nd Follow-up) did not calculate an overall response rate or did not mention ways of calculating this type of response rate.

Item Response Rates

Item nonresponse has the effect of diminishing the number of observations that can be used in calculating statistics from affected data elements and thus increases sampling variances (Ingels et al. 1994). NCES standards stipulate that item response rates "are to be calculated as the ratio of the number of respondents for which an inscope response was obtained divided by the number of completed interviews for which the question was intended to be asked." Below we present responses to three questions relevant to item response rates, followed by two questions examining nonresponse research and availability of nonrespondents on data files.

H. Are unweighted item response rates calculated consistently?

SASS, BPS, HS&B, NSOPF, and B&B used the NCES standard as the means of calculating unweighted item response rate, although the exact wording varied. B&B and NSOPF defined item nonresponse. (It should be noted that the documentation for three of those surveys--BPS, HS&B, and B&B--did not explicitly identify the item response rate definition as unweighted or weighted.) A look at four surveys examine "Don't know" responses as a source of possible difference when calculating inscope responses. B&B provided separate tabulations for refusals and "don't know" responses, and presented a combined nonresponse rate integrating the two. NELS:88 used "don't know" as a valid response to certain questions, so it did not classify "don't know" as a nonresponse. In RCG, item nonresponse included responses of "don't know," "refused," and "not ascertained." However, there were no questions where "don't know" was considered a response (Westat, Inc. 1994). Finally, for NSOPF, "don't know" was included as an item nonresponse even in cases where "don't know" was an explicit response category for the item (Abraham et al. 1994).

I. Are weighted item response rates calculated consistently?

Only three surveys, NELS:88, RCG, and TFS, defined weighted item response rates. All used the standard definition, although exact wording varied. Considering unweighted and weighted item response rates together, all eight surveys which provided definitions used the NCES standard definition.

J. Are item response rates tabulated by subgroups?

Presentation of item response rate information varied considerably. At one end of the spectrum, RCG and NSOPF presented item response rates for all questions. At the other end, NAEP, NALS, and NLS-72 did not tabulate any item response rates. NHES and NPSAS are the two surveys which used subgroups in their presentations on item response rates. The tabulated subgroups on NHES were participation items, course or activity items, and sociodemographic items. NPSAS used four subgroups: student characteristics, enrollment variables, costs, and aid eligibility variables. The other surveys took one of two approaches. B&B and HS&B simply presented item response rates for a selected number of items. The rest--SASS, TFS, BPS, and NELS:88--presented information only on those items which exceeded a designated response rate (or nonresponse rate) threshold.

K. Is there any research dealing with nonresponse rates; e.g., adjustment, incentives, etc.?

We identified research done on three surveys--SASS, NSOPF, and NHES. For SASS, there were several reports (often in the form of memos or articles) examining characteristics of nonrespondents. NSOPF included an experimental design to examine the effect incentives and prompts can have on nonresponse rates. For NHES, there were internal memos and a report examining telephone undercoverage. One reason there may be so little research on nonresponse in NCES surveys is that response rates are generally high. As the following table shows, the majority of unit response rates exceed 80 percent (see Table 1).

Table 1: Unit Response Rates, by NCES Survey

Survey Name	Unweighted (%)	Weighted (%)
Elementary/Secondary Education		
Schools and Staffing Survey (SASS)		
School Administrator (public)	96.9	96.7
School Administrator (private)	91.1	90.1
TDS (public)[1]	93.7	93.5
TDS (private)	84.8	83.9
School (public)	95.0	95.3
School (private)	85.1	83.9
Teacher (public)[2]	91.5	90.3
Teacher (private)[2]	83.1	83.6
Teacher Follow-up Survey(TFS)		
Current (public)	not avail.	97.4
Current (private)		92.4
Former (public)		96.2
Former (private)		94.1
Postsecondary Education		
National Household Education Survey (NHES)	82.1[3]	not avail.
National Postsecondary Studen t Aid Study (NPSAS)		
Institutions	95	89
Students	77	76
National Survey of Postsecondary Faculty(NSOPF)	not avail.	not avail
Recent College Graduates Study (RCG)	83.1	83.2
Educational Assessment		
National Assessment of Educational Progress (NAEP)		
School	86.0	not avail.
Student	87.4	not avail.
National Adult Literacy Survey (NALS)	89.1	not avail.
National Longitudinal Studies		
Baccalaureate and Beyond Longitudinal Study (B&B)	85.4	not avail.
Beginning Postsecondary Student Longitudinal Study (BPS)	96.1	not avail.
High School and Beyond (HS&B)	not avail.	86.1
National Education Longitudinal Survey of 1988 ((NELS:88)	92.5	91.5
National Longitudinal Study of 1972 (NLS-72) Target Sample (4th Follow-up)	89.3	not avail.

[1] Combined School and TDS [3] Using Business office method

[2] Percent of eligible teachers in sample responding

L. Has any information on nonrespondents been included on the data file?

Eight out of the 13 NCES surveys examined include information on nonrespondents. For five out of those eight (SASS, TFS, BPS, HS&B, and B&B) however, this information was contained only on the restricted-use data file. Only two surveys (RCG and NELS:88) contain information on nonrespondents on the public-use data file.

Analysis of Response Categories

Researchers using more than one NCES database soon discover that there is minimal uniformity in demographic data collected: either the question wording or the response categories differ. We have identified nine common demographic survey items, five representing institutional characteristics and four representing individual characteristics.

Institutional variables

A. Sector

Twelve of the 13 NCES surveys collected data about the school sector (public, private, etc.). Unlike other variables described in this chapter, school sector was often not directly asked to respondents, but was a sampling frame variable. School sector was asked on five of the NCES surveys examined (NELS:88, RCG, HS&B, B&B, and NHES).

B. Region

Four of the 13 NCES surveys examined did not provide a region designator (NSOPF, RCG, BPS, and B&B). The remaining nine surveys may be divided into five categorization schemes: Four surveys--SASS, TFS, NALS, NHES--used the FIPS (Federal Information Processing Standards) categorization: Northeast, Midwest, South, and West. NAEP, NELS:88, and NLS-72 also provided four categories, but used slightly different categories. One of the two region categories provided on HS&B also provides four categories. NAEP used Northeast, Southeast, Central, and West, however, the part of Virginia that is included in the Washington, DC metropolitan statistical area (MSA) is included in the Northeast region, while the remainder of the state is included in the Southeast region. NELS:88, NLS-72, and NHES use Northeast, North Central, South, and West.

C. Urbanicity/locale

Six surveys provided documentation on urbanicity/ locale (NLS-72, NELS:88, NAEP, TFS, HS&B, and NHES). Three of these surveys (NLS-72, TFS, and HS&B) presented very similar categories: a rural or farming community, a small city or town of fewer than

50,000 people that is not a suburb of a larger city, a medium-sized city (50,000 to 100,000 people), a suburb of a medium-sized city, a large city (100,000 to 500,000 people), a suburb of a large city, a very large city (over 500,000 people), and a suburb of a very large city. TFS also included Indian reservation and military base, while HS&B only included military base. NELS:88 included only three categories: urban, suburban, and rural, developed from a composite variable created directly from QED (Quality Education Data), using the FIPS designator, utilized by the U.S. Census. NAEP collapses an urbanicity/locale variable into three categories: urban, suburban and rural. It also provides more detailed categories based on 1980 Census information. These categories included: rural, disadvantaged urban, advantaged urban, big city, fringe, medium city, and small place. Since 1990, SASS has replaced the self-reported community type with a 7-category scheme determined by the ZIP Code of the school and matched to the Census community size for that ZIP Code (Johnson, 1989).

D. School level

School level identifies whether the school is primary, secondary, or a combination of the two. (This analysis is not applicable to postsecondary schools.) Only three of the 13 NCES surveys examined (NELS:88, NHES, and SASS) provided such designation and all use different categories. NELS:88 does not provide school level exactly, but classifies the type of school by the grades spanned, which were collapsed into seven categories, using school data first. NHES classifies by lowest grade (prekindergarten to 11th) and highest grade (3rd to 12th). SASS provided four choices: elementary (if the school has only grades below 8th grade), middle school/junior high, secondary (if the school has grades between 7th and 12th, and combined elementary and secondary (if the school has any other combination of grades).

E. School/Institution size

Six surveys provided information on school/institution size: no two surveys used the same categories. NLS-72 indicated school size by enrollment of seniors--less than 400 or greater than 400. NELS:88 provided a composite variable, categorizing the entire school enrollment as reported by the school. These values were 1-199, 200-399, 400-599, 600-799, 800-999, 1000-1199, and 1200^{+}. On the public school questionnaire, SASS asked for the total number of students enrolled in grades K-12 or comparable ungraded levels. RCG had three categories: less than 1,500, 1,500 to 5,999, and 6,000 or more. NPSAS set its categories at less than 1,000, 1,000 - 2,499, 2,500 - 4,999, 5,000 - 9,999, 10,000 - 19,999, and 20,000 or more. NHES defined school size as under 300, 300 - 599, 600 - 999, and 1,000 or more.

Individual characteristics

F. Race/ethnicity

All 13 NCES surveys inquired about respondents' race; however, differences were found in categories from one survey to another. The first difference is the order of the race response categories. Some surveys begin with a minority response category such as black, American Indian, Asian, etc. (NLS-72, NSOPF, NELS:88, and HS&B), while others begin the response categories with white. Six surveys (NLS-72, NAEP, RCG, BPS, HS&B, and NHES) provide an other race category, while the remaining surveys do not. Race categories also varied by whether a Hispanic item was provided. RCG and NAEP combine race and Hispanic origin, e.g., white, non-Hispanic. Seven surveys ask for race information, followed by asking about Hispanic origin. On SASS, TDS (Teacher Demand and Shortage) and the School Survey include Hispanic origin as part of the race item, while the Administrator Survey and Teacher Survey ask this item separately. HS&B provides Hispanic as a type of race, not distinguishing white, black or other race. Only NLS-72 does not include a Hispanic designator.

G. Socioeconomic status

Surveys rarely ask respondents to provide their socioeconomic status (SES). Instead, this variable was constructed by combining various sociological and economic data. Only two surveys (NLS-72 and NELS:88) provided a specific SES composite variable on the data file. For NLS-72, SES was derived from an equally weighted linear composite of father's education, mother's education, father's occupation, family income, and household items (such as newspaper, dictionary, encyclopedia, etc.) from the first follow-up and/or base year student questionnaire. NELS:88 used the same composite variables; however, mother's occupation was used, rather than household items. The remaining surveys do not contain an SES composite.

H. Degree

All NCES surveys examined inquired about respondents' level of education/degree. However, a variety of different questions and response categories were used to gather them. In general, we may group survey responses into three major categories: responses with detailed lower degree levels, responses with detailed higher degree levels, and those with broad categories. Surveys with detailed lower degree levels include NPSAS, BPS, NALS, and NHES. Surveys with

detailed higher response categories include NLS-72 and B&B. NSOPF used seven detailed categories utilizing not only the names of various degrees, but mentioning words such as equivalent or certificate as completing one's degree. TFS used six categories: associate's degree, bachelor's degree, master's degree, doctorate, education specialist or professional diploma, or professional degree. Three surveys used broad categories. NELS:88 offered three: less than a bachelor's, bachelor's, and master's. RCG provided three categories: bachelor's, master's, or some other degree. NAEP provided four categories (among parents education): did not finish high school, graduated high school, some college, graduated college, or don't know. SASS asked about degree types earned on two of its surveys, the Teacher Survey and the Administrator Survey.

I. Respondents' Age Group

ALL NCES surveys inquired about age, but few provided age groupings. Only NAEP provided for age groupings for children, specifically, students who were either in the fourth grade or 9 years old; students who were either in the eighth grade or 13 years old; and students who were either in the twelfth grade or 17 years old. On the TFS survey, the restricted use file provided actual ages; however, the public release file provided four categories: Under 30, 30 to 39, 40-49, and 50 and above. On RCG, actual ages are provided for respondents and categories are provided for newly qualified teacher of: 23 or younger, 24 to 25, and 26 or older. All other surveys inquired about respondents' exact year of birth or actual ages so that researchers may combine specific ages and convert them to age groupings.

Conclusions

This paper examined two major topics: consistency of response rates information/calculation and consistency of response categories. Most NCES surveys provided detailed information on unit and item response rates and defined these consistently across surveys. The amount of documentation on the intensive follow-up of nonrespondents was minimal. Some of the response categories showed large variation across surveys, such as those used for urbanicity and race/ethnicity. Different questionnaire wording (some of which also had different response categories), were also prevalent especially for those used for degree and race/ethnicity.

Recommendations and suggestions

Several additional studies could be explored to further elaborate on information provided in this report: 1) more efforts are needed to examine the impact of response rates on baseline statistics related to two major issues: what bias is generated by differential nonresponse rates on estimates of school resources and student outcomes across geographic or socioeconomic categories? How much bias can be measured or adjusted, if differential response rates are found? 2) the most recent surveys could be considered for the nonresponse issues since response rates change over time due to different reasons. Techniques for calculating response rates may change over time, too. Higher nonresponse rates might be due to the mode of administration or economic status of respondents. These issues could be addressed in the further studies. 3) Additional response categories may be examined, such as Likert scales (3- 5- or 7-point, response categories from low to high or vice versa, etc.).

References

Abraham, S., Suter, N., Spencer, B., Johnson, R., Zahs, D., Myers, S. and Zimbler, L. (1994), 1992-92 National Study of Postsecondary Faculty Field Test Report. Technical Report NCES 93-390. U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

Council of American Survey Organizations (CASRO) (1982), "On the Definitions of Response Rates." Port Jefferson, New York.

Ingels, S., Dowd, K., Baldridge, J., Stipe, J., Bartot, V. and Frankel, M. (1994), "Second Follow-up: Student Component Data File User's Manual, Volume I." NCES 94-374. U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

Johnson, F., (1989), "Assigning Type of Locale Codes to the 1987-1988 Public School Universe," Technical Report: Assigning Type of Locale.

Kalton G. (1983). Compensating for Missing Survey Data, Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.

Westat, Inc. (1994), "1991 Survey of Recent College Graduates: Methodology Report." Washington, DC: National Center for Education Statistics.

MULTIVARIATE MODELING OF UNIT NONRESPONSE FOR 1990-91 SCHOOLS & STAFFING SURVEYS

Sameena Salvucci, Fan Zhang, and David Monaco, Synectics for Management Decisions Inc., Kerry Gruber, National Center for Education Statistics, Fritz Scheuren, George Washington University
Fan Zhang, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd #305, Arlington VA 22201

KEY WORDS: Nonresponse modeling, Complex sampling design, Hierarchical response patterns, Multivariate modeling

I. Introduction

This paper presents selected results of a study which analyzed unit nonresponse for the components of the 1990-91 Schools and Staffing Surveys (SASS): schools, principals, teachers, and school districts. SASS is a periodic, integrated system of sample surveys on elementary and secondary schools in the United States sponsored by the National Center for Education Statistics (NCES) of the U.S. Department of Education and administered by the U.S. Bureau of the Census.

The study was motivated by the need to identify potential sources of nonsampling error in the SASS estimates associated with nonresponse. Nonresponse is a concern depending on the amount of incompleteness that exists in the data and the difference in the characteristics between respondents and nonrespondents. We developed a multivariate model of unit nonresponse to try to explain the relationship of these factors to the level of unit nonresponse for each of the components of SASS. We also studied the results of the modeling effort across the SASS components.

One of the reasons that it is so hard to evaluate nonsampling error from unit nonresponse in a survey is the lack of data from nonrespondents, which is critical in the evaluation. As a result, the scope of our study is limited to the few frame variables for which data were collected for all sampled schools, teachers, administrators, and districts. It was conjectured that these variables might have a plausible effect on nonresponse. As will be seen, this conjecture was at least somewhat optimistic.

II. The Surveys and Sample Design

SASS is comprised of four interrelated national surveys:

1. The School Survey included data on school programs and services, student characteristics and staffing patterns. For private schools additional information was collected on aggregate demand for both new and continuing teachers.
2. The School Administrator Survey collected background information from principals on their education, experience, and compensation, and their perceptions of the school environment and educational goals.
3. The School Teacher Survey collected information on demographic characteristics of public and private school teachers, their education, qualifications, income sources, working conditions, plans for the future, and perceptions of the school environment and the teaching profession.
4. The Teacher Demand and Shortage Survey (TDS) targeted public school district personnel who provided information about their district's student enrollment, number of teachers, position vacancies, new hires, teacher salaries and incentives, and hiring and retirement policies.

The target populations for the 1990-91 SASS surveys included U.S. elementary and secondary public and private schools with students in any of grades 1-12, principals and classroom teachers in those schools, and local education agencies (LEAs) that employed elementary and/or secondary level teachers. (In the private sector, since there is no counterpart to the LEAs, information on teacher demand and shortages was collected directly from individual schools. Nonresponse in the Teacher Demand and Shortage data was analyzed for the public sector only.)

Three primary steps in the sample selection process were followed during the 1990-91 SASS. The School Survey sample forms the basis for all other survey samples.

1. A sample of schools was selected first for the School Survey. The same sample was used for the School Administrator Survey.
2. For each school in the School Survey, a list of teachers was obtained from which a sample was selected for inclusion in the Teacher Survey.
3. The sample for the Teacher Demand and Shortage Survey was formed from responses from all private schools selected in the School Survey and all LEAs administrating public schools already in the School Survey sample.

Details pertaining to the frame, stratification, sorting, and sample selection for each of the four surveys of SASS are presented in Kaufman and Huang (1993).

III. Weighted Unit Response Rates

For each survey of SASS, weighted unit response rates were calculated. The weighted unit response rates were derived by dividing the sum of the weights for the interviewed cases by the sum of the weights for the eligible cases (the number of sampled cases minus the number of out-of-scope cases). In other words, the weighted unit response rate specifies what proportion of a population is covered by the respondents.

The simplest weighted response rate uses the unit of collection as the population. However, other populations can be used. For example, the public school survey collects many characteristics: some are specific to the school as an entity; some relate to the teaching staff or to the student body. Therefore, for the School and Administrator components of SASS, three alternative adjusted response rates were calculated:

- School-based response rate: This measure is calculated by weighting the responding (R) and nonresponding schools (N) by the inverse of their base sample selection probabilities (or base weights). Once the schools are so weighted, the rates are determined for each group being considered by calculating the ratios R/(R+N) and multiplying by 100 to convert them to percents. For example, a 90% school-based response rate for the public school survey means that 90% of public schools are covered by the respondents.
- Teacher-based response rate: This measure is calculated in the same way as the school-based response rate, except a school's base weight is multiplied by the number of teachers in the school before calculating the response ratio as above. For example, a 90% teacher-based response rate for the public school survey means that 90% of the teacher population is covered by the responding schools.
- Student-based response rate: This measure is calculated in the same way as the school-based response rate, except a school's base weight is multiplied by the number of students in the school before calculating the response ratio.

Similarly, LEA-based, school-based and student-based weighted unit response rates were calculated for the Teacher Demand and Shortage component of SASS. However, for the teacher component only one weighted unit response rate was calculated using an adjusted base weight.

For each of the SASS surveys, the three different weighted response rates were examined graphically and it was determined that little differences existed between the simple weighted response weight and the alternative measures. Therefore, for modeling purposes we confined our analysis to using the most simple weighted response rate, i.e., using the unit of collection as the population.

Overall unit response rates are high for the SASS surveys and, as expected, better for the public rather than for the private component (see table 1). However, unit nonresponse remains a concern because of the complex, hierarchical nature of the SASS design, and there is room for improvement (Moonesinghe, Smith and Gruber, 1993). Also, unit response rates vary considerably across the states within each of the public surveys and across affiliations within each of the private surveys (see "highest" and "lowest" columns in table 1).

Table 1: Response Rates for 1990-91 SASS Surveys

Survey Component	Overall	Highest	Lowest
Public School	95.30	99.61	80.99
Private School	83.95	97.89	59.03
Public Administrator	96.69	100.00	82.35
Private Administrator	90.05	98.85	72.39
Public Teacher	90.33	97.88	69.40
Private Teacher	84.31	94.83	57.12
TDS (public LEAs)	93.49	100.00	76.96

Multiple regression techniques were employed in order to examine the combined effects of other stratification variables, such as urbanicity. school size, and school level within each of the components.

IV. Methodology

Exploratory Analysis:

In the first stage of this study we undertook an exploratory analysis of unit nonresponse behavior within each of the SASS surveys. We focused only on a limited number of variables for which we conjectured a plausible effect on response rates, and used comparable, simple structure, complete logistic regression models for each analysis. Here the goal was to develop a model of response rates by state or affililiation for each of the survey components-- not just to see how frame variables such as urbanicity vary in their effects by state or affiliation. We used the simple base school weight divided by the mean base weight for the state for public components and the simple base school weight divided by the mean base weight for the affiliation for private components. We modeled nonresponse on urbanicity, school level, school size for the School, Administrator, and Teacher surveys and on urbanicity, number of schools in the LEA, number of students in the LEA for the School Teacher Demand and Shortage survey. Within each of the survey components, we selected a final model which included an additional categorical variable which grouped either states or affiliations into clusters through a stepwise, modeling procedure. The objective was to

reduce the variability in response due to the states/associations in order to concentrate on the variation caused by the frame variables.

The stepwise modeling procedure began by fitting the data to a complete, baseline model which contained all categorical frame variables for each of the states/affiliations separately without any clustering. No interactions were modeled. The goodness-of-fit of the model fitted was evaluated on the basis of how well it estimated response at the state/affiliation level. A t-value was calculated for each state using the observed and fitted response rate. The variance was adjusted using the average design effect for proportions at the state/affiliation level (Salvucci and Weng, 1995) as follows:

$$\frac{\text{Response Rate - Estimated Response Rate}}{\sqrt{(\text{Design Effect})\dfrac{(\text{Response Rate})(1 - \text{Response Rate})}{\text{Sample Size}}}}$$

The design effects used in the calculation of the t-values for states or affiliations within each of the survey components in this stepwise modeling process were:

Public School Survey:	1.7433
Public School Administrator Survey:	1.7807
Public School Teacher Survey:	2.8493
Private School Survey:	2.0488
Private School School Administrator Survey:	2.3694
Private School Teacher Survey:	1.9053
TDS (public LEA's):	1.8603

Successive models fitted included all frame variables and differed only in how they clustered states/affiliations into groups The criterion used for segregating states/affiliations in the successive models was that the t-value be smaller than -2 or greater than 2 -- a two-tail t-test at the .05 percent significance level. If the t-value criterion by state/association cluster was not violated the modeling procedure was terminated; otherwise the plot of the estimated response rate versus the actual response rate was used to identify outliers, the clusters were redefined, a new model was fitted, and the cycle was repeated.

Final model specifications

Final logistic regression models (developed as above) were fitted for each of the surveys. These involved all of the frame variables studied for the particular survey and an additional categorical variable which divided states or affiliations into clusters for the public and private component surveys respectively.

For example the final multiple logistic model used for the Public School Survey was:

$$g(x) = b_0 + \sum_{i=1}^{2} b_{1i} x_{1i} + \sum_{j=1}^{2} b_{2j} x_{2j} + \sum_{k=1}^{3} b_{3k} x_{3k} + \sum_{g=1}^{4} b_{4g} x_{4g}$$

where $P(Y=1|x) = \pi(x)$ is defined as the conditional probability that the outcome is present and

$$\pi(x) = \frac{e^{g(x)}}{1 - e^{g(x)}}$$

where x_{1i}, i=1,2, 3 are the variables coding urbanicity, x_{2j}, j=1,2,3 the variables coding school level, x_{3k}, k=1,2,3,4 the variables coding school size and x_{4g}, g=1,2,..,m the variables coding state/affiliation groupings. No variable interactions (the combined effect of two or more variables) entered into the model.

V. Findings

For each final model we fitted the odds ratios to define more closely subpopulations with significant nonresponse differentials. A summary of our findings follows:

- For the public component of the School Administrator, School and Teacher surveys, only urbanicity and state were significant in modeling unit nonresponse. (Tables 2-7)

Table 2: State cluster odds ratios, Public School Survey

- Group 1: The District of Columbia, Maryland, New Jersey, New York (81.0% through 88.3%)
- Group 2: Alaska, Massachusetts (91.1% through 92.0%)
- Group 3: Hawaii, Illinois, Indiana, Utah (98.7% through 99.6%)
- Group 4: Connecticut, Delaware, North Carolina, Virginia, Washington (92.2% through 93.3%)
- Referent Group: The Remaining 36 States (93.9% through 98.7%)

(in parentheses are the response rate intervals for the cluster)

Group Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Group 1 vs Referent	0.27	0.17	0.44
Group 2 vs Referent	0.38	0.17	0.83
Group 3 vs Referent	3.53	0.92	13.55
Group 4 vs Referent	0.49	0.28	0.86
Group 1 vs Group 2	0.73	0.31	1.72
Group 1 vs Group 3	0.08	0.02	0.31
Group 1 vs Group 4	0.56	0.29	1.07
Group 2 vs Group 3	0.11	0.02	0.49
Group 2 vs Group 4	0.77	0.31	1.92
Group 3 vs Group 4	7.25	1.76	29.82

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Public School Questionnaires).

Table 3: Urbanicity odds ratios, Public School Survey

Urbanicity Type Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Urban Fringe / Large Town vs Rural / Small Town	0.52	0.35	0.79
Central City vs Rural / Small Town	0.47	0.31	0.70
Urban Fringe / Large Town vs Central City	1.12	0.75	1.66

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Public School Questionnaires).

Table 4: State group odds ratios for the Public School Administrator Survey

- Group 1: The District of Columbia, Maryland, New York (82.3% through 89.5%)
- Group 2: Idaho, Illinois, Indiana, Montana, Utah, West Virginia (99.3% through 100.0%)
- Group 3: Louisiana, New Jersey, Washington (92.4% through 93.7%)
- Referent Group: The Remaining States (94.4% through 99.2%)

Group Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Group 1 vs Referent	0.21	0.12	0.36
Group 2 vs Referent	10.77	1.12	103.95
Group 3 vs Referent	0.38	0.20	0.73
Group 1 vs Group 2	0.02	0.00	0.19
Group 1 vs Group 3	0.55	0.26	1.17
Group 2 vs Group 3	28.35	2.77	290.31

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Public School Administrator Questionnaires).

Table 5: Urbanicity odds ratios for the Public School Administrator Survey

Urbanicity Type Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Urban Fringe / Large Town vs Rural / Small Town	0.50	0.29	0.86
Central City vs Rural / Small Town	0.33	0.20	0.55
Urban Fringe / Large Town vs Central City	1.52	0.94	2.44

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Public School Administrator Questionnaires).

Table 6: State group odds ratios for the Public Teacher Survey

- Group 1: The District of Columbia, New York (68.5% through 79.6%)
- Group 2: Alabama, Alaska, California, Connecticut, Florida, Hawaii, Kentucky, Maryland, Nevada, New Mexico, New Jersey, Ohio, Rhode Island, Washington (86.3% through 91.0%)
- Group 3: Massachusetts, Michigan (84.3% through 84.8%)
- Group 4: Illinois, Utah (96.4% through 97.7%)
- Group 5: Texas, Virginia (91.6% through 91.7%)
- Referent Group: The Remaining States (89.6% through 96.5%)

Group Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Group 1 vs Referent	0.24	0.17	0.33
Group 2 vs Referent	0.50	0.41	0.60
Group 3 vs Referent	0.34	0.25	0.48
Group 4 vs Referent	2.06	1.11	3.84
Group 5 vs Referent	0.72	0.51	1.00
Group 1 vs Group 2	0.48	0.35	0.66
Group 1 vs Group 3	0.70	0.46	1.07
Group 1 vs Group 4	0.12	0.06	0.23
Group 1 vs Group 5	0.34	0.22	0.51
Group 2 vs Group 3	1.45	1.05	2.01
Group 2 vs Group 4	0.24	0.13	0.45
Group 2 vs Group 5	0.70	0.50	0.97
Group 3 vs Group 4	0.17	0.08	0.33
Group 3 vs Group 5	0.48	0.31	0.74
Group 4 vs Group 5	2.88	1.46	5.69

Source: U.S. Department of Education, National Center for Education Statistica, Schools and Staffing Surveys: 1990-91 (Public School Teacher Questionnaires)

Table 7: Urbanicity odds ratios for the Public School Teacher Survey

Urbanicity Type Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Urban Fringe / Large Town vs Rural / Small Town	0.74	0.62	0.88
Central City vs Rural / Small Town	0.63	0.53	0.75
Urban Fringe / Large Town vs Central City	1.17	0.98	1.39

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Public School Teacher Questionnaires)

- For the private components of the School Administrator and Teacher surveys, only affiliation was significant. (Tables 8-9)

Table 8: Affiliation group odds ratios for the Private School Administrator Survey

- Group 1: Area Frame, National Society for Hebrew Day Schools, Other Jewish, American Association of Christian Schools, All Else (72.4% through 86.1%)
- Group 2: Solomon Schechter Day Schools, Lutheran Church - Missouri Synod, Evangelical Lutheran Church - Wisconsin Synod, Evangelical Lutheran Church in America, Other Lutheran (97.3% through 98.9%)
- Referent Group: Association of Military Colleges and Schools of U.S., Catholic, Friends, Episcopal, Seventh-Day Adventist, Christian Schools International, National Association of Private Schools for Exceptional Children, Montessori, National Association of Independent Schools (92.2% through 96.2%)

Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Group 1 vs Referent	0.28	0.16	0.50
Group 2 vs Referent	1.98	0.53	7.44
Group 1 vs Group 2	0.14	0.04	0.52

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Private School Administrator Questionnaires)

Table 9: Affiliation odds ratios for the Private School Teacher Survey

- Group 1: National Society for Hebrew Day Schools, Other Jewish, American Association of Christian Schools (59.8% through 63.5%)
- Group 2: Association of Military Colleges and Schools of U.S., Lutheran Church - Missouri Synod, Evangelical Lutheran Church - Wisconsin Synod, Other Lutheran, Christian Schools International (90.3% through 94.8%)
- Group 3: Area Frame, Montessori (75.0% through 76.9%)
- Group 4: Catholic, Solomon Schechter Day Schools (85.7% through 88.0%)
- Referent Group: Friends, Episcopal, Evangelical Lutheran Church in America, Seventh-Day Adventist, National Association of Private Schools for Exceptional Children, National Association of Independent Schools, All Else (79.2% through 86.0%)

Group Comparison	Odds Ratio	Confidence Interval	
		Lower 95%	Upper 95%
Group 1 vs Referent	0.37	0.25	0.54
Group 2 vs Referent	2.37	1.46	3.83
Group 3 vs Referent	0.68	0.48	0.96
Group 4 vs Referent	1.44	1.03	2.02
Group 1 vs Group 2	0.16	0.09	0.27
Group 1 vs Group 3	0.54	0.35	0.83
Group 1 vs Group 4	0.26	0.17	0.39
Group 2 vs Group 3	3.49	2.10	5.79
Group 2 vs Group 4	1.64	1.00	2.69
Group 3 vs Group 4	0.47	0.32	0.70

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Private School Teacher Questionnaires)

- For the Private School survey, the significant variables were affiliation and school level. (Tables 10-11)

Table 10: Affiliation odds ratios for the Private School Survey

- Group 1: Area Frame, National Society for Hebrew Day Schools, Other Jewish, American Association of Christian Schools (59.0% through 74.0%)
- Group 2: Lutheran Church - Missouri Synod, Evangelical Lutheran Church - Wisconsin Synod, Evangelical Lutheran Church in America (95.5% through 97.9%)
- Group 3: Solomon Schechter Day Schools, National Association of Private Schools for Exceptional Children, Montessori, National Association of Independent Schools, All Else (81.1% through 86.5%)
- Referent Group: Association of Military Colleges and Schools of U.S., Catholic, Friends, Episcopal, Other Lutheran, Seventh-Day Adventist, Christian Schools International (89.4% through 94.2%)

Group Comparison	Odds Ratio	Confidence Interval Lower 95%	Confidence Interval Upper 95%
Group 1 vs Referent	0.24	0.14	0.42
Group 2 vs Referent	2.35	0.63	8.68
Group 3 vs Referent	0.54	0.31	0.96
Group 1 vs Group 2	0.10	0.03	0.38
Group 1 vs Group 3	0.45	0.28	0.73
Group 2 vs Group 3	4.33	1.16	16.10

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Private School Questionnaires).

Table 11: School level odds ratios for the Private School Survey

School Level Comparison	Odds Ratio	Confidence Interval Lower 95%	Confidence Interval Upper 95%
Elementary vs Combined	1.53	1.03	2.27
Secondary vs Combined	2.35	1.05	5.26
Elementary vs Secondary	0.65	0.29	1.45

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Private School Questionnaires).

- For the public component of the Teacher Demand and Shortage Survey the significant variables were state and the number of students in the LEA. (Tables 12-13)

Table 12: State group odds ratios for the Teacher Demand and Shortage Survey

- Group 1: Connecticut, Maryland, New Jersey, Vermont (77.0% through 87.5%)
- Group 2: Delaware, The District of Columbia, Kansas, Nevada, Tennessee, Colorado, Iowa, Nebraska, Oklahoma, South Dakota, West Virginia, Washington (97.0% through 100.0%)
- Group 3: California, Montana, North Dakota, Oregon (91.2% through 95.1%)
- Referent Group: The Remaining States (90.1% through 100.0%)

Group Comparison	Odds Ratio	Confidence Interval Lower 95%	Confidence Interval Upper 95%
Group 1 vs Referent	0.40	0.20	0.81
Group 2 vs Referent	4.78	1.51	15.12
Group 3 vs Referent	1.27	0.56	2.87
Group 1 vs Group 2	0.08	0.02	0.30
Group 1 vs Group 3	0.32	0.12	0.82
Group 2 vs Group 3	3.76	1.00	14.07

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Teacher Demand and Shortage Survey Questionnaires).

Table 13: Odds ratios for the number of students in local education agency for the Teacher Demand and Shortage Survey

Number of Students in LEA Comparison	Odds Ratio	Confidence Interval Lower 95%	Confidence Interval Upper 95%	Number of Students in LEA Comparison	Odds Ratio	Confidence Interval Lower 95%	Confidence Interval Upper 95%
0 to 299 vs 25,000 Plus	0.62	0.09	4.47	300 to 599 vs 1,000 to 2,499	1.14	0.42	3.04
300 to 599 vs 25,000 Plus	1.73	0.22	13.45	300 to 599 vs 2,500 to 4,999	1.1	0.30	4.07
600 to 999 vs 25,000 Plus	1.18	0.16	8.91	300 to 599 vs 5,000 to 9,999	1.31	0.29	5.99
1,000 to 2,499 vs 25,000 Plus	1.52	0.23	9.89	300 to 599 vs 10,000 to 24,999	1.00	0.16	6.40
2,500 to 4,999 vs 25,000 Plus	1.57	0.28	8.79	600 to 999 vs 1,000 to 2,499	0.78	0.30	1.98
5,000 to 9,999 vs 25,000 Plus	1.32	0.23	7.74	600 to 999 vs 2,500 to 4,999	0.75	0.21	2.68
10,000 to 24,999 vs 25,000 Plus	1.73	0.24	12.57	600 to 999 vs 5,000 to 9,999	0.89	0.20	3.96
0 to 299 vs 300 to 599	0.36	0.14	0.90	600 to 999 vs 10,000 to 24,999	0.68	0.11	4.24
0 to 299 vs 600 to 999	0.53	0.21	1.29	1,000 to 2,499 vs 2,500 to 4,999	0.97	0.34	2.78
0 to 299 vs 1,000 to 2,499	0.41	0.18	0.90	1,000 to 2,499 vs 5,000 to 9,999	1.15	0.32	4.16
0 to 299 vs 2,500 to 4,999	0.40	0.12	1.29	1,000 to 2,499 vs 10,000 to 24,999	0.88	0.17	4.63
0 to 299 vs 5,000 to 9,999	0.47	0.11	1.93	2,500 to 4,999 vs 5,000 to 9,999	1.18	0.40	3.50
0 to 299 vs 10,000 to 24,999	0.36	0.06	2.10	2,500 to 4,999 vs 10,000 to 24,999	0.91	0.20	4.06
300 to 599 vs 600 to 999	1.47	0.50	4.30	5,000 to 9,999 vs 10,000 to 24,999	0.77	0.16	3.63

Source: U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Surveys: 1990-91 (Teacher Demand and Shortage Survey Questionnaires).

VI. Conclusions

Our study carries implications for handling nonresponse during data collection -- by either undertaking intensive follow up studies where nonresponse is significant or simply increasing sample size where nonresponse is random -- and/or the analysis level -- by adjusting for nonresponse along significant variables. For example, our results focus attention on the states for the public school survey and affiliations for the private school survey as variables accounting for variation in nonresponse. The U.S. Bureau of the Census has rightfully selected these variables for nonresponse adjustments, as already noted by Shen, Palmer and Tan (1992) in a similar study.

The results of our study, however, suggest that variability in nonresponse can be accounted for by only those variables which were shown to be significant in our modeling. Given these findings some re-evaluation might be in order on how nonresponse adjustments are made with regard to variables from which nonresponse bias does not appear to arise. For example, adjusting for school level in the Public School Administrator, School and Teacher surveys might lead to overadjustments if one considers the results of our analysis which suggest that variation in nonresponse along this particular variable may be random when adjustments are made for state clusters and urbanicity.

Our study, although preliminary, shows how statistical modeling can be of assistance in defining subpopulations with nonresponse differential. Nonresponse bias can then be reduced using poststratification techniques. Further statistical modeling examining the effect of additional covariates should lead to a better understanding of unit nonresponse. This will have considerable practical consequences for improving the SASS data base at the collection stage and for adjusting for nonresponse while conducting analysis.

Bibliography

Fay, R. (1986), "Causal models for patterns of nonresponse." *Journal of the American Statistical Association* 81: 354-365.

Groves, R. M. (1989), *Survey errors and survey costs*. New York: John Wiley & Sons.

Hosmer, D. W. and Lemeshow, S. (1989), *Applied Logistic Regression*. New York: John Wiley & Sons.

Jabine, T. (1994), *Quality Profile for SASS, Aspects of the Quality of Data* in *the Schools and Staffing Surveys (SASS)*. Technical Report, NECS 94-340. Washington, DC: National Center for Education Statistics.

Kaufman, S. and Huang, H. (1993), *1991 Schools and Staffing Survey: Sample Design and Estimation*, Technical Report, NECS 93-449, Washington, DC: National Center for Education Statistics.

Madow, W. G., Nisselson, H. and Olkin, I. (1983), *Incomplete Data in Sample Surveys, Vol. 1, Report and Case Studies*. New York: Academic Press.

Madow, W.G. and Olkin, I. (1983), *Incomplete Data in Sample Surveys, Vol. 3, Proceedings of the Symposium*. New York: Academic Press.

Madow, W. G., Olkin, I. and Rubin, D. B. (eds) (1983), *Incomplete Data in Sample Surveys, Vol 2, Theory and Bibliographies*. New York: Academic Press.

Smith, W., Moonesinghe, R., Smith, W. and Gruber, K. (1993), "Characteristics of Nonrespondents to the 1990-91 Schools and Staffing Survey." *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 698-703. Alexandria, VA: American Statistical Association.

National Academy of Sciences (U.S.) (1992), *Combining Information. Statistical Issues and Opportunities for Research in the Combination of Information*. Washington, D.C.: National Academy Press.

Oh, L. H. and Scheuren, F. J. (1983), "Weighting Adjustment for Unit Nonresponse," in Madow, W. G., Olkin, I. and Rubin, D. B. (eds), *Incomplete Data in Sample Surveys, Vol 2, Theory and Bibliographies*, pp. 143-184. New York: Academic Press.

Pregibon, D. (1984), "Logistic Regression Diagnostics," *Annals of Statistics* 9.

Salvucci, S. and Weng, S. (1995), *Design Effects and Generalized Variance Functions for the 1990-91 Schools and Staffing Surveys (SASS) - Volume 1*, Washington, DC: National Center for Education Statistics.

Shen, P., Palmer, R.J., and Tan, A.I. (1992), "Characteristics of Nonrespondents in the Schools and Staffing Surveys' School Sample," *Proceeedings of the Section on Survey Research Methods, American Statistical Association.* Alexandria, VA: American Statistical Association.

EVALUATION OF IMPUTATION METHODS FOR STATE EDUCATION FINANCE DATA

David Monaco and Stanley Weng, Synectics for Management Decisions, Inc., Frank Johnson, National Center for Education Statistics

David Monaco, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd #305, Arlington VA 22201

KEY WORDS: Missing values, Imputation distributions, Percent error, Log transformation

INTRODUCTION

The purpose of this study is to identify, develop, and analyze appropriate methods for imputing missing data in the National Public Education Financial Survey (NPEFS), collected by the National Center for Education Statistics (NCES).

NPEFS is part of the Common Core of Data (CCD), a series of surveys collected annually from State education agencies. NPEFS provides detailed State-level information about revenues and expenditures for public elementary and secondary education. These data are used to allocate $7 billion in federal funds for education to the states, therefore all states submit data for this survey. The need for imputation is not to correct for non-reporting states, but to correct for missing items in the states' submissions. The goal is a complete dataset that is comparable across states.

Each state has a unique accounting structure for tracking revenues and expenditures for public education. Even in states following the most recent 1990 accounting handbook there are revenues or expenditures which are reported as aggregate amounts with other items. NCES works with states to improve reporting and have developed state specific software to crosswalk finance data from states' accounting systems to NCES's. However even with these efforts, imputation operations were required for 37 states for the FY 1992 collection.

In most cases these imputations were used to disaggregate a single value reported for two or more items. For example a state may not distinguish between student fees for transportation, textbooks, and summer school but only track student fees in general which they might report as student fees for transportation because state officials know that transportation fees are larger than book or summer school fees. NCES would then perform an imputation to disaggregate the reported single value and distribute it to the three separate student fee items. NCES performed 148 separate imputation operations for the FY 1992 collection, of which 129 involved disaggregating a reported value to two or more items. The remaining operations involved imputing values for items that states do not track.

This study looks at the two similar methods for imputing data that were developed by NCES (NCES I and NCES II) along with a variation of this method (NCES III). In addition, time series, regression, and nearest neighbor methods are discussed.

These methods were analyzed in order to determine the affects of each and to select one method as being "better" in disaggregating the data. The analysis focuses on the distribution of Revenues from Nonproperty Taxes (R1D) to Revenues from Tuition (R1F) and Summer School (R1N). This particular operation was chosen because of the variability of these revenues across states. Unlike expenditures for education where the proportions spent for salaries, instruction, etc. are fairly consistent across states, revenues for education come through a variety of revenue collecting activities.

IMPUTATION METHODS

NCES I Imputation Method

The NCES I method for distributing aggregate amounts is to calculate a ratio of each appropriate item in the distribution to the sum of the items in the distribution. For example, one state reports tuition fees (R1F) and summer school fees (R1N) as a Non-property tax (R1D), then the ratio of R1D to the sum of R1D + R1F + R1N is calculated for each state reporting both items. The ratios of R1F to the sum, and R1N to the sum are also calculated and then the average of each set of ratios across states is determined. This ratio is then used to disaggregate the reported amount.

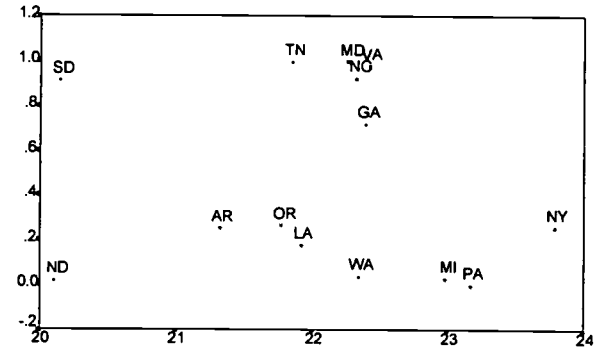
Table 1 demonstrates this method. State A is the state reporting the three revenues as R1D. States B through E etc. are the states whose reported amounts are used for the imputation. The R1D ratio for state B is 18.0/(18.0+3.2+3.9). The average ratio is for all of the states used in the imputation, of which only four are shown in the table. The average ratio times the amount reported for R1D yields the imputed amounts for each of the three variables (at the bottom of the table.)

Table 1. NCES I Method (amounts in $ millions)

State	R1D	R1F	R1N	R1D Ratio	R1F Ratio	R1N Ratio
State A	302.8	—	—			
State B	18.0	3.2	3.9	0.72	0.13	0.16
State C	1,069.1	3.2	2.5	0.99	0.01	0.00
State D	55.1	156.3	2.5	0.26	0.73	0.01
State E	500.5	1.1	2.3	0.99	0.00	0.06
ect.						
Average Ratio	—	—	—	0.42	0.51	0.06
State A Imputed	R1D 127.90	R1F 155.50	R1N 19.40			

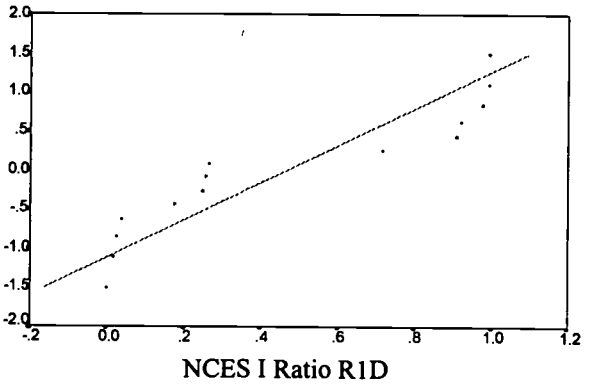
The ratio containing R1D is plotted against the natural logarithm of Total Revenues for every state. This gives us an indication of the characteristics of this model and the weight that the variable R1D has on the imputation involving the three variables (R1D, R1F and R1N). The plot for Method I is presented in Figure 1. The ratio plotted on the Y axis is that of R1D/(R1D+R1F+R1N), and the natural logarithm of Total Revenues is plotted on the x axis. Two groups of states are apparent, one group of six states where the ratio of R1D to the sum of the three variables is between .70 and 1.00 and another group where the ratio is .30 or less.

Figure 1. NCES I Method: Plot of R1D Ratio to Log of Total Revenues



The large gap that exists between the two groups of states would indicate that an average of these ratios would not be representative of the data for either of the two groupings. This conclusion is supported further by a normal probability plot (Figure 2), where the ratios are arranged in increasing order of magnitude and then plotted against normal distributed values.

Figure 2. NCES I Method: Normal probability plot of R1D ratio



If the data are from a normal distribution, this plot will resemble a straight line. The state with the lowest R1D value is approximately 1.5 standard deviations below the mean. The state reporting a slightly higher value for R1D is found to be slightly higher than 1 standard deviation below the mean, and so on. The resulting plot is curved. The departure of the data points from the straight line exhibits the departure of the data from normality.

NCES II Imputation Method

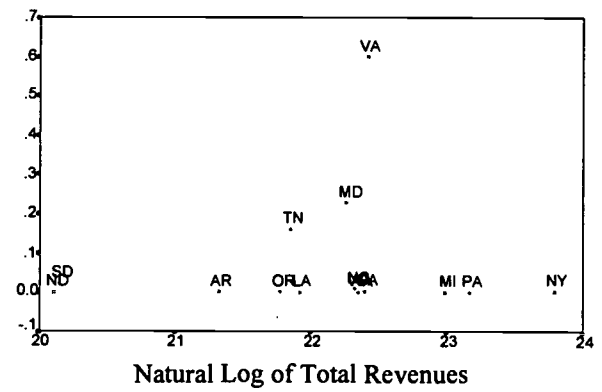
NCES II method was developed as an improvement over NCES I, but is very similar. We will use the same example where State A reports the value for R1F and R1N aggregated in the value reported for R1D. This time the ratios calculated are of each value divided by total revenues (TR). (If the items were expenditures the ratios would be calculated with total expenditures as the denominator.) Only states reporting values greater than 0 for each of the 3 revenues are used in the operation. States in which any of the 3 revenues are changed by other imputations are excluded from the operation. The average of these ratios is calculated, and then the relative distribution of these averages is determined. This distribution is then used to disaggregate the reported revenue amount.

Table 2. NCES II Method (amounts in $ million's)

State	RID	RIF	RIN	TR	RID Ratio	RIF Ratio	RIN Ratio
State A	302.8	—	—				
State B	18.0	3.2	3.9	5,332	0.00	0.00	0.00
State C	1,069.1	3.2	2.5	4,692	0.23	0.00	0.00
State D	55.1	156.3	2.5	21,574	0.00	0.01	0.00
State E	500.5	1.1	2.3	3,094	0.16	0.00	0.00
ect.							
Average Ratio	—	—	—	—	0.04	0.01	0.00
Percent distribution of avg. ratios					0.85	0.15	0.01
State A	RID	RIF	RIN				
Imputed	255.8	44.5	2.5				

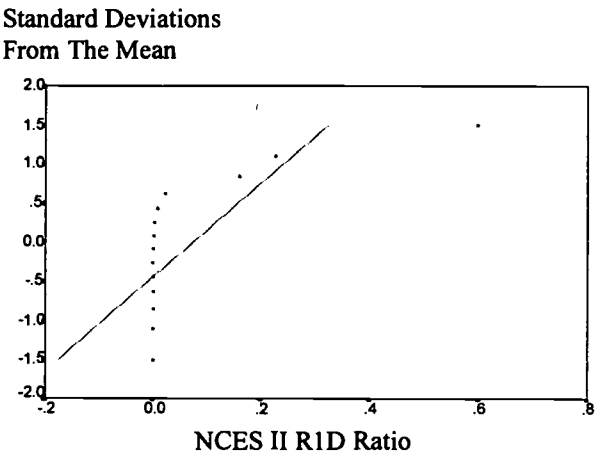
The plot of R1D/TR (NCES II ratios) by the natural logarithm of TR is presented in Figure 3. This plot shows most points are scattered about a horizontal level with a few outliers. The average ratio would shift from that stable level and therefore not represent the majority of the ratios.

Figure 3. NCES II Method: Plot of R1D ratio vs. Log of Total Revenues



In addition, the normal probability plot for the NCES II ratios for R1D shows a curve differing from the expected straight line, indicating the data are not normally distributed. (Figure 4)

Figure 4. NCES II Method, Normal probability plot of R1D ratio



NCES III Method

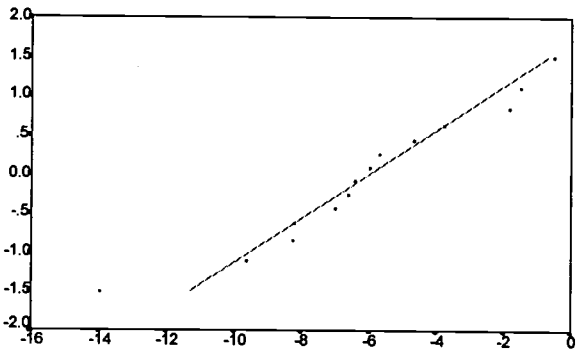
A variation of Method II was designed for this analysis. This method calculates the natural logarithms of the ratios (of item to total revenues (or expenditures)). The average of these logs is computed, and then the natural exponent of the average is determined. The distribution of these exponents is calculated, and the resulting values are used to distribute the aggregated amount. The log transformation of the ratios should stabilize the variance of the ratios. An example of the NCES III method is shown in Table 3. Note that the averages are calculated from more data than are shown.

Table 3. NCES III Method (amounts in $ million's)

State	RID	RIF	RIN	TR	RID Ratio	RIF Ratio	RIN Ratio
State A	302.8	—	—				
State B	18.0	3.2	3.9	5,332	0.00	0.00	0.00
State C	1,069.1	3.2	2.5	4,692	0.23	0.00	0.00
State D	55.1	156.3	2.5	21,574	0.00	0.01	0.00
State E	500.5	1.1	2.3	3,094	0.16	0.00	0.00
ect.							
					Log of R1D Ratio	Log of R1F Ratio	Log of R1N Ratio
State B					-5.6890	-7.4261	-7.2080
State C					-1.4791	-7.2886	-7.5391
State D					-5.9701	-4.9272	-9.0689
State E					-1.8214	-7.9184	-7.2218
ect.							
Average					-6.4080	-6.0309	-7.9524
Natural exponent (of average Log)					0.0016	0.0024	0.0004
Distribution					0.4743	0.5458	0.0799
State A	RID	RIF	RIN				
Imputed	255.8	44.5	2.5				

A normal probability plot of the logs of the ratios is presented in Figure 5. This figure demonstrates that random discrepancy and normality is significantly improved with the log-transformation model.

Figure 5. NCES III Method, Normal Probability plot of R1D ratio



ADDITIONAL METHODS EXPLORED

The following sections cover Time Series, Regression, and the Nearest Neighbor methods which, after initial exploration, were found not to be suitable candidates for CCD Finance Data Imputations.

Time Series Method

The problem encountered using time series is that there is not enough data to get good diagnostic plots which are critical in determining which model should be fit. At present there are only 4 years of CCD Finance Data were available to fit a model and at least 6 to 8 more years are needed in order to determine what model should be fit.

Regression Method

In employing the regression method, individual regression relationships need to be identified for each variable to be imputed and the auxiliary variable have to be identified. These variables in turn may have to be imputed. In addition, the imputed values for the missing components of an aggregate, provided by separate regressions, would not sum up to the reported value of the aggregate. Though, seemingly, a proportional adjustment can be taken to the imputed values to make their sum matching the aggregate value, the validity of such adjustment is in question.

Nearest Neighbor Method

The Nearest Neighbor method uses the financial data to group States in order to apply separate imputation distributions. Each group of States would have its own imputation distribution. As recognized on the Original ratio plot for R1D for the fiscal year 1992 data, two clusters of points appear. (Figure 1) This pattern displays a classification of the States.

The Nearest Neighbor method incorporates this information of classification into the imputation operation. The reporting States are grouped into two classes, and imputation distribution would be found for each class of States. When imputing for a missing value, the imputing State's class needs to be identified before applying the corresponding imputation distribution.

This model has been rejected in the past because of the difficulty in determining the class of states. In addition, for some survey items, there are only a small number of states which have the specific revenues or expenditures for which we are imputing. Dividing this small number of observations (states) into groups results in too small a grouping upon which an imputation can be based.

ANALYSIS AND RESULTS

The selection of the best imputation method depends on the uses to which the data are to be put. For each of the imputation methods described in the previous sections, groups of variables of importance to NPEFS were evaluated across fiscal years 1989 through 1992. The objective used in the evaluation was to minimize the average percent error across the largest set of variables of interest. Percent error is defined as the absolute difference of the reported value from the imputed value, divided by the reported value.

Three groups of variables are used in the evaluation of three NCES methods using data for fiscal years 1989-1992. Group 1 is a small group of revenue variables which consisted of R1D, R1F, and R1N (which were used in demonstrating the methods). Group 2 is a larger group of revenue variables. Group 3 consists of expenditures for Food Services.

For Group 1 variables, the NCESII method performed best across all years, yielding the smallest average percentage error as highlighted in Table 4.

Table 4. Results of Analysis using group 1 variables

Method	Year	Percent Error RID	Percent Error RIF	Percent Error RIN	Average Percent Error
NCES I	1989	3.69	40.65	3.19	15.84
NCES II	**1989**	**10.35**	**14.06**	**0.74**	**8.38**
NCES III	1989	3.71	40.59	3.10	15.80
NCES I	1990	3.04	32.07	15.05	16.72
NCES II	**1990**	**7.84**	**5.93**	**0.80**	**4.68**
NCES III	1990	3.94	29.12	6.63	13.23
NCES I	1991	3.20	34.18	12.12	16.50
NCES II	**1991**	**5.73**	**6.82**	**0.86**	**4.47**
NCES III	1991	4.29	22.45	6.50	11.08
NCES I	1992	4.90	26.97	10.95	14.27
NCES II	**1992**	**8.99**	**5.44**	**0.88**	**5.10**
NCES III	1992	6.14	20.87	6.56	11.19

Similar analysis was performed for Group 2 and Group 3 variables. The resulting average percent errors from this analysis and from the Group 1 analysis is presented in Table 5.

Table 5. Summary of analysis using groups 1, 2 and 3

		Average Percent Error		
Method	Year	Group 1	Group 2	Group 3
NCES I	1989	15.84	274.49	---
NCES II	1989	**8.38**	170.89	---
NCES III	1989	15.80	**36.56**	---
NCES I	1990	16.72	10.17	---
NCES II	1990	**4.68**	7.16	---
NCES III	1990	13.23	**2.34**	---
NCES I	1991	16.50	15.11	---
NCES II	1991	4.47	10.78	---
NCES III	1991	11.80	**5.65**	---
NCES I	1992	14.27	2.24	2.25
NCES II	1992	**5.10**	2.48	2.27
NCES III	1992	11.19	**1.35**	**1.28**

Conclusion

The NCES III method of imputation appears to be the best method for imputing data for the NPEFS survey. It does better a majority of the time and always does better than the other methods for larger groups of variables. The overall average percent error is the smallest using the NCES III method for the majority of the variable groups considered. The logarithmic transformation works to minimize the amount of variability encountered in the data.

References

National Center for Education Statistics. (1995). *The National Education Finance Survey Booklet.* Washington DC.

National Center for Education Statistics. (1995). *Statistics in Brief: Revenues and Expenditures for Public Elementary and Secondary Education: School Year 1992-93* . Washington DC.

Weisburg, S. (1985). *Applied Linear Regression. 2nd ed.* New York: John Wiley.

DISCUSSION

David L. Hubble, U.S. Bureau of the Census
Demographic Statistical Methods Division, Washington, DC 20233

Prior to reading these papers, I was quite illiterate in the subject of these NCES surveys, their associated components, and the non-sampling error issues addressed in these papers. Now, unfortunately, I dream of their issues and problems and how they might find solutions. I commend (or is it condemn) the authors for doing such a fine job of drawing me into their world.

As with a child, please forgive the silly questions I may ask, as I feel my way through this new maze, but with the same gentle manner please straighten me out if I stray off course and am missing the point.

I will discuss each paper in the order in which they were presented.

"Comparisons Across and Within NCES Surveys" by Salvucci, et al.

Unfortunately, Salvucci is dealing with one of the messiest problems in statistics. That is, collect the same information by two different methods and then try to reconcile the differences that arise in your basic count statistics.

I know of only one clean solution to this kind of problem and that is to ignore one of the data sources. Foregoing this rather simplistic solution, Salvucci attempts to identify the various differences between the Core of Common Data (CCD) surveys and the Schools and Staffing Surveys (SASS). If I understand it correctly, the CCD surveys are actually censuses and serve as the sample frame for the SASS.

After establishing the existence of many significant differences in the count of students, teachers, schools and local education agencies between the CCD and SASS data, Salvucci provides a good discussion enumerating the coverage differences between the CCD and SASS surveys that may explain the observed differences in the basic count statistics.

However, while the reconciliation process explained away the differences for many states, several states still had significant differences. And in fact, some states showed significant differences only after adjustments were made.

Personally, I felt like many of the differences were still unexplained.

Some questions come to mind:

(1) Do the same people respond to the CCD and the SASS? I wasn't sure.

(2) Are the CCD and SASS data collected at the same time of the year?

In the final section and in the conclusion, an attempt is made to assess who's right and who's wrong between the CCD and SASS. SASS was assessed to be the major source of the differences.

While there may be merit in such work, I have some reservations with the method. The method compared the SASS to CCD percentage difference with CCD year-to-year percentage differences. If the SASS to CCD percentages was larger than the CCD year-to-year percentages then the discrepancy was attributed to the SASS estimate.

However, my guess is that the reporting error for CCD estimates by state are highly correlated from year to year, while the correlations between SASS and CCD reporting errors are not nearly as high.

The problem scenario I envision is this: the same CCD reporting error is made year after year while the SASS figure is virtually error free. However, under the decision criteria stated earlier, SASS would be labeled as the source of the discrepancy because the SASS to CCD difference would almost always be greater than the CCD year-to-year differences.

Whether SASS is the major source of the differences, or not, I don't know, but I do believe this assessment technique needs to be rethought. Possibly something along these lines could be done once SASS 93-94 data is available.

But beyond this issue of blame is a larger question. Does NCES want these differences in estimates to disappear or at least be substantially reduced in the future? If so, then plans for modifying the CCD or SASS to bring them in line need to be developed. And while being able to accomplish this completely is extremely difficult, if not impossible, at least those differences that are little more than arbitrary should be removed.

Also, I wonder if some of the issues presented and discussed at an earlier session apply to possibly bringing these CCD and SASS estimates into agreement.

"Documentation of Nonresponse Across NCES Surveys" by Saba, et al.

Saba has tackled the rather daunting task of reviewing the detailed documentation of 13 different NCES surveys.

Saba has amassed a great deal of information on how each of these surveys address 4 issues:

- The calculation of unit response rates.
- The calculation of item response rates.

- Methodological and analytical issues of addressing nonresponse in estimation, and
- Categorization of demographic variables.

Saba's work was quite comprehensive. The only possible omission was that the discussion of methods on how to deal with nonresponse only included unit nonresponse. There was no mention of how these surveys dealt with item nonresponse. Maybe the documentation was lacking, but still I was left wondering just the same.

However, even if this had been covered, my overall feeling after reading this paper would have been the same. And that is...

WHERE DO WE GO FROM HERE?

Saba's work clearly demonstrates that the methods are very different from survey to survey.

BUT WHAT HAPPENS NOW?

Certainly you want to avoid this document living a quiet life (or death) in a dozen or so filing cabinets. I have two thoughts here:

(1) Create a database with all this information. The paper stated that this information would be useful for users of the NCES data and especially to those making comparisons across surveys. Possibly this database could be updated and, therefore, not force others to read through tons of material in search of a few pieces of information.

(2) A second thought is to establish even greater uniformity between surveys in the future (a goal Saba pointed out in the paper). In terms of nonresponse, it may be helpful to establish guidelines for defining and classifying nonresponse for different types of surveys (such as mail, RDD, or personal interview). This would probably also lead to greater uniformity in methods used to adjust for nonresponse.

In terms of categorization of demographic variables it may be helpful to develop forms (or at least specific question and answer categories) that are used across several surveys. An example of this is the Census Bureau's attempt to create what is generally referred to as a "Uniform Control Card" for all its major household-based demographic surveys (CPS, NCVS, SIPP, CE). In doing so analysts can be fairly sure when comparing statistics on, let's say, race that the question was worded the same way and that the answer categories were presented in the same order.

I believe doing this would increase the utility of all the NCES surveys.

"Multivariate Modeling of Unit Nonresponse for SASS 1990-1991" by Gruber, et al.

Gruber's paper is another good example of using modeling techniques to gain a more complete understanding of nonresponse in surveys and, in this case, unit nonresponse. The surveys included SASS's Public School Surveys and Private School Surveys.

Assessments of the available variables of state, urbanicity, school level, and school size for public school surveys and the variables of urbanicity, school level, school size, and affiliation for private school surveys have been made before through the modeling work of Shen, Parmer, and Tan in 1992.

The new twist in the paper is the combining of final models across the surveys that is school district, school and teacher surveys. This was done separately for public and private schools.

These combined models were referred to as "Cross Component Models". While the idea is intriguing, I believe the paper needs to provide more information in terms of the motivation for combining the models. It wasn't clear to me.

The conclusion from the paper was that fewer variables appeared to be significant in modeling unit nonresponse than from the Shen, Parmer, and Tan modeling research.

I wondered why this outcome. Does it relate to the fact that the "unit" is different in the various surveys with the cross component models? That is, the unit ranges from school administrator to school to teacher.

Or is it because states were collapsed into strata in this paper, while no state collapsing took place in the Shen, Parmer, and Tan research?

One final side note: while the collapsing of states may improve the model, I get a little nervous when one of the defined state strata consists only of Alaska and Massachusetts. Maybe regional constraints are needed.

"Evaluation of Imputation Methods for the CCD Finance Data" by Johnson, et al.

Johnson's paper deals with the problem that for certain reported aggregate amounts the corresponding components are not reported by many states. This was addressed both in terms of revenues and expenditures.

Specifically, the paper evaluated different approaches to impute the missing internal values. Before discussing the specific techniques used for imputation, I would like to raise a few issues.

An example in the paper describes a situation in which the components require imputation because one of the components is missing or zero. Having worked on a similar problem, I know we grappled with the proper treatment of reported zeros. I wonder did you always impute for a reported zero? Or were states ever contacted about missing values or reported zeros for clarification?

Also, out of curiosity, I wondered if states ever report values for some components that do not add to the reported total value? And, if so, how are these situations handled?

In the year investigated (91-92) only 14 states reported the desired detail of the revenue variable examined. One concern I have is; are the 14 states representative of the 36 who did not report the detailed information. Generally, it is problematic if correlations exist between whether a state reports the detailed distribution and the distribution itself. And when response rates are low the likelihood of this occurring increases. Specifically, of the 14 states there appears to be some geographic clustering.

This is less of a concern for the expenditure variable which had 38 states report the components.

As I was reading the paper I had another "just wondering" question. Is longitudinal imputation a possibility? Even for just some of the states needing imputation? If previous years distributions were available it would make sense to use them. I get the impression though, that the details of the reported finances are fairly constant from year to year.

In terms of the possible techniques evaluated in the paper, I also have a few comments.

One is that, the ability to interpret the plots is increased if the X- axis is always in terms of the sum that was used to compute the ratio on the Y-axis. This was not always done.

Of all the specific techniques compared, I agree that the NCES III Alternate Herriot is probably the best. Also, the number of states evaluated was quite small, 6 or 7 states at times.

However, one potential drawback to all these techniques is that the distributions for all the imputed states will be exactly the same and probably looking different than any of the reporting states. This is a problem if one needs to estimate variances associated with these distributions. Though, I am not sure if such needs arise.

Finally, I was interested in knowing more about why the regression method with its promising "highly significant linear relationship" failed in warranting further consideration.

CONCLUSION

As I said in the beginning, I have learned a lot about the NCES surveys and some of their current issues. The authors are commended in their fine work. My hope is that you may find some of my comments to be of some use. And finally, I appreciate the opportunity to have participated in this session.

VARIANCE ESTIMATES COMPARISON BY STATISTICAL SOFTWARE

Stanley S. Weng and Fan Zhang, Synectics for Management Decisions[1]
Michael P. Cohen, National Center for Education Statistics[1]
Stanley S. Weng, Synectics for Management Decisions, Inc., Arlington, VA 22201

Key Words: Complex survey, Variance estimation, Balanced repeated replication, Jackknife, Taylor linearization

This article reports a comparison analysis which, involving six statistical software routines in wide use for variance estimation for complex surveys, examined the variance estimates produced by those routines in a sample data setting from an NCES (National Center for Education Statistics) complex survey. This study helps identify reliable and capable statistical software for variance estimation, and, perhaps more meaningfully, is an effort to raise the standard of practice in the analysis of complex survey data.

1. Introduction

Most of the surveys of the NCES are large complex surveys. As well known, the sampling and weighting processes of complex surveys have much changed the methodology and algorithms of variance estimation.

Conventional statistical software packages such as SAS and SPSS can be only used to provide variance estimates for simple random samples. Naive use of such software for variance estimation on complex survey data, as often made in practice, may lead to underestimating the variances.

There are three methods widely used for variance estimation for complex surveys: the *balanced repeated replication* (BRR) method, the *jackknife* (JK) method, and the *Taylor series* method (Wolter, 1985). The first two methods are under the *replication* approach, and the third one under the *linearization* approach. A number of statistical software have been developed to perform these methods.

The BRR method has been used to estimate the sampling errors associated with estimates for all of the 1990-91 Schools and Staffing Survey (SASS) samples. In the BRR method, within each stratum, sampled schools are paired by the order they were selected. One school from each pair is placed into each replicate. Each replicate includes approximately half the total sample. The choice of when to place a school from a pair into a replicate is done in a balanced manner to reduce the variability of the variance estimates. See Kaufman and Huang (1993) for more detailed information on how SASS units are placed into balanced half-sample replicates. SASS uses 48 replicates for variance estimation, giving a reasonable degrees-of-freedom cushion for the validity of the z-test approximation when making inference. Each SASS public use file includes a set of 48 weighted replicates for BRR variance estimation.

The jackknife method has been used by the 1990 National Assessment of Educational Progress (NAEP) to estimate all sampling errors as presented in the various reports and provided good quality estimates of sampling variance for most statistics. A set of 56 jackknife replicate weights for students was developed, for the purpose, in the manner that models the design as one in which two PSUs were drawn with replacement per stratum (Johnson and Allen, 1992).

The Taylor series method has been used by the National Education Longitudinal Study of 1988 (NELS 88) and follow-ups to calculate standard errors as presented in various reports (Spencer et al., 1990, and Ingels et al., 1994).

However, NCES recently reported several occurrences where unexpected differences in variance estimates were produced by different statistical software routines. This resulted in a concern: if reliable results could be expected from available variance estimation routines, including their estimating approach, portability, and capabilities to accommodate the features of various complex surveys. This study was conducted to address the computational as well as methodological issue: whether different statistical programs, using different estimation methods and with different designs, produce significantly different results for complex surveys. The study compared the variance estimates, produced by six statistical software routines in wide use, from descriptive as well as regression analysis using the same data from an NCES complex survey.

We will present the analysis and results in Section 2, and make some discussions in Section 3 to serve the purpose of this study. The remainder of this section is a brief description of the six software routines selected for this study. They are:

SUDAAN (Shah, et al., 1992). Uses Taylor series approximations in conjunction with textbook-type variance formulas to calculate variance estimates.

PC CARP (Fuller, et al., 1986). Uses Taylor series method. It uses a general framework of linearization for the calculation of variance estimates, which could cover most sampling designs used in practice.

VPLX (Fay, 1995). Performs replication methods (BRR, JK, etc.). VPLX can create the jackknife replicate weights in general algorithm, and has the full

computational ability to handle hundreds of PSUs within a stratum without the need of grouping.

WESVAR (Westat, 1993a) and WESREG (Westat, 1993b). WESVAR handles basic survey estimates. WESREG handles regression analysis. Both programs perform either BRR or JK. The jackknife procedure of WESVAR and WESREG assumes a two-per-stratum sampling design.

REPTAB(Liebman, 1993). A SAS procedure, uses replication methods (BRR and JK).

STRATTAB (Ogden and Liebman, 1991). A SAS procedure using a Taylor series approximation.

2. Analysis and Results

2.1 Data

Data from the Teacher Survey of the 1990-91 School and Staffing Survey (SASS), as recommended by NCES statisticians, were used to apply the software routines for the variance estimates comparison. Below is a brief description of the SASS Teacher Survey.

The SASS Teacher Survey is a two-stage stratified probability sample. The school survey is the first stage of the sampling. Schools are selected within strata by a probability proportional to the number of teachers within the school. Within the first-stage school sample, a second-stage teacher sample is selected stratified by teacher experience level.

The SASS Teacher Survey sample design is very close to the standard two-stage sampling design, as the one adopted in the design by all statistical software for variance estimation for complex surveys: the stratified probability sampling with replacement at the first stage and simple random sampling at the second stage. For the analysis purpose of this study, because of the small sampling rates of schools within strata, it should not cause concern to treat the sample as from with-replacement sampling at the first stage. Stratum and PSU variables, as required for performing Taylor series and jackknife procedures, are well included in the data files, and the BRR replicate weights for teachers are also available in the files.

2.2 Analyses

Our analyses used the public school sample in the Teacher Survey. Variance estimates were produced for basic survey statistics, including means, percents/ proportions, and ratios, as well as regression coefficients, using the six selected software routines. Two analyses with different sets of variables were conducted for each kind of statistics (see Table 1, the first three columns).

Here is a list of some questions with abbreviated wording in the column Variables of Table 1:

Percent:

Master's degree--Do you have a master's degree?

Look forward to day--I usually look forward to each working day at this school.

Mean:

Salary--What is your academic base year salary for teaching in this school?

Ratio:

Schl hrs extra--School-related activities involving student interaction

Othr hrs extra--Other school-related activities

Regression (first):

Independent--Have you ever taken any undergraduate or graduate courses in the following subjects.

Before entering analysis, the data were necessarily shaped. For instance, the strata which contained only one PSU were appropriately collapsed. Missing values were also handled appropriately according to the design of the software routine applied. For those routines which do not have the capability of handling missing values, missing variables were handled in an external data step.

There are two versions of the jackknife procedure used in variance estimation for survey data: the *simple jackknife* (JK1) and the *stratified jackknife* (JK2). The simple jackknife is the basic algorithm of the jackknife procedure. The stratified jackknife is a generalization of the simple jackknife to stratified samples. For multi-stage stratified sample variance estimation, the simple jackknife is considered generally not able or not suitable to perform. Only the stratified jackknife was performed in this study.

The Taylor series procedure, as understood, is performed in conjunction with the selected sampling design. For SUDAAN, a number of standard designs as options are available. By the sampling design of SASS, the appropriate design option would be "without replacement (WOR)". However, under this design option, SUDAAN requires data on the number of PSUs for each stratum in the population. The variable, NUMSCH, in 1990-91 SASS public school file for this purpose, had some problems in its data. For instance, all PSUs in the same stratum should have the same values of NUMSCH, but this is not always the case. Therefore, our analysis used the design option "with replacement (WR)" which appeared suitable to the survey design and the data. In using PC CARP, the sampling design is identified by three components: the design variables entered into the analysis, the "Two-Stage" option, and the optional data of the sampling rates of strata. Since there were no sampling rates data available, our analysis also used the "with replacement" sampling design for PC CARP. STRATTAB was designed only for the standard sampling scheme assuming sampling with replacement at the first stage. Thus, the same design option was used for the three routines to perform the Taylor series procedure.

Some features of the software were noticed with the

conduction of the analyses.

(a) For WESVAR, if a given variable has a missing value for an observation, that observation is not used in the calculation of the total for that variable only. Effectively, this treats the missing value as zero in all computations. However, even to estimate the mean for a missing variable, this way of handling missing values will yield incorrect results. In fact, the WESVAR mean is treated as a ratio. Thus, the numerator will be calculated using only non-missing values, while the denominator will sum up to the weights for all observations. The same problem will occur when calculating ratio estimates. When the two variables involved have missing values in different cases, the resulting ratio estimate will be misleading. To avoid the problem, we handled the missing data in a SAS data step, prior to using WESVAR.

WESREG was supposed, as a regression procedure, to handle missing values in the usual way, as also stated in its manual: "Observations having missing values for the dependent variable or any of the independent variables are excluded from all estimates." However, our analysis showed that it is not the case with WESREG. There is no further information available to clarify how WESREG handles missing values. In our analysis, we then handled missing values in a SAS data step prior to using WESREG.

(b) The version of WESVAR used in our study does not have the ability to perform jackknifing from the stratum and PSU variables in the data. Moreover, the jackknife procedure in WESVAR is in a simplified form. As documented in the manual (Westat, 1993), the jackknife procedure is formulated only to the special case that there are two PSUs in each stratum. WESVAR cannot handle more than two PSUs in a stratum. When there are more than two PSUs in a stratum, even if the jackknife replicate weights are supplied, a simple use of WESVAR will give wrong results. In such a situation a grouping procedure must be conducted to make two (pseudo) PSUs in each stratum in order to meet WESVAR's jackknife frame.

Remark The new WesVarPC (Westat, 1995) can create the jackknife replicate weights from the design variables, however, still in the two-per-stratum form, remaining from the design of WESVAR.

2.3 Variance estimates

The estimates of the statistics and associated standard errors are presented in Table 1. The different variance estimation procedures were not involved in the estimation of the survey statistics. All the software routines produced identical estimates for all the statistics in the analysis. In Table 1, one column is used to present the estimates of the statistics, and the body of the table is for the variance estimates (standard errors) presented by the estimation method and the software used. For the analysis not available due to capability limitation of the software, an N/A indicator is put in the table. In the following, we examine the variance estimation results, mainly under same estimation method and also across the methods.

(1) BRR variance estimates

The three statistical software packages, VPLX, WESVAR, and REPTAB, provide BRR variance estimates for descriptive survey statistics. As generally designed for software performing BRR, replicate weights need to be supplied with the input data for all the three programs. With replicate weights supplied, only simple and standard calculations need to be conducted to obtain the BRR variance estimates. As Table 1 shows, for all the descriptive statistics (means, percents, and ratios), the three routines produced identical BRR variance estimates.

Among the six software packages, WESREG is the only one providing BRR variance estimates for regression coefficients. No comparison could be made for the BRR variance estimates for regression coefficients. However, some comparisons between the results by WESREG and by SUDAAN and PC CARP (both using Taylor series method) may be of interest, and are discussed later in this section.

(2) Jackknife variance estimates

The data set does not supply replicate weights for the jackknife procedure and thus WESVAR and REPTAB are not applicable. VPLX is the only software which provided jackknife variance estimates in this study. By using all PSUs in the jackknifing, VPLX reached great precision. The VPLX jackknife variance estimates appear the same, except for a minor difference for the mean of SALARY, as those produced by SUDAAN and PC CARP using the Taylor series method. This coincidence, as expected from the asymptotic property that the jackknife variance estimate tends to be close to the linearized variance estimate if both calculations employ the same PSUs and the statistic is smooth (Wolter, 1985), is an indication that VPLX has sound statistical design and is computationally reliable.

(3) Taylor series variance estimates

Three statistical software routines, SUDAAN, PC CARP, and STRATTAB, produced variance estimates using the Taylor series method.

Experience with large, complex sample surveys has shown that the Taylor linearization approximation often yields satisfactory results, except for highly skewed populations. Generally speaking, if the nonlinear estimator can be expressed as a smooth continuous function of population totals, the Taylor linear approximation would be valid (Wolter, 1985).

SUDAAN and PC CARP produced identical variance estimates for the descriptive survey statistics, means, percents, and ratios. For the first regression analysis, the variance estimates for the coefficients produced by the two programs appear slightly different. The differences may be due to different computational procedures handling complex functions of survey estimates. As for the second regression analysis which is simpler than the first one, the variance estimates by the two programs are similar.

The variance estimates produced by STRATTAB, only available for means and percents, appear to be of quite different magnitude (smaller) compared to those by SUDAAN and PC CARP. The differences could not be considered as within a reasonable range of error due to different computational procedures.

(4) Comparison across estimation methods

Though there seems no general comparison based on rigorous theoretical justification between the BRR, jackknife, and Taylor series methods for variance estimation - appraisal of their performance with different estimators and types of surveys has relied on empirical studies, an important property has been established that for nonlinear statistics that can be expressed as functions of estimated means of p variables - such as ratios, regression and correlation coefficients, the variance estimators from the linearization, the jackknife, and the BRR methods are asymptotically consistent (Krewski and Rao, 1981). This result is valid for any multistage design in which the primary sampling units (PSUs) are selected with replacement and in which independent subsamples are selected within those PSUs sampled more than once (Rao and Wu, 1988). The sample data used in our study can be considered under this situation, and are of large size. Meaningful information could be drawn.

For the descriptive survey statistics, the BRR variance estimates are very close to that produced by the Taylor series (using SUDAAN and PC CARP) and jackknife methods. And for most of them, the BRR variance estimate appears slightly lower.

For the first regression analysis, for six out of the eight regression coefficients, the BRR standard error (by WESREG) appears significantly different from, mostly higher than, the Taylor series estimate (by SUDAAN and PC CARP). The differences range from 14 percent to over 50 percent compared to the Taylor estimates. For the second regression, the BRR standard errors appear almost the same as the Taylor estimates. Since the first regression involves more regressors than the second one, the behavior of BRR method when performed to complex functions of survey estimates, such as regression coefficients, needs to be further explored. The comparison of jackknife estimates and Taylor series estimates was already made above between the results from VPLX and from SUDAAN and PC CARP.

3. Conclusion

For estimating variances, it would be expected that statistical software routines performing the same estimation method produce same results; while for routines using different methods it may not be expected that same results would be reached. Thus, identical variance estimates produced by two statistical routines performing different methods provide an indication of their reliable performance; while significant difference in the results produced by routines using the same method implies error existent with some of them.

This study thus helps identify reliable and capable statistical software for variance estimation for complex surveys. Reliable statistical software routines are available to perform all the three variance estimation methods.

This study may also be a motivation for further development of comprehensive statistical software for variance estimation of survey data.

To perform the BRR, the creation of the BRR replicate weights is an issue. All the statistical software routines for performing BRR, included in this study, require the replicate weights be supplied in the data input. This situation certainly limits the practice of calculating the BRR estimates. As already made available for general jackknifing, VPLX is going to make available a general algorithm for creating BRR replicate weights within the program. The new WesVarPC (Westat, 1995) has the capability of creating the BRR replicate weights. Such capability will expand the usability of the software and promote the use of the BRR method.

The progress of computing ability in recent years has been changing the consideration in designing statistical software for variance estimation for complex surveys. Computing cost seems no longer a big issue as it was years ago. The computing-intensive methods, like BRR and jackknife, can be performed in general versions as a usual matter. Unnecessary simplification in the estimation algorithm would merely limit the applicability of the software and reduce the power of the performance of the method.

Many NCES surveys use more complex sampling designs than the standard one as assumed for the BRR and the jackknife to apply. It seems necessary to make available the statistical software using more general algorithms for variance estimation, for example, the more general resampling procedure (Rao and Wu, 1988; and Kaufman, 1993a, 1993b, and 1995), and also the combination of linearization and replication methods, if the Taylor linearization does not bring the estimate to an appropriate form to which standard variance estimation formulas are applicable.

Table 1: Standard Errors Associated with Survey Estimates by Statistical Software

Analysis				Variance Estimation Method						
				BRR			JK2	Taylor series		
Data type	Survey statistics	Variables	Esti-mate	VPLX	WESVAR/ WESREG	REPTAB	VPLX	SUDAAN	PC CARP	STRAT .TAB
Categorical	Percent(%)	Master's Degree								
		1: YES	46.980	.326	.326	.326	.393	.393	.393	.0259
		2: NO	53.020	.326	.326	.326	.393	.393	.393	.0259
		Look forward to day								
		1: ST AGREE	51.37	.341	.341	.341	.385	.385	.385	.019
		2: AGREE	40.39	.313	.313	.313	.363	.363	.363	.017
		3: DISAGREE	6.23	.163	.163	.163	.180	.180	.180	.014
		4: ST DISAGREE	2.01	.121	.121	.121	.107	.107	.107	.000
Continuous	Mean	Salary ($)	30,751	93.494	93.494	93.494	102.849	102.798	102.798	7.099
		AGE (=91-BIRTHYR)	42.576	.0751	.0751	.0751	.0732	.0732	.0732	.0028
	Ratio	Schl hrs extra/Hrs requ	.0886	.001	.001	.0011	.001	.0010	.0010	N/A
		Othr hrs extra/Hrs requ	.223	.0013	.0013	.0013	.0014	.0014	.0014	N/A
	Regression Coefficients	Independent:		N/A		N/A	N/A			N/A
		Math	72.152		155.231			188.72	194.46	
		Computer Science	232.656		397.865			258.61	258.10	
		Biology-Life Science	221.769		170.345			126.36	128.73	
		Chemistry	-27.725		379.888			355.94	359.63	
		Physics	323.148		369.148			323.39	319.11	
		Earth/Space science	339.624		345.140			309.12	310.74	
		Other nat science	435.344		369.957			264.33	266.24	
		AGE	451.914		44.107			48.32	44.26	
		Dependent: Salary								
		Independent:		N/A		N/A	N/A			N/A
		Look forward to day	-1.274		.0716			.0726	.073	
		BIRTHYR	-0.745		.0054			.0056	.006	
		Depen.: Years to retire								

N/A: Not available due to capability limitation of the software

NCES recently issued a note from the chief statistician regarding the technical approaches to performing analyses on NCES survey data (Ahmed, 1993) in the desire to perform more complex statistical analyses on NCES data taking account of the complex survey designs. In practice it is not unusual that analysis on complex survey data does not account for the complex design. As reported by a recent survey by the Census Bureau, for instance, many, if not most, journal articles in the social sciences do not account for the complex survey. More effort needs to be made to promote the survey data analysis practice, including the further development and employment of advanced statistical software for variance estimation and other analysis purposes. With today's computing ability and facilities, it is necessary and possible to raise, with our great effort, the standard of practice in the analysis of complex surveys.

[1] This paper reports the general results of research undertaken by staff members of Synectics for Management Decisions, Inc. and the National Center for Education Statistics (NCES). The views expressed are attributable to the authors and do not necessarily reflect those of Synectics or NCES.

References

Ahmed, S. W. (1993), "Technical Approaches to Performing Regression and Other Multivariate Techniques on NCES Survey Data - Where We Stand," A Note from the Chief Statistician. Washington, DC: National Center for Education Statistics.

Fay, R.E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," *Proceedings of the Survey Research Methods Section*, 212-217. Alexandria, VA: American Statistical Association,

Fay, R.E., (1995), *VPLX*. Washington DC: U.S. Bureau of the Census.

Fuller, W.A., Kennedy, W., Schnell, D., Sullivan, G., and Park, H.J. (1986), *PC CARP*. Ames, IA: Statistical Laboratory, Iowa State University.

Ingels, S.J., Scott, L.A., Rock, D.A., Pollack, M.J., Rasinski, K.A., and Wu, S.-C. (1994), *National Education Longitudinal Study of 1988, First Follow-up Final Technical Report, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Johnson, E.G. and Allen, N. (1992), *The NAEP 1990, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Kaufman, S. (1993a), "A Bootstrap Variance Estimator for the Schools and Staffing Survey," *ASA 1993 Proceedings of the Section on Survey Research Methods.*

Kaufman, S. (1993b), "Properties of the Schools and Staffing Survey's Bootstrap Variance," *ASA 1994 Proceedings of the Section on Survey Research Methods.*

Kaufman, S. and Huang, H. (1993), *1991 Schools and Staffing Survey: Sample Design and Estimation, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Kaufman, S. (1995), "Properties of the School and Staffing Survey's Bootstrap Variance Estimator," presented at the 1995 ASA Meeting.

Kish, L. and Frankel, M. (1974), "Inference from Complex Samples," *Journal of the Royal Statistical Society: Series B* (Methodological), 36: 2-37.

Krewski, D. and Rao, J.N.K. (1981), "Inference from Stratified Samples: Properties of Linearization, Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics*, 9, 1010-1019.

Liebman, E. (1993), *PC REPTAB* (with *PROC REPTAB*). Berkeley, CA: MPR Associates, Inc.

Ogden, C. and Liebman, E. (1991), *PC STRATTAB* (with *PROC STRATTAB*). Berkeley, CA: MPR Associates, Inc.

Rao, J.N.K. and Wu, C.F.J. (1988), "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association*, 83, 231-241.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, Vol 1, No. 4, Statistics, Sweden.

Sarndal, C.E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shah, B.V., Barnwelll, B.G., Hunt, P., and LaVange, S.C. (1992), *SUDAAN User's Guide* (software version 6.00, 1992). Research Triangle Park, NC: Research Triangle Institute.

Spencer, B.D., Frankel, M.R., Ingels, S.J., Tourangeau, R., and Owings, J.A. (1990), *National Education Longitudinal Study of 1988, Base Year Sample Design Report, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Westat,Inc. (1993a), *The WESVAR SAS Procedure, Version 1.2*. Rockville, MD.

Westat, Inc. (1993b), *The WESREG SAS Procedure*. Rockville, MD.

Westat, Inc. (1995), *A User's Guide for WesVarPC, Version 1.1*. Rockville, MD.

Wolter, K.M. (1985), *Introduction to Variance Estimation*. New York: Springer-Verlag.

TEACHER SUPPLY AND DEMAND IN THE U.S.

Richard M. Ingersoll, University of Georgia
Department of Sociology, Baldwin Hall, UGA, Athens, GA 30602

Key Words: teacher supply, teacher shortage, SASS

Introduction

Beginning in the early 1980s, a series of highly publicized reports focused national attention on the imminent possibility of widespread shortages of elementary and secondary school teachers in the U.S. (e.g. Darling-Hammond 1984; National Commission on Excellence in Education 1983). These predictions came as a complete surprise to many. Throughout much of the 1970s, there had appeared to be a surplus of school teachers. Indeed, reductions in the teaching force through layoffs had been common to many schools and districts in the U.S. But, this new research on teacher supply and demand made a compelling case that beginning in the 1980s teacher supply would drastically decrease, while demand for new teachers would steadily increase, resulting in shortages.

Those predicting shortages held that fewer and less qualified college graduates were choosing to teach, while more children of the "baby boom" generation were entering the school system, driving enrollments and, hence, hiring of teachers up. Moreover, a growing imbalance between teacher supply and demand would be exacerbated, according to this view, because of problems of teacher retention. A high level of teacher attrition, in this view, was a large source of demand for new teachers and a key factor behind the predicted shortages (e.g. Haggstrom et al. 1988; Grissmer and Kirby 1987).

These reports arrived in a context of widespread concern and criticism surrounding the adequacy of the elementary and secondary school system as a whole. Critics linked declining U.S. economic performance, especially in the international arena, to declining school performance (e.g. National Commission on Excellence in Education 1983). The apparent inability of schools to attract and retain qualified teachers appeared to be one more in a host of symptoms of the "crisis" besetting schools. As a result, the imminent possibility of teacher shortages gained widespread coverage in the national media.

The education research community was, however, not unanimous in its assessment of the threat of teacher shortages. Some analysts argued that teacher supply was and would continue to be adequate and that attrition was not particularly high (e.g. Feistritzer 1986). A study conducted of Indiana in the late 1980s seemed to provide empirical support for these arguments. It suggested that due to higher salaries and increased re-entry of former teachers, teacher supply had increased, and that due to a stable work force and a decline in turnover among new teachers and women, attrition was actually at its lowest point in years (Grissmer and Kirby 1992).

Currently, research and policy concerning teacher supply and demand seems to be in a state of limbo. Little research has been done to resolve the above contradictory claims. Indeed compared to the 1980s, interest in teacher shortages in the research community, the policy community and in the media seems to have largely disappeared. As a result, it is not at all clear what happened to the teacher shortage.

Almost all involved have agreed that one source of the confusion and irresolution, has been a lack of data, especially at the national level, on the disputed phenomena: the demand for teachers, the supply of teachers and the gap between the two (e.g. Haggstrom et al. 1988). Indeed, it was in order to address these shortcomings, that the National Center for Education Statistics (NCES), the statistical agency of the U.S. Department of Education, fielded a major new survey of schools and teachers in the late 1980s - the Schools and Staffing Survey (SASS).

This paper presents data from SASS that directly address the debate as to whether there are or are not shortages of teachers in elementary and secondary schools in the U.S. in recent years Our analysis examines what has happened to demand for new teachers, and whether the supply of teachers has been adequate to meet this demand. It examines to what extent schools have difficulty meeting their needs for new teachers, and how they cope with the difficulties they do have. This paper is drawn from a larger ongoing investigation of teacher supply, quality and demand in the U.S. sponsored by NCES. The results presented here build on two previously published documents reporting results from this larger investigation (see Ingersoll 1994 and 1995a).

Data and Methods

The Schools and Staffing Survey is the largest and most comprehensive data source available on the staffing, occupational and organizational aspects of schools in the U.S. It includes a wide range of information on the characteristics, work, and attitudes of school faculty, and on the characteristics and conditions of schools and districts. SASS was designed to be administered triennially; at this point three cycles are publicly available - for the 1987-88, 1990-91 and 1993-94 school years.[1] This analysis used data from the first two cycles.

SASS includes four sets of integrated questionnaires: a school survey; a central district office survey for public schools; a principal survey; and a teacher survey. Response rates have been high, ranging from about

84 percent for private school teachers to 95 percent for public school administrators. The samples utilized in this analysis contain about 4,800 public school districts, 9,000 public schools, 2,600 private schools, 46,700 public school teachers, and 6,600 private school teachers. All of the data reported here are weighted to be representative of the national populations of teachers, principals and schools in the year of the survey.

Each cycle of SASS obtained a rich array of information on issues at the heart of the shortage debate: the numbers and fields of teaching position vacancies in schools; the degree to which schools experienced difficulties in filling vacancies; the numbers of unfilled positions; the methods that schools used to respond to difficulties in filling vacancies; the sources of new teachers; and the background, characteristics, qualifications and assignments of newly hired and already employed teachers. Theliterature on teacher supply and demand has held that shortages and staffing problems vary greatly depending upon the type of teacher, school and locality. Typically, analysts have argued that particular fields, such as math, science and special education, and particular kinds of schools, such as those serving poor communities, have borne the brunt of teacher supply and staffing problems in the U.S. (e.g. Darling-Hammond 1984). Following the literature, this analysis will focus on similar comparisons. Our analysis will examine a series of indicators related to teacher supply and demand across different subject fields (math, science, English, social studies, special education, English as a second language, etc.), across school sector (public and private) and, within the public sector, across school poverty level. The poverty level of schools is based on the percentage students enrolled that receive the federal reduced or free lunch program (Low poverty: less than 15%; Medium poverty: 15% to 50%; High poverty: 50% or more).

In order to provide additional context, the analysis also utilizes selected historical and other data from several other large-scale surveys: class size data from the National Education Association's Status of the American Public School Teacher survey; salary and supply data from NCES' Recent College Graduates Survey (RCG); data on student enrollment, teachers employed, and pupil-teacher ratios from NCES' Common Core of Data survey (CCD). These data sources will be noted where discussed in the text.

Results

Shortages of teachers, most simply put, occur where demand, or the number of teaching positions funded, outstrips supply, or the number of teachers available. Analyses of shortages then must begin by assessing demand and supply.

What has happened to the quantity of demand for new teachers?

Demand for teachers appears to be on the rise. Since the mid 1980s, after a decade and a half of decline, school enrollments have steadily increased and are projected to continue to do so (CCD). Total public school enrollment, for example, rose about 5 percent from 1984 to 1990. As a result, schools are hiring teachers. At the beginning of both the 1987-88 and 1990-91 school years, an overwhelming majority of schools had job openings for teachers. Moreover, this hiring was not simply done to replace teachers who moved or retired. The number of employed elementary and secondary teachers has steadily increased since the mid 1980s (CCD). For example, from 1987-88 to 1990-91, the total population of elementary and secondary teachers jumped from 2,630,000 to 2,915,000.

Has the quantity of teacher supply been adequate?

Unlike demand trends, changes in the adequacy of teacher supply are far more difficult to assess. As a result, they have proven to be the focus of the bulk of research on teacher shortages and, hence, will be the focus of this paper.

Much of the research on teacher supply has focused on the teacher reserve pool - the quantity of potential teachers. But, the reserve pool of potential teachers is large, diverse and probably, unknowable. Newly qualified teachers who have recently graduated from state-approved teacher training programs at colleges and universities are perhaps the most obvious and quantifiable source of supply. But, newly qualified teachers comprised only about 20 percent of those hired in 1987-88 and 1990-91.

There are numerous other sources of teachers for teaching jobs. Substantial numbers of newly hired teachers in both 1987-88 and 1990-91 were re-entrants - former teachers who were returning. There were also substantial numbers of delayed entrants - trained teachers who did not seek a position immediately after their schooling. Indeed, as many as 40 percent of newly trained and qualified teachers do not seek teaching positions immediately after their schooling (RCG). Some delay their entrance into teaching and some never teach. All of these newly qualified teachers are potential members of the reserve pool.

The real issue for assessing the adequacy of teacher supply is, however, not the number of potential teachers, but how many trained candidates are available and willing to apply to teaching vacancies. One manner of assessing this "actual" teacher supply is to determine how often schools had hiring problems.

The data suggest that despite the large reserve pool, many schools do, indeed, not find it easy finding qualified candidates to fill openings. For instance, in 1987-88, principals in 40 percent of all public and 47 percent of all private schools reported experiencing some difficulties in finding qualified applicants to fill their teaching vacancies in at least one field. The situation was comparable in 1990-91,

as illustrated in table 1.

Table 1 —Percentage of secondary schools reporting difficulties filling teaching vacancies with qualified teachers, by field and school characteristics: 1990-91

	English	Math	Life Sciences	Physical Sciences	ESOL or Bilingual	Special Education
Public	11	18	11	15	6	21
Poverty Enrollment						
Low	9	17	10	14	5	19
Med	13	18	11	16	6	22
High	14	21	15	14	7	26
Private	8	21	16	16	4	6

In a number of fields at the secondary level, significant numbers of schools had some trouble filling their teaching vacancies. Even for English positions, often considered a surplus-ridden teaching field, over 10 percent of all schools had some difficulty getting a qualified candidate. This represents about a quarter of all those schools which had vacancies for English teachers. There were also some differences between schools in the likelihood of having hiring difficulties, but not as much as one might expect. In several fields, poor schools, for example, more frequently indicated hiring problems, but often the differences were slight.

Hence, large numbers of school principals had some degree of trouble finding qualified candidates to fill openings in their schools. Moreover, in 1990-91, 11 percent of all principals in the U.S. reported that they had openings that simply "could not be filled with a teacher qualified in the course or grade level to be taught." Despite these widespread difficulties in finding suitable candidates, however, there were very few teaching positions left unfilled in the U.S. In both the 1987-88 or 1990-91 school years, public districts reported, on average, less than 5 percent of their new openings were left vacant or were withdrawn because suitable candidates could not be found. For private schools, the proportion was less than 3 percent. Together these represented far less than 1 percent of all k-12 teachers employed. But, if there were extensive hiring difficulties, suggestive of shortages, why were there so few unfilled teaching positions in the U.S. - perhaps the most concrete indicator of a lack of shortages?

In reality, schools often simply cannot and do not leave teaching positions unfilled, regardless of supply. School staffing is legally mandated - public schools are obligated to provide teaching in subjects required by state law for graduation. Faced with this legal obligation, there are two general strategies by which school officials can reduce shortfalls between the supply of, and demand for, particular kinds of teachers. One involves altering the quantity of teachers demanded and the other involves altering the quantity of teachers supplied.

The first strategy is to decrease the demand for certain kinds of teachers by eliminating positions. This would inevitably result in increases in teachers' course loads, school class sizes, or pupil-teacher ratios. The second strategy is to increase or alter the quantity of teachers supplied. One version of this strategy alters the quantity supplied by filling a position with an underqualified candidate. This could be accomplished by shifting existing staff to areas of greater need; that is, assigning teachers trained in one field to teach in another. For example, social studies teachers could be assigned to teach mathematics courses. Alternatively, school officials could hire the available teacher candidates, regardless of qualifications.

The survey asked principals what means they actually used to cover a vacancy that could not be filled with a qualified teacher. These data for 1990-91 are displayed in table 2.

Table 2 - Percentage of secondary schools that used various methods to compensate for difficulties in filling vacancies, by school characteristics: 1990-91

	Public-Low Pov.	Public-Med. Pov.	Public-High Pov.	Private
Added sections	16	10	12	30
Expanded class size	13	11	9	15
Canceled classes	10	12	5	11
Used PT or itinerant teacher	14	6	12	18
Assigned another teacher	25	22	29	30
Hired less qualified teacher	25	27	21	18
Used substitute teachers	40	51	46	48

Interestingly, principals infrequently turned to the decrease-demand strategy to cope with hiring difficulties. Of public schools that experienced hiring problems, only about 10 percent either expanded class sizes, added additional class sections or canceled classes in order to cover their staffing shortfalls. There were some differences among schools. Poor public schools were slightly less likely to use these three methods, while private schools were slightly more likely to use them, especially the addition of class sections to existing staff.

Data from NCES' Common Core of Data survey corroborate that the decrease-demand strategy has not been used with frequency in recent years. The pupil-teacher ratio

in both public and private schools actually slightly dropped from 1980 to 1991. For public schools, the ratio decreased from 18.7 to 17.3. For private schools, the ratio decreased from 17.7 to 14.6. Moreover, data from the National Education Association show that the average number of students taught per day by public secondary school teachers, for example, declined from 118 to 93 between 1981 and 1991.

In contrast to the decrease-demand strategy, the data indicate that the increase-supply strategy has been commonly used. For both public and private schools, among the most common methods of coping with difficulties in filling openings in both 1987-88 and 1990-91 were to hire less qualified teachers, to assign teachers trained in another field or grade level to teach the understaffed subjects, and to use substitute teachers. For instance, in 1990-91, 50 percent of public secondary school principals who indicated they had difficulty filling openings, reported using substitute teachers as a remedy. Again, there were not large differences between different types of schools (see bottom half of table 2).

The cumulative effect of these 3 methods is to decrease the numbers of unfilled positions, and to increase the numbers of underqualified staff. Hence, the widespread use of this increase supply strategy necessitates a shift in focus for teacher supply assessments. Rather than focus on whether or not there are, or will be, sufficient numbers of potential or available teachers, these data suggest shortage assessments need to examine the actual fit between the needs of schools and the qualifications of the teachers currently employed. That is, the focus shifts from assessing the adequacy of the quantity of available teachers to assessing the adequacy of the quality of employed teachers.

Has the quality of teacher supply been adequate?

Assessing levels of teacher qualifications and quality, like assessing quantity, is a difficult and ambiguous task. How to define and measure a qualified teacher and quality teaching are subjects of great controversy (Ingersoll 1995b). There is, however, almost universal agreement that one of the most important characteristics of a qualified teacher is training and preparation in the subject or field in which they are teaching. Research has shown moderate but consistent support for the reasonable proposition that subject knowledge (knowing what to teach) and teaching skills (knowing how to teach) are important predictors of both teaching quality and student learning (for a review of this research, see Darling-Hammond and Hudson 1990). Knowledge of subject matter and of pedagogical methods do not, of course, guarantee qualified teachers nor quality teaching, but they are necessary prerequisites.

Hence, one method of assessing the adequacy of teacher supply is to focus on levels of basic teacher qualifications and training. But, it must be noted that the issue for assessing the adequacy of teacher supply is not a lack of basic training and qualifications on the part of teachers. The data indicate that most teachers in the U.S. have basic training. For example, 98 percent of all teachers newly hired in the 1990-91 school year held a bachelor's degree and over a third had obtained a graduate degree. Moreover, 88 percent of these newly hired held teaching certificates. The issue in question is the phenomenon of out-of-field teaching - teachers assigned to teach subjects that do not fit their fields of training. The last portion of this analysis will focus on out-of-field teaching as an indicator of inadequacies in the available supply of teachers.

Of course, it must be noted that some degree of out-of-field teaching may be unavoidable and may not always be an indicator of a shortage of qualified and available teaching candidates. School administrators charged with the task of offering programs in a range of required and elective subjects may often be forced to make spot decisions concerning the assignment of available faculty to an array of changing course offerings. But even low levels of out-of-field teaching are meaningful to teacher supply assessments. This is especially true for the case of high schools and for the core academic fields. In high schools, teachers are divided by fields into departments; faculties are thus more specialized than in elementary schools, and therefore the differences between fields are more distinct and, perhaps, greater. Moreover, the level of mastery in different subjects is higher in high schools, and hence a clear case can be made that teachers ought to have adequate background in the subjects they teach. Hence, the remaining portion of this section focuses on the levels of and variations in out-of-field teaching in high schools.[2]

Table 3 — Percentage of high school teachers who taught one or more classes in a field without at least a minor in that field, by field and school characteristics: 1990-91

	Math	Science	Social Studies	English
Total Overall	32.1	18.7	18.9	23.2
Public	30.5	16.9	16.9	21.9
Poverty Enrollment				
Low	27.7	14.0	15.7	19.2
Medium	31.8	20.3	19.2	24.5
High	40.0	20.2	18.0	30.7
Private	41.0	28.6	30.3	32.0

In fact, substantial numbers of high school teachers were assigned to teach out of field or out of department in both 1987-88 and 1990-91. While most high school teachers had a undergraduate or graduate major in their main teaching assignment field, large numbers of teachers

were assigned to teach courses in additional fields for which they did not have a major or even a minor. In 1990-91, public high school teachers taught, on average, about 15 percent of their class schedules in fields for which they did not have even a minor. This amounted to about one course in six. Private high school teachers taught far more of their classes without minimal qualifications. On average, for about one-quarter of their scheduled classes, teachers did not have at least a minor in the field. These percentages all substantially increase (sometimes double) if the standard is raised from a minor to a major in the field taught. As a result, substantial numbers of high school students were taught core academic classes by teachers without even minimal training in the field. These levels of out-of-field teaching, however, varied substantially by field, as shown in table 3.

In 1990-91, 23 percent of all high school English teachers did not have at least a college minor in English, language arts, journalism or communication. Thirty two percent of all high school mathematics teachers did not have at least a minor in mathematics or mathematics education. Nineteen percent of high school science teachers did not have at least a minor in any of the biological, physical or natural sciences or science education. Nineteen percent of high school social studies by teachers did not have at least a minor in history, any of the social sciences or social studies education.

Out-of-field levels also varied considerably across different types of schools. Notably, public schools with a high proportion of poverty-level students had a higher proportion of out-of-field faculty in mathematics, science, and English than schools with less than 20 percent poverty-level students. In several fields, these high levels were overshadowed by those in private schools, in which, for example, 59 percent of mathematics teachers and 47 percent of English teachers out of field.

Conclusion

This paper addresses the ongoing debate as to whether there are shortages of teachers in the U.S. If one accepts the premise that adequate staffing requires high school teachers to hold at least a college minor in the fields which they teach, then this analysis suggests that many of the nation's schools have not been adequately staffed.

Analysts have offered three possible explanations for inadequacies in the supply of teachers. Some have suggested that inadequacies are due to insufficient training of teacher candidates. Some have suggested that shortages are due to insufficient numbers of trained teachers. Finally, others have suggested that staffing inadequacies are due to an inability of many schools to attract adequate numbers from the pool of existing trained teacher candidates to seek positions.

First, are staffing inadequacies, such as out-of-field assignments, due to inadequacies in the qualifications of the supply of teachers? That is, is out-of-field teaching a problem of poorly trained teachers? In fact, the data suggest that the prevalence of out-of-field teaching is not due to a lack of basic teacher training. Most high school teachers in the United States had completed a college education and, indeed, over half had acquired graduate degrees. The inadequacies lay in the fit between teachers' fields of training and their teaching assignments. Many teachers were assigned to teach classes which did not match their education or training. Hence, increased and improved teacher training, while a worthwhile goal and the object of much current research and reform, may not reduce levels of out-of-field teaching.

Second, are staffing inadequacies, such as out-of-field assignments, due to inadequacies in the quantity of the supply of teachers? That is, is out-of-field teaching a problem of too few teachers? In fact, the data suggest that the supply of potential teachers in the larger population is both large and diverse. Only a small proportion of the newly hired come directly from training institutions; a large proportion are either re-entrants or delayed entrants. This suggests that out-of-field teaching assignments are not due to insufficient numbers of trained teachers and, thus, for example, increasing enrollments in teacher training programs, the goal of some current education reforms, may not be an effective method of reducing levels of out-of-field teaching.

But, despite the large and diverse reserve pool and the widespread extent of basic training held by teachers, many school principals report experiencing difficulties in hiring qualified candidates, As a result, they turn to the use of substitute teachers, in-school reassignments and hiring of the underqualified as strategies for coping with these difficulties. Hence, although there may be many reasons for out-of-field assignments, a leading factor appears to be the inability of schools to obtain or retain sufficient numbers of candidates from the existing pool of trained teachers.

The data, however, do not establish the sources of this inability. For example, it is unclear if out-of-field assignments are an emergency condition resulting from spot shortages of particular types of teachers at particular times in particular places, whether they are a short-term condition due to fiscal constraints in particular settings, or to what extent they are a chronic condition because this is a normal and ongoing practice in particular schools. Moreover, if most out-of-field teaching is a remedy for difficulties in hiring, it is not at all clear whether the root of the problem is the unwillingness of existing trained teacher candidates to seek positions, or whether the root of the problem is the unwillingness of schools to attract, effectively utilize and retain existing trained teacher candidates, or both.

Whatever the reasons, the data suggest a story that is both provocative and unsettling: There has not been shortages in the quantity of available elementary and

secondary school teachers in this country. But, our analysis suggests there have been, in fact, distinct inadequacies in the quality of available elementary and secondary school teachers in this country. Schools have filled teaching positions, but only at the expense of minimal standards of teacher qualification. The result: teacher quality has been sacrificed for teacher quantity, rendering the teacher shortage "invisible."

Endnotes

1. SASS data tapes, survey questionnaires and user's manuals are available from NCES, U.S. Department of Education, 555 New Jersey Ave., Washington, D.C. 20208-5641. For an extensive report summarizing the items used in this investigation and providing an overview of the entire survey see Choy et al. (1993).

2. This analysis of out-of-field teaching borrows heavily from the larger study on teacher supply, qualifications and turnover mentioned earlier (see Ingersoll 1995a). As the report shows, out-of-field teaching can be empirically measured in a number of ways. The measure of out-of-field teaching used here focuses on whether each of those, who taught one or more classes, in each of 4 broadly defined fields, had a minimum of substantive training in that field. More specifically:

(A.) substantive training - I focus whether teachers had formal training in a discipline, rather than formal training in teaching methods and pedagogy i.e. certification.

(B.) minimal levels - I focus on whether teachers had at least a college minor in the field.

(C.) broadly defined fields - Fields are defined parallel to conventional departmental divisions in high schools. That is, fields include all within-department disciplines. Hence, for example, a minor in any of the natural, physical or biological sciences is considered adequate training to teach any science course.

References

Choy, S., Henke, R., Alt, M., Medrich, E. & Bobbitt, S. (1993). Schools and Staffing in the U.S: A Statistical Profile, 1990-91. Washington, DC: US Department of Education, National Center for Education Statistics.

Darling-Hammond, L. (1984). Beyond the commission reports: The coming crisis in teaching. Santa Monica, CA: Rand Corporation.

Darling-Hammond, L., Hudson, L. (1990). "Pre-college Science and Mathematics Teachers: Supply, Demand and Quality. Review of Research in Education. Washington, DC: American Educational Research Association.

Feistritzer, E. (1986). Teacher Crisis: Myth or reality? Washington, D.C.: National Center for Education Information

Grissmer, D. & Kirby, S. (1987). Teacher attrition: the uphill climb to staff the nation's schools. Santa Monica, CA: Rand Corporation.

Grissmer, D. & Kirby, S. (1992). Patterns of attrition among Indiana teachers, 1965- 1987. Santa Monica, CA: Rand Corporation.

Haggstrom, G. W., Darling-Hammond, L., & Grissmer, D. (1988) Assessing teacher supply and demand. Santa Monica CA: Rand Corporation.

Ingersoll, R. (1994). "Teacher Shortages and Teacher Quality." In The Proceedings of the American Statistical Association: 1994. Alexandria, Va: American Statistical Association. .

----. (1995a). Teacher Supply, Teacher Qualifications and Teacher Turnover:1990-1991. Washington, DC: US Department of Education, National Center for Education Statistics.

----. (1995b). National assessments of teacher quality. Washington, DC: US Dept.t of Education, National Center for Education Statistics.

Listing of NCES Working Papers to Date

Number	Title	Contact
94-01	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05	Cost-of-Education Differentials Across the States	William Fowler
94-06	Six Papers on Teachers from the 1990-91 SASS and Other Related Surveys	Dan Kasprzyk
94-07	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01	Schools and Staffing Survey: 1994 papers presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk

Listing of NCES Working Papers to Date (Continued)

Number	Title	Contact
95-04	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings
95-06	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08	CCD Adjustments to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12	Rural Education Data User's Guide	Samuel Peng

Listing of NCES Working Papers to Date (Continued)

Number	Title	Contact
95-13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17	Estimates of Expenditures for Private K-12 Schools	Steve Broughman
95-18	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk
96-02	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

TM028246

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").